DIVERSITY @ ECAI 2016

INTERNATIONAL WORKSHOP
ON DIVERSITY-AWARE ARTIFICIAL INTELLIGENCE

The Hague, Netherlands, 29th August 2016

Workshop Proceedings

Diversity is pervasive in human nature and culture, and is deeply rooted in the variation of natural traits and experience among individuals, the collectives they form, and the environments they inhabit. When humans reason individually, they maintain different representations, conceptualisations, and theories, and apply different rules of inference, learning, and decision making. When they interact with each other to combine their skills or resources, to coordinate their activities, and to resolve conflicts between their individual objectives, they exchange information and knowledge, negotiate and align their individual views, and adapt to each other's behaviour dynamically. Arguably, diversity is not only a phenomenon that humans have to deal with, but it is also the vehicle for achieving some of the most impressive products of human intelligence.

Artificial Intelligence, on the other hand, has so far largely relied on a certain degree of homogeneity, not necessarily in terms of the components involved in a method or system, but in terms of the process that combines them. Various areas within AI have already developed methods that can cope with and/or exploit diversity to some extent, for example:

- electronic markets where individual agents have different goals and aim to maximise their own profit;
- hybrid robot architectures that involve different layers of representation and reasoning;
- knowledge sharing infrastructures where different agents use different domain ontologies; and
- machine learning systems that combine different sources of data and/or learning units.

However, more often than not, these systems still involve a 'monolithic', global approach to integration. This usually derives from a global task context, a common intermediate representation layer, or a global output to be produced by the integrated system.

We believe that there is a huge potential in bringing the insights from work on problems that involve diversity—like those listed in the examples above—together to gain a deeper understanding of the phenomenon of diversity, as well as to develop principled methodological approaches that will enable us to better utilise diversity in future AI systems.

— The Organisers

Organisers

Michael Rovatsos, The University of Edinburgh
Ronald Chenu-Abente, University of Trento

Steering Committee

Alan Bundy, University of Edinburgh, United Kingdom
Peter Gardenfors, University of Lund, Sweden
Fausto Giunchiglia, University of Trento, Italy
Asunción Gómez Pérez, Universidad Politécnica de Madrid, Spain
Ben Kuipers, University of Michigan, USA
Ariel Procaccia, Carnegie-Mellon University, USA
Carles Sierra, IIIA-CSIC Barcelona, Spain
Luc Steels, Vrije Universiteit Brussels, Belgium
Michael Wooldridge, University of Oxford, United Kingdom
Gerhard Weiss, University of Maastricht, The Netherlands

Programme Committee

Yoram Bachrach, Microsoft Research Cambridge, United Kingdom
Gábor Bella, University of Trento, Italy
Sofia Ceppi, University of Edinburgh, United Kingdom
Jérôme Euzenat, INRIA Grenoble, France
Kobi Gal, Ben-Gurion University of the Negev, Israel
Fabien Gandon, INRIA Sophia-Antipolis, France
Mark Hartswood, University of Oxford, United Kingdom
Nick Hawes, University of Birmingham, United Kingdom
Catholijn Jonker, Technical University of Delft, The Netherlands
Ian Kash, Microsoft Research Cambridge, United Kingdom
Oliver Lemon, Heriot-Watt University Edinburgh, United Kingdom
Nicolas Maudet, Université Pierre et Marie Curie Paris, France
Fiona McNeill, Heriot-Watt University Edinburgh, United Kingdom
Roberto Navigli, University of Rome “La Sapienza”, Italy
Luc Moreau, University of Southampton, United Kingdom
Iyad Rahwan, MIT, USA
Subramanian Ramamoorthy, University of Edinburgh, United Kingdom
Katharina Reinecke, University of Washington, USA
Robert van Rooij, ILLC University of Amsterdam, The Netherlands
Carlos Ruiz, TAIGER S.A., Spain
Marco Schorlemmer, IIIA-CSIC Barcelona, Spain
Onn Shehory, IBM Haifa Labs, Israel
Pavel Shvaiko, Informatica Trentina, Italy
Remi van Trijp, Sony Computer Science Labs Paris, France

Acknowledgment

The Organisers are grateful for the support received from the SmartSociety project (www.smart-society-project.eu) and the ESSENCE Marie Curie Initial Training Network (www.essence-network.com), both funded by the European Commission's 7th Framework Programme under grant agreements no. 600854 and 607062.

Agenda

Towards Building Ontologies with the Wisdom of the Crowd

Paula Chocron[1] and **Dagmar Gromann**[2] and **Francisco José Quesada Real**[3]

Abstract. Crowdsourcing provides a valuable source of input that reflects the human diversity of domain knowledge. It has increasingly been used in ontology engineering and evaluation, however, few approaches consider different types of crowdsourcing for data acquisition. In this paper, we compare two crowdsourcing techniques - a mechanized labor-based task and a game-based approach - to acquire shared knowledge from which we semi-automatically build an ontology. This paper focuses on the first two steps of ontology engineering, the forming of concepts and their hierarchical relations. To this end, we adapt a distributional semantic and class-based word sense disambiguation approach and a knowledge-intensive tree traversal algorithm. Each step along the process and the final resources are evaluated manually and by a gold standard created from Wikipedia data. Our results show that the ontology resulting from data obtained with the mechanized labor-based approach provides a higher level of granularity than the game-based one. However, the latter is faster and seems more enticing to participants.

1 INTRODUCTION

Creating knowledge resources manually is a time- and cost-intensive task [26], and the resulting resources are in general difficult to maintain. Moreover, when resources are created by individuals (*experts* in the domain and the technique), in many cases they are not free from arbitrariness. The default alternative to manually crafting knowledge resources is to develop techniques that automate the process, or at least parts of it. The ontology learning community has developed different automated approaches, using tools that range from machine learning [16] to NLP-intensive approaches [22]. These methods extract information from either a structured (e.g. WordNet) or unstructured (e.g. text) existing corpus, and are therefore strongly dependent on the existence and quality of such a corpus. As an alternative, and paired with a general growing interest in these kind of techniques, in the past years the community has proposed different applications of crowdsourcing methods to ontology engineering (e.g. [7, 15, 28]). We contribute to this community effort by comparing two distinct crowdsourcing approaches to the task of knowledge acquisition for building ontologies semi-automatically.

Crowdsourcing is a problem-solving method that relies on a collective of non-experts (a *crowd*) performing short and accessible tasks that are then combined to tackle a larger problem. Crowdsourcing methods are particularly well suited for tasks that are difficult to automatize completely, but are at the same time too large to be completed by just one person, or that benefit from the diversity of the participants, as is the case with our approach. This includes, for example, many information retrieval or classification tasks, often in complex human domains, such as natural language. The question of how to increase the attractiveness of crowdsourcing methods to make the participation more appealing has received much attention as of late. While one way is to provide explicit, in general monetary, incentives, other methods rely on intrinsic rewards, such as learning a language [23], helping a cause, or having fun. This last category is particularly exploited via the *Games With a Purpose* approach [21, 24].

This paper proposes an ontology learning technique that combines crowdsourcing to retrieve data with automated methods to organize it. Instead of crowdsourcing the ontology building process as it is frequently done, we leverage diversity by crowdsourcing the data acquisition step. Thereby, we obtain domain knoweldge that reflects the human diversity of domain knowledge and brings ontologies closer to their initial aim of representing shared knowledge. We build two ontologies from scratch using the data obtained from two separte crowdsourcing methods and then compare them to each other as well as to a third gold standard ontology obtained from Wikipedia data. While this knowledge production technique has all the advantages of collaborative methods, the obtained data is usually not organised, which represents a technical challenge when building an ontology with it. Thus, we implement and compare different methods to disambiguate the retrieved data categories and we build a taxonomical structure with it.

We focus on the task of building an ontology for a particular concept, identifying all the related categories that could be used when describing an instance. We chose to perform our experiments using the concept of *city*, mainly for three reasons. First, it is a topic with which the crowds are in general familiar. Second, it belongs to a category of particularly fuzzy, collectively constructed concepts, which makes it ideal to be crowdsourced. Third, a sound representation of city has become something particularly necessary in the last years, with the growing interest in visions such as the one of Smart Cities [3]. The ability of a city to share and re-use data has become a key indicator for a Smart City and a domain ontology that contains categories typically characterizing a city can facilitate this task as well as the integration of data across Smart Cities.

To obtain these ontologies, we first implement two crowdsourcing methods (a direct and a game-based one) in which we ask participants to describe instances of a city on $CrowdFlower$[4] and in a game we developed. We consider this kind of crowdsourcing *implicit*, since participants have to solve a different problem from which the desired data are then extracted in a post-processing phase. An *explicit*

[1] Artificial Intelligence Research Institute (IIIA-CSIC) and Universitat Autònoma de Barcelona, email: pchocron@ iiia.csic.es
[2] Artificial Intelligence Research Institute (IIIA-CSIC), email: dgromann@iiia.csic.es
[3] University of Edinburgh, email: fquesada@inf.ed.ac.uk

[4] `https://crowdflower.com/`

approach would consist in asking people for characterizations of the general concept of city itself. Implicit crowdsourcing techniques are useful in order to make the task more attractive, fun, or *gamifiable* than the explicit approach. We also believe that it can lead to richer and more fine-grained ontologies than the explicit one. However, the direct comparison between explicit and implicit crowdsourcing is yet to follow. The kind of techniques we propose here is particularly applicable when describing abstract concepts that do not have a clear physical correspondence, where the properties are less evident.

Our post-(crowdsourcing)-processing phase consists in extracting categories related to cities from the crowdsourced description by analyzing the obtained natural language expressions. To this end, we first disambiguate the senses of these expressions, for which we implement two techniques - a distributional semantics and a class-based approach. We also consider the next step in ontology building, which is adding a taxonomical backbone to the resource by relating the disambiguated concepts hierarchically and extracting their superordinate classes. Finally, we evaluated our approach by comparing its results to an existing, also crowdsourced, description of cities that we extract from the Wikipedia Tables of Contents (TOCs) of individual city pages.

After discussing related work, we describe our approach following the traditional structure of method, results, and discussion. We first explain the techniques that were implemented for each step, then present the results obtained with each of them, and finally compare them and discuss advantages and drawbacks of each one. In the last section we present future work and some concluding remarks.

2 RELATED WORK

Due to the difficulties that the manual crafting of ontologies present, the field of ontology learning has been extensively studied in the past years [12]. Many of these approaches, particularly those in the first years of the area's development, rely on predefined patterns and rules or static resources, such as WordNet [26]. However, these static approaches have two drawbacks, namely they are neither scalable nor easily portable between domains. Recent approaches seek to be more dynamic, for example by using machine learning to extract relations from an existing seed ontology [16] or to develop axioms extracted from text [22].

Using static resources in ontology learning is not straightforward due to the multiplicity of senses associated with each word. To address this problem, Bentivogli et al. [2] associate senses with a WordNet domain ontology they create and which we also use herein to classify words. A similar idea is presented by [8] who associate the Kyoto ontology of the project with WordNet senses and also a number of upper level ontologies. Those associations are then used to present a class-based word sense disambiguation method we adapt in this paper. Alternatively, distributional semantic approaches have been investigated for word sense disambiguation with context-poor data sets. For instance, Basile et al. [1] extract DBpedia glosses for each word in tweets and then compute the cosine similarity between the context of the word in the tweet and each gloss to find the most related one(s), a second approach we adapt in this paper. Similarity between sets of words can be computed by composing their vectors in different ways; in [1] the authors use addition.

The use of crowdsourcing techniques has received considerable attention across research fields in the past few years [27]. For instance, crowdsourcing is highly popular in Natural Language Processing (NLP), such as for named entity recogntition [9]. In ontology learning and building, crowdsourcing has mainly been used in an explicit fashion, asking users to relate concepts hierarchically [6] or evaluate already learned relations and term clusters [7]. Additionally, it has been used as a method to align ontologies with each other [17, 20]. Most frequently, crowdsourcing has been applied to ontology evaluation both for verifying subsumption hierarchies [15] as well as entire ontology statements [28].

Among these crowdsourcing techniques, games are particularly important since they offer an interesting way to motivate humans to solve large-scale problems that are currently beyond the ability of computers [18, 24]. Some well-known examples are Duolingo [23], an approach to crowdsourcing the translation of the Web, and re-CAPTCHA [25], a method for digitizing paper copies of documents. Approaches that use 'Games with a Purpose' build on the instrinsic motivation of participants to learn something new. For instance, Dumitrache et al. [5] use gamification and crowdsourcing to create a gold standard for annotations of medical texts. Luengo-Oroz et al. [11] develop a game for counting malaria parasites in images of thick blood films, while Deng et al. [4] and Zou et al. [29] focus on feature discovery and image categorization. Individual ontology engineering tasks have been crowdsourced as games as well, such as for classification and population [19]. In [14] a game is proposed to obtain attributes for concept descriptions. Their approach is explicit in that it asks players to name properties directly. In combination with ontologies, a specific part of the ontology building task is usually crowdsourced but not the knowledge acquisition step that precedes the ontology building as in our approach.

3 METHOD

In this paper we present a method to build ontologies for the concept of *city* from data obtained with crowdsourcing techniques. We use two different implicit crowdsourcing methods, in which we ask participants to describe specific instances of cities as direct question and in a game to obtain a general characterization of *city* as a general concept. We consider *city* to be a particularly good concept to perform this experiment, since it has clear instances which are in general well-known by a random crowd. In addition, although a city can be uniquely identified by means of its coordinates, these are in general not the most immediate characteristics that come to mind, and the resources used when describing an instance are very varied.

From the descriptions obtained with the crowdsourcing methods, we extract general categories on which we build a hierarchical taxonomy to obtain a preliminary ontology for the concept of *city*. We consider the results obtained to be seed ontologies that can be used for further ontology learning rather than fully formalized ontologies; nevertheless, they can be seen as a schema of a city characterization. To evaluate this claim, we compare them to a gold standard ontology that we manually and collaboratively build from Wikipedia TOCs of pages describing specific instances of cities, countries, regions, and continents. The complete process of our approach is depicted in Figure 1, where rectangles are steps and circles are the different techniques that we explore.

3.1 Data Collection

Our method for collecting data by means of crowdsourcing can be subdivided into two separate techniques: (1) mechanized labor-based knowledge acquisition, and (2) game-based knowledge acquisition. Mechanized labor refers to popular crowdsourcing platforms where people complete mechanical tasks in exchange for monetary rewards, e.g. *CrowdFlower* or *Mechanical Turk*. In this type of data collection

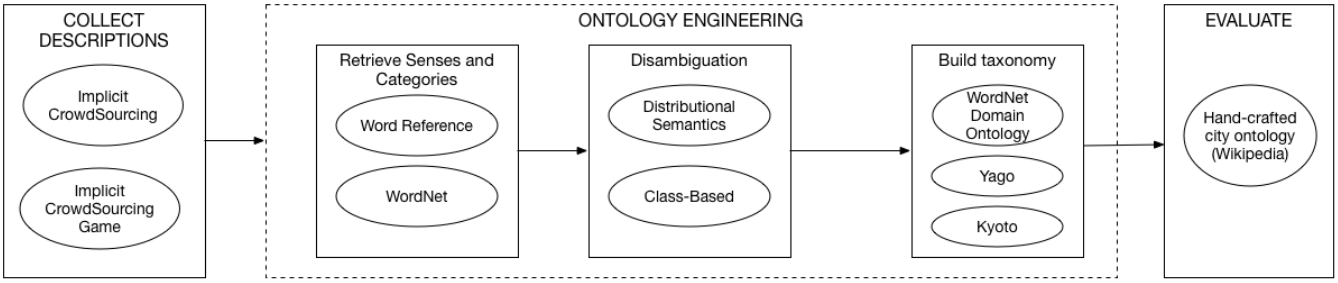


Figure 1: Steps followed in the ontology building procedure

method participants were asked to provide the first ten words they associate with a city name displayed to them. In contrast, in a game-based elicitation of knowledge participants provide the desired information while playing a game without being asked direct questions. We utilized a list of 300 city instances derived from online listings of popular cities that were retrieved by a search engine query. In both tasks it was possible to skip to the next city if a participant was not familiar with a specific city instance.

Both tasks focus on the collection of common nouns in combination with verbs and adjectives. There are two main reasons for this restriction: (1) proper nouns trivialize the identification of cities as they uniquely identify them, e.g. Eiffel for Paris, and (2) we were interested in ontology building from common language and not based on instances or named entities. In both types of activities participants were explicitly instructed to comply with this input restriction. Additionally, measures were taken in both tasks to enforce this restriction and the non-conforming characterizations were omitted from the final data set. The results from the first data set are distinct from the second data set since the nature of the game required us to provide descriptions of the city obtained from the first technique as input for the game.

3.1.1 Mechanized Labor-Based Knowledge Acquisition

To obtain first characterizations of cities, we uploaded the list of 300 cities to the crowdsourcing platform *CrowdFlower*. Questions presented to the crowd provided the name of the city, its country name, latitude and longitude, ten input fields for city descriptions, and the option to skip to the following city if the city was not known to the worker. In addition, each worker was asked 20 test questions to ensure their ability to comply with the instructions regarding the input restrictions, such as use of a common noun or noun phrases with adjective and verbs, use of loan words but no words that are not English, and omission of personal opinions. For instance, we asked workers whether Breaking Bad is an adequate description of Albuquerque, USA. Since this is the title of a TV series and thus, a named entity, this question had to be negated. The ability to comply with instructions was also tested by using misleading descriptions, such as the description of Liverpool with U2. Each test question was equipped with a detailed explanation for the correct answer so that participants who did not fully read or understand the instructions were prepared for the actual question of the task.

For quality assurance four measures were taken: (1) the actual run was preceded by a test run, (2) each worker was asked twenty test questions, (3) only workers with an accuracy exceeding 70% on the test questions could participate in the task, and (4) only workers who spend more than ten seconds on each question apart from the test questions would be considered. Furthermore, we limited this task to workers with English as their first language since we required an English data set and such word association tasks are difficult in a second language. An initial test run with a subset of the cities helped evaluate the kind of results we were to expect and modify the test questions and project settings based on the feedback from the crowd. In fact, those modifications strongly improved the quality of the results as well as the time needed to obtain them in the actual second run.

Obtained city characterizations were deduplicated automatically by applying similarity measures from the WordNet Similarity for Java (WS4J) library[5] combined with the Levenshtein distance [10]. On this basis the most frequently provided and deduplicated city descriptions were identified and then evaluated manually.

3.1.2 Game-Based Knowledge Acquisition

In this second crowdsourcing technique, participants played a Taboo game of cities adapted from the popular board game *Taboo*. There are two roles a player might assume: describer and guesser. The describer provides hints to the guesser that describe a given city and the guesser responds with a city name that is believed to be the correct result. The objective of the game is to obtain the name of the described city from the guesser. As a further restriction, the describer may not use any of the phrases that are provided as taboo words along with the city.

The taboo words of this game were obtained from the first data collection method. Thereby, it was ensured that there is no overlap between the data set gathered with the first collection method and this second crowdsourcing method. Additionally, in order to play this game, Taboo words are needed. For the hints, the same conditions as in the first method were applied for the same reasons. This meant that we needed to limit the type of hints people provide when playing the game.

Players were recruited at the University of Edinburgh by means of internal mailing lists and personal contacts of our local colleagues within the ESSENCE project [6]. As with the first technique, we restricted the participation to native English speakers. Each participant obtained a small shopping voucher in return for their participation. The number of games per participant was not limited.

To ensure that the input complied with our restrictions and to enable several simultaneous games, we developed an online platform [7] and pre-scheduled game sessions with up to nine players at a time.

[5] `https://code.google.com/archive/p/ws4j/`
[6] `http:\essence-network.com`
[7] `http:\taboo.iiia.csic.es`

The first player to log onto a game would be assigned the guesser role. The second player to join a game session would be the describer, who in contrast to the guesser would see the city name, country, and Taboo words. The game commences by the describer providing a hint and ends with the correct guess from the guesser. Players were newly assigned automatically and anonymously to each game. This should prevent participants from providing clues based on previous experiences in case of acquainted players.

The final data set is limited to successful games that follow the restrictions of the initial instructions. A successful game is one were the city was guessed correctly based on the provided hints. This ensures the quality of the hints, i.e., they are indeed associated with the city being described to a degree that allows a human player to identify the city. Naturally, there might be many reasons for the inability of a guesser to provide the correct city name, which, however, we did not investigate for this paper and instead relied on the quality-assured hints of successful games.

3.2 Ontology Engineering

The task of building ontologies, known as ontology engineering, is commonly divided into four major steps that can be implemented with different engineering methods:

1. concept formation
2. concept hierarchy building
3. building non-taxonomic relations
4. axiom discovery

This list is non-exhaustive, and some approaches also include, for instance, ontology population as another step. In this section, we present the methods we implemented for building ontologies for the *city* concept from the data sets that resulted from the two crowdsourcing methods described previously. In this paper we focus on the first two steps of ontology engineering: forming the concepts and building a hierarchy. As we discuss later, the third step can be initiated with the methods we use but will be the subject of another paper since we do not evaluate it here.

Concept formation refers to the process of clustering terms based on their more general categories. Thus, for this step we required the general concepts related to *city* that were represented by the descriptions of city instances. For example, if sun was related to Barcelona, we wanted to extract weather as a general characteristic of a city. To this end, we retrieved the available senses and classifications for each noun and noun phrase from WordNet and an online dictionary. We noticed that, although these methods return adequate categories, an unexpected level of complexity arises given the multiplicity of senses that exist for each hint. Thus, our method required a step of word sense disambiguation for which we implemented two ideas: (a) a distributional semantic approach taking the city as context, and (b) a class-based approach that does not consider the city. Finally, we present our method to build the taxonomy, extracting more general concepts for the categories obtained. In this section we explain the techniques we used for each of these ontology engineering steps depicted in Figure 1.

3.2.1 Sense Extraction and Classification

The first of our approaches to extract general categories from descriptions of specific cities uses Word Reference [8], an online dictionary that associates words with general labels that can be generally seen as its superordinate class. For example, Sushi is labeled as Food. To use this information, we first extract all nouns in city descriptions and retrieve all existing glosses and categories from Word Reference. A second approach consists in using WordNet to obtain the categories. Due to the fine-granular nature of WordNet senses, it was necessary to use ontologies associated with WordNet synsets to obtain general categories, as described in detail below.

3.2.2 Sense Disambiguation

With context-poor and highly ambiguous input data, word sense disambiguation for the purpose of term clustering and concept formation is a highly challenging task. At times the disambiguation is not even easy for human users, e.g. curse for Cairo could relate to a film, urban legends, or verbal expressions. Both data sets derived from the described crowdsourcing techniques consist of single common nouns or noun phrases with their associated city name as the only context. To address this challenge, we tested two different approaches to word sense disambiguation: (1) a distributional semantic approach, and (2) a class-based approach. In both cases the objective is the identification of the sense that is most closely related to a city, which is then used to form ontology concepts.

All initial input data were submitted to an NLP preprocessing step to identify all common nouns in the data set and lemmatize them. For this we used the NLTK[9] in Python for the distributional approach and the Stanford CoreNLP library[10] in Java for the class-based approach for no reason other than the personal preference of the developers. The former used individual tokens only, while the latter approach first queried noun phrases. If the noun phrases returned no result, the head noun of the phrase was identified by CoreNLP and submitted to the sense query component.

Distributional Semantics-Based Disambiguation Due to the nature of our data there is no real context for the words used. Therefore, our approach consists in computing the similarity of each definition of a word extracted from the lexical resources with the vector of the city. For example, if Paris was described with love, for which we retrieved three definitions, we compute the vector for each definition and their similarity with the vector for Paris, and chose the one with the highest score. After some initial experiments we combined the vector of the city with the vector for each data element from the crowdsourcing techniques since it substantially improved the disambiguation of the word's senses. For instance, in the example above we would combine Paris and love and then compare the result to the three glosses retrieved from the lexical resource. Instead of DBPedia, we opted for an extraction of senses from Word Reference and WordNet since it is faster and less noisy. Furthermore, the categories retrieved along with the senses in Word Reference seemed promising for the classification task. We implemented two ways of composing vectors: the addition used in [1] and a simple average of individual vectors, that is the standard way to compute similarity between sets of words in the word2vec Python package. We chose this last option after performing a general initial comparison.

Class-Based Word Sense Disambiguation To follow up on a second idea, we investigated a class-based sense disambiguation approach adapted from [8]. Although the use of WordNet to disambiguate words is wide-spread, one of the major issues is the high

[8] http://www.wordreference.com

[9] http://www.nltk.org/

[10] http://stanfordnlp.github.io/CoreNLP/

granularity of its senses. For instance, querying architecture returns five distinct senses ranging from architecture as a profession to computer architecture. One method to alleviate this situation is the semantic classification of WordNet senses by using associated ontologies. The approach in [8] associates WordNet senses semi-automatically with the ontology Kyoto[11].

In this three-step algorithm, we first extract all senses associated with an input noun or phrase from YAGO[12] and query Kyoto for each association with each retrieved sense. In a second step, the algorithm traverses the sense hierarchy in YAGO and searches for categories by again querying Kyoto and searching for WordNet domains associated with individual senses. Since the mapping to WordNet domains is not consistent in YAGO, each sense label queries the WordNet domain ontology for string matches and adds them to the resulting collection of categories and senses. The third step consists of extracting all tokens from each label of a category and ranking them according to frequency. To find the best sense, the most frequent word of all senses and the previously evaluated Word Reference categories from the distributional semantics approach are utilized as determining factor on which sense to return. The extracted and evaluated Word Reference category is added as an additional weight to the decision of which category to chose as the final one and the same approach could be done without this additional weight. Queries to WordNet that immediately return a WordNet domain along with the senses are not submitted to this process but instead classified by the domain directly.

3.2.3 Taxonomy Building

In order to build a hierarchical backbone for an ontology, we query YAGO, WordNet domains, and Kyoto relations. Although some of the upper ontologies in Kyoto are highly useful, we exclude DOLCE since it is too abstract for our purpose, namely building a resource that represents categories associated with the general concept of *city*. The senses we obtain from the word sense disambiguation tasks are utilized to retrieve the Kyoto concepts and WordNet domains directly associated with the sense. Additionally, we traverse the YAGO sense hierarchy up two levels to obtain all senses and domains associated with the disambiguated and evaluated sense. If the word or sense is directly associated with a WordNet domain we extract the superordinate level of the respective domain in the WordNet domain ontology where available. The WordNet domain ontology currently only provides one hierarchical level associating domains with their more general level. If there is no WordNet domain we query Kyoto and extract all concepts that are associated with a sense by means of a subclass relation. The focus here due to the data set is on nouns, which is why we do not extract any concepts related to verbs or adjectives. In case this step returns several concepts, we manually select the best hierarchy for a given WordNet sense.

3.3 Evaluation

Each word sense disambiguation approach is evaluated manually by at least two fluent/first language English speakers. For WordNet, the senses were rated regarding their correct specification of the input description as either *correct* or *incorrect*. For Word Reference, both the categories and the definitions were rated since the former was used as a weight in the taxonomy building task. Only senses and data on which both raters agreed were submitted to the ontology engineering task. The resulting seed ontologies with a hierarchical backbone were again manually evaluated by two ontology engineers.

In a second evaluation step for the ontologies, we compared them with another crowdsourced classification of concept properties that is obtained from describing city instances: one obtained from the TOCs of Wikipedia. The usefulness of TOCs of Wikipedia for building knowledge resources has been acknowledged before [13]. Each Wikipedia page of a specific city is organised in a tree of sections (for example, dog has the subtree Biology → Anatomy → Size and Weight). These TOCs work naturally as an organisation of categories that are important to describing something. Moreover, although Wikipedia establishes certain patterns that authors should follow, TOCs are mostly originated from a collaborative attempt at describing things in the world, here cities.

To build a general ontology for city descriptions we chose 20 random cities from the list of cities we used for the crowdsourcing and merged the TOCs in their Wikipedia pages, keeping the most general ones. In this way, we removed categories that were very specific to one city or region (such as "2.1.1 Legend of the founding of Rome" for the city of Rome). This was done by four ontology engineers in a collaborative shared task to avoid personal biases. Each created an ontology from five different cities, and then all together collaboratively discussed how to merge them to get a common taxonomy. In general it was easy to achieve an agreement, which suggests a high degree of consistency in Wikipedia's TOCs.

We repeated the same process for countries, regions, and continents and merged the final result to a four-layered knowledge resource reflecting the four main levels we found in the city descriptions. At times people utilize those levels of granularity to describe a city. For instance, nasal vowel relates to Portuguese and Portugal rather than Lisbon while wall clearly relates to the city of Berlin. However, both are used to describe the respective cities in the crowdsourcing tasks.

4 RESULTS

Results are structured in line with the method section to facilitate their traceability. We first report on the obtained data sets from the two distinct crowdsourcing techniques before we detail the results of the ontology engineering method. The evaluation of the word sense disambiguation methods was done by human users and the resulting seed ontologies from both crowdsourcing data sets were evaluated by using a manually curated gold standard ontology based on Wikipedia TOCs.

4.1 Data Collection

From the CrowdFlower platform we derived a total of 6,238 descriptions for 275 of the 300 cities, 25 not being described by a single user. For 244 cities the number of descriptions exceeded 5, which meant they could be kept for the game of Taboo. Similarity measures and simple string-matching techniques were employed to deduplicate the results and identify the most frequent words from this set. This resulted in 576 descriptions for 226 cities where frequent meant that more than one user provided the same characteristic. For the Taboo game we manually chose several additional salient descriptions as Taboo words, while for this task of ontology building we decided to keep only the most frequent ones as a quality assurance measure. We kept duplicates across the data set but de-duplicated the descriptions

[11] http://weblab.iit.cnr.it/kyoto/xmlgroup.iit.cnr.it/kyoto/index.html

[12] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

of each city since many cities are, for instance, a capital. This was particularly crucial for the distributional word sense disambiguation approach that considers the city as the context for individual words. To ensure the comparability of the two data sets, both contain the same instances of cities and their descriptions and all other city instances that were not described with the second method were also omitted in the first data set for this paper. This drastically reduced our data set to 322 descriptions of 62 cities, with an average of 5.65 descriptions per city.

Each city was described by a total of 12 participants, who had the option to skip a city shown to them in case they were not familiar with it. We obtained a total of 3,616 trusted judgments over six days the task was online, where trusted refers to workers with more than 70% accuracy on the test questions and an answer time exceeding ten seconds for each question. In fact, the trust level for this task was extremely high with 91% on average. The 80 participating workers were mainly based in the United States with 65% followed by Great Britain with 30% and the remaining workers came from New Zealand and Australia. We did not limit the number of descriptions that could be provided by an individual participant and some top contributors provided up to 85 descriptions.

The most frequent input data from this first task and a predetermined number of handpicked most salient properties for each city were utilized as taboo words for the remaining 226 cities. The predetermined number of taboo words per city depended on the number of descriptions provided for each city by the crowd: $>$25 descriptions = 12 taboo words, 20-25 descriptions = 10 taboo words, $<$20 descriptions = 8 taboo words. Those benchmarks were based on the assumption that more descriptions in the first task require more taboos in the game since people seem to have more associations with those cities and we wanted to keep the game challenging.

The remaining 226 cities were provided to a total of 30 users in five online sessions on the platform. This resulted in 316 games, of which 174 were successful, i.e., the city was guessed correctly in 174 cases. Those successful games were manually evaluated by 12 ontology engineers and researchers regarding their conformance to the restriction to common nouns and the rules of the game, e.g. not containing a taboo word. Two of those engineers evaluated all selected successful and compliant games in a final quality assurance task, to ensure that all games corresponded to the established quality criteria. This process reduced our data set to 73 games of 62 cities and a total number of 202 descriptions of cities. This set of 202 descriptions of the second technique and the 322 descriptions from the first crowdsourcing technique provided the input to our ontology building method.

4.2 Ontology Engineering

4.2.1 Category Extraction

In general, querying Word Reference (WR) and WordNet (WN) provided definitions for the words in the descriptions, although there were some exceptions, most of which were actually words in foreign languages. From the found words, in most cases the correct sense (the one intended by the describer) was available as a definition, as shown in Table 1. 'Hints' refer to data obtained by the game by the describer and 'taboos' are the result of the first mechanized labor-based task. By 'not available' we mean that the word is in the resource but the required sense is not. In some other cases, the describer used the word in a very complex or informal way, which was not included in our resources. This is the case of, for example, using sack to describe Sacramento.

	Available	Not Available
WN (hints)	112	6
WN (taboos)	199	1
WR (hints)	109	9
WR (taboos)	194	6

Table 1: Availability of correct senses for WordNet and Word Reference

We also measured the number of available correct categories in Word Reference depicted in Table 2. In this case the value is lower, because many glosses in Word Reference are not classified into a category. At times, the categorization in the resource is not entirely accurate, as for instance beer is classified as wine instead of alcoholic beverage. Nevertheless, in most cases the quality of the categories is surprisingly high.

	Available	Not Available
hints	77	42
taboos	132	64

Table 2: Availability of categories in Word Reference

4.2.2 Sense Disambiguation

The distributional semantics approach provided the sense in WordNet that was closely related to the pair $(word, city)$ for each hint. We used only the words of which the intended sense was in WordNet, and classified the results as *correct* or *incorrect*. For the Word Reference data we additionally retrieved the closest sense between the ones that were related with a category, since this would be the category assigned to the word. We followed the same criterion for evaluating the categories. Each of the results was evaluated by two ontology engineers, and we considered as correct the intersection of those in both evaluations. The results of this process are depicted in Table 3, where WN refers to WordNet and WR to Word Reference and only the senses available in the resource have been considered. In brackets we indicate the data set, which is 'taboos' for the mechanized labor-based approach and 'hints' for descriptions obtained from the game-based crowdsourcing technique.

As can be seen from the summary of the results in Table 3, the F-Measure or accuracy of retrieving the correct WordNet sense for both data collections almost reaches 80%. Given the highly ambiguous nature of the input data and the lack of context, we consider this a good result. For instance, boot has 7 different senses and no obvious connection to Wellington. Our algorithm identifies the correct sense of 'footwear', since the description most likely hinted at the famous 'Wellington rubber boot'.

In this approach every exact duplicate for different cities was kept. We considered the description of one city with the term architecture different from the same string for another city. And in fact our results proved this point since Agra was associated with the profession of designing works of architecture, while Berlin returned the discipline of architecture as a field. This shows that the city vector has an effect on the selection of a sense. While we noticed this fact as part of our study, the number of duplicates in our data set was not sufficient to provide a proper analysis of the impact of the city vector on the sense selection, which is definitely interesting for further experiments.

In contrast to the senses, the retrieved results of the categories depicted in Table 4 from Word Reference were considerably lower,

	Correct	Incorrect	F-Measure
WN (hints)	89	23	0.79
WN (taboos)	164	35	0.82
WR (hints)	74	35	0.68
WR (taboos)	122	72	0.63

Table 3: Sense Evaluation for DS approach in WordNet and Word Reference

since not all glosses were classified in the resource itself. Nevertheless, if categories were available, the quality and accuracy was very high. For instance, star for the city Cannes provided the category 'show business'. In order to retrieve meaningful categories, several restrictions had to be added to the algorithm. Firstly, we decided to ignore all categories that classified language usage, such as 'slang term'. Secondly, since one of our crowdsourcing restriction was use of only common nouns, we omitted all categories that referred to proper nouns of any kind. With those restrictions in place, the retrieval of categories led to an accuracy of more than 84% for both data sets.

	Correct	Incorrect	F-Measure
hints	65	12	0.84
taboos	119	13	0.90

Table 4: Category Evaluation for DS approach in Word Reference

As a class-based approach to disambiguating words, the city of the description is not taken into accounting as no difference in the sense selection could be expected. The approach queries the data in YAGO and Kyoto and returns a result irrespective of the city. Thus, the results presented in Table 5 sum to a different total than the results of the distributional approach provided in Table 3. Furthermore, instead of categories this approach considers WordNet domains (WND) for both types of data sets that are directly associated with data as they are queried in YAGO. This is based on the assumption that such domains provide an excellent basis for disambiguating our city descriptions. The results support this point since all the domains that were directly obtained on the first query were accurate categories for the input data.

One further difference between the distributional semantic and the class-based approach is the type of input data. While the former queries individual words in combination with cities, the latter first attempts to retrieve senses for noun phrases, such as red carpet, and, only if it does not retrieve any result, queries the head noun of each phrase. This head noun identification succeeded in 44 out of 48 cases of compounds in all data sets using the Stanford CoreNLP parser. Failures can be attributed to imperfect input, such as *embargo lift instead of lifted embargo to refer to Havanna or to be precise to Cuba and the U.S. trade embargo that has been recently lifted. The head noun that was queried for this example was *lift* which returned a sense related to skiing. Specific symbols equally constituted a problem for the parser since ex-empire remained unchanged and thus did not return a sense, which would have been achieved by only querying empire.

All results were rated by two experts in a separate task and only the ones that were agreed upon are presented in Table 5. The accuracy for each input depended on the sense that was provided or in case of domains on the domain label as well as its superordinate class. For a total of 27 input phrases the raters did not agree and thus those data are neither considered here nor in the taxonomy building task.

	Correct	Incorrect	F-Measure
WN (hints)	78	19	0.80
WN (taboos)	78	8	0.91
WND (hints)	5	0	1
WND (taboos)	16	0	1

Table 5: Class-Based sense disambguation results

4.2.3 Taxonomy Building

When building the hierarchical backbone of the two ontologies for the two different data sets, we utilized the disambiguated senses from the previous task. Our approach consisted of following the sense up the hierarchy for two levels and extracting all `subClassOf` relations from Kyoto. Only for the 21 WordNet domains directly associated with the first query no further disambiguation was necessary since each domain returned exactly one additional hierarchy level subordinate to the domain. For instance, the sense skyscraper_104233124 returned the domain wordnetDomain_building_industry, which is in turn narrower in meaning than the wordnetDomain_architecture. One issue we faced in this regard is the poor coverage of WordNet domains in YAGO, which is why we always performed a string matching of sense labels and WordNet domains, which returned twice as many domains as querying YAGO alone. In the final ontology, the retrieved WordNet domain hierarchy was still evaluated manually against the other hierarchies for accuracy and adequacy.

From Kyoto, we retrieved up to 32 senses on the second level of hierarchy. For instance, `victim` provided mostly `person` on the first level but then explodes on the uppermost level to 32 different types of agents and social figures. This number already excludes DOLCE concepts and senses related to any other part-of-speech type than nouns in Kyoto. Thus, the manual effort involved in deduplicating the retrieved hierarchies is rather high and it is definitely worth investigating automated methods in the future. The first and the second level of hierarchy in Kyoto are frequently identical but still relate by means of a subclass relation. Those were eliminated as well.

One step to further reduce unnecessary complications was to deduplicate hierarchies in the ontologies obtained from identical senses, since we also in this step abstract away from the city and thus this context. This means each sense is only included once across the data set and with one specific hierarchy. This step reduced the number of obtained superclass concepts from 426 to 60 for the hints and from 301 to 48 for the taboo words and phrases.

The differences in level of granularity were kept in this experiment. The following two example hierarchies for animals show this difference: `crocodile` $\sqsubseteq$ `animal_fauna` $\sqsubseteq$ `organism_being` as opposed to `dingo` $\sqsubseteq$ `mammal` $\sqsubseteq$ `animal_fauna`. While the first animal is directly mapped to `animal` and then a general concept of `organism`, the Australian representative is first mapped to `mammal` and would require one more hierarchical level to reach the same level of abstraction as the first.

4.3 Evaluation

This section describes the evaluation of our seed ontologies against each other and the Wikipedia ontologies we created for the general concept of city and for the evaluation of their semantic equivalence, their 'citiness'. First, we evaluate the correctness of the extracted Word Reference categories. Then we compare the two taxonomies with a four-layer ontology extracted from Wikipedia TOCs of cities.

This step serves to evaluate whether the crowdsourcing techniques provided semantic classes that are closely related to descriptions of the general concept city by comparing them to a manually created gold standard, our four-layer Wikipedia ontology describing a city on the city, country, continent, and region level.

4.3.1 Word Reference Categories

We evaluated the Word Reference categories by comparing them with the city ontology that we built from Wikipedia TOCs of specific cities. We performed this evaluation only over the categories that were disambiguated correctly with the distributional semantics approach, since we are interested in how far the data collected by crowdsourcing reflect a proper description of the general concept city. The categories that were not correctly classified were left out since they did not refer to the correct meaning of the descriptions obtained by crowdsourcing and consequently could not reflect on the level quality of the description of city. These categories are not organized in a taxonomy and thus only the number of semantically equivalent categories with the Wikipedia resource was analyzed.

When removing duplicates in the categories from Word Reference for the hints, we obtained a total of 29 categories. From those, 18 (62%) directly corresponded to categories in the Wikipedia taxonomy for cities, modulo clear term alignment (like Food $\equiv$ Cuisine). One other category corresponded to the Wikipedia taxonomy for regions. From the remaining 10 categories which did not have clear matches in Wikipedia, 8 were subconcepts of a category in the Wikipedia taxonomy (for example Mammal), while 2 were not present. For the taboos the total was 31 categories, from which 15 were in the city ontology, 6 on the other layers (region or country), 5 were subconcepts of categories in the city ontology, and 5 were not present. In both cases, 9 of the 12 first-level categories in the Wikipedia taxonomy were represented, either by themselves (in 6 cases) or by one of their subcategories.

Since we apply two different crowdsourcing techniques in this paper, it is also interesting to evaluate any differences in the resulting data sets of those techniques. From both datasets a total of 60 categories were obtained of which 60% are identical. Half of all non-corresponding categories for each data set represented specific concepts that would occur on a lower level of hierarchy and be subsumed by corresponding concepts, such as Eastern Religion is a subcategory of religion. The other half are categories that are very general, such as Sport, and are thus likely to occur on the highest level of hierarchy.

4.3.2 Ontology Alignment

The first evaluation step of Wikipedia was similar to the evaluation of the Word Reference categories in that we only considered semantically equivalent categories. For instance, 'meteorology' ' and 'climate' would be considered semantically equivalent categories, while 'snow' clearly is more specific. In addition, we also consider the level of hierarchy on which the categories co-occur and whether those correspond. This step serves to evaluate whether the semi-automatically built resources based on data from crowdsourcing approaches offers a similar level of detail as the manually created resource for describing the general concept of city.

The results of this evaluation are quantified in Table 6, which only includes correct hierarchy extractions. Thus, the number of deduplicated results was further reduced from the disambiguation step to 60 taboos and 48 hints with two correct levels of hierarchy. 'N' refers to no correspondence in Wikipedia, 'L0' to the most specific hierarchy level, 'L1' to the intermediate level, and 'L2' to the most general meaning of the Wikipedia city ontology.

	N	N (%)	L0	L1	L2
taboos	15	25%	11	19	15
hints	30	62%	7	7	4

Table 6: Comparing Resulting Ontology with Gold Standard Ontology from Wikipedia TOCs

When comparing the two different data sets, the seed ontology deriving from the mechanized labor-based data set, the taboos, shows a larger variety of types of categories that correspond to Wikipedia categories of cities with 45 in total. The categories and subclass relations obtained from the game-based crowdsourcing task correspond in 18 cases to Wikipedia elements on different levels of hierarchy. One reason for this lower coverage of hint categories in the gold standard ontology is the fact that they relate to more abstract Kyoto concepts. For instance, Kyoto#activity for cooperation is more general than categories that could be found in TOCs, such as Politics or Governance. Furthermore, WordNet Domains proved to show a high correlation with our gold standard ontology. Thus, the fact that more taboos directly correspond to WordNet domains is a second reason for the stronger correlation of the Wikipedia TOC and the Taboo seed ontology. It is interesting that the distribution of both types of ontologies is rather even across the three levels of hierarchy.

A direct comparison of the two data sets with each other, however, indicated a stronger variation of the results of the two crowdsourcing approaches. We compared the first level of hierarchy to each other, that is, the Kyoto concepts and WordNet domains either directly associated with the sense we obtained from the disambiguation techniques or associated with it on the next level of hierarchy. Deduplicating those concepts resulted in 46 concepts for the taboo words obtained from the mechanized labor-based approach and 31 concepts for the game-based approach. This comes as a little surprise since the first data set is larger than the second one. In total, 49% of the obtained 77 concepts are identical across both data sets on the first level of hierarchy. We classified the non-identical concepts into concepts and subconcepts. Concepts would likely be found on the most general level of hierarchy, while subconcepts would be found on a lower level, such as soccer as a subconcept of sport. The distribution of this classification is identical across the two data sets with 56% subconcepts and 44% concepts likely to be found on the highest level of hierarchy in a city ontology. Nevertheless, the type of subconcepts that can be found in the results obtained from the taboo dataset shows a slightly higher level of granularity. For instance, it contains concepts such as mountaineering that could only be found as sport in the hint data set.

5 DISCUSSION

The two distinct crowdsourcing techniques utilized in this approach both proved to be a valid and valuable source of input for the ontology engineering process. We found that the time needed to obtain data from the mechanized labor-based approach strongly exceeded the time for obtaining the same amount of data in a game-based approach. The former was running for more than a working week, while the latter achieved the same in just five sessions each a bit more than an hour. The incentive to participate in a game of Taboo seemed much higher. In fact, participants asked for the permission to

play again after the first session, and four of the thirty participants joined a second session.

The nature of the game required the creation of Taboo words. When we investigated descriptions of cities online, we found very little useful data. Thus, we decided to crowdsource this first step. The fact that the results of this first method, the descriptions of the cities we called taboo words, are then used as input for the game-based approach is not ideal. While it reduces the overlap of the two data sets, it also creates an unwanted bias. Without this step the overlap of the data set might be much stronger and thus the resulting seed ontology much more similar than in this mutually exclusive way we propose. On the other hand, our major goal was not the comparison of the two techniques with each other. We were rather interested in the degree of overlapping concepts and hierarchical relations obtained from each data set and our manually created gold standard ontology from Wikipedia data.

The two approaches that we implemented for the extraction of senses from two different resources are accurate in that they have the correct sense for most of the words in the city descriptions. Word Reference is convenient because it already provides a classification of the senses in the form of a general domain label, however, there are many senses for which that classification is missing, which results in a great loss of useful data. A resource like this one but with a complete classification would be ideal for our purposes. For WordNet, the labeling feature is not immediately available, so more complex techniques need to be implemented to retrieve a classification of our data. In both cases there were many other senses available, so some kind of sense disambiguation is necessary.

The distributional approach returned impressively accurate results in some cases (for example, the pair hint-city *(star, Cannes)* matches with the definition *a prominent actor, singer, or the like, esp. one who plays the leading role in a performance*, which has the category *'Show Business'*). However, there are also some issues that should be resolved in future work. For example, using a simple comparison with the vector obtained from the average of the hint and the city causes that if any of the definitions includes the hint as a word, it will rank very high. Consequences of this are, for example, that *(go) jump in the lake, (used as an exclamation of dismissal or impatience)* ranks higher for *(lake, Lausanne)* than the actual sense of *boy of water*. This should be resolved by either exploring different ways of composing words in vectors, or by using a more complex combination, for example giving more weight to the city than to the hint when combining them.

The class-based approach is strongly biased by the static resources it utilized and thus, less attractive than the more dynamic distributional semantic approach. Moreover, the number of senses obtained from WordNet on only two levels of hierarchy can be very overwhelming. For highly ambiguous terms with several senses and upon querying YAGO, Kyoto, and the WordNet domain ontology, the number of retrieved categories and senses for an input word quickly reached more than 400. With the automated frequency-based and Word Reference category-weighted disambiguation we still obtained comparable results to the distributional approach. WordNet domains proved to be a highly reliable and disambiguating part of this approach as also found and suggested by [2]. First of all, their occurrence in the sense repository of a query shows that the queried word is very close to this high-level ontology associated with WordNet senses. Thus the queried word itself can be assumed to be more general in meaning than those in the same category that are not directly associated with a domain. Secondly, the domain proved to be highly accurate for the kind of disambiguation we needed. Finally, it associates the category of the description with exactly one further superordinate category and thus makes it more comparable to our Wikipedia ontology.

The strong difference of level of granularity in WordNet unfortunately propagates to the ontology concepts that are retrieved from Kyoto. For instance, DOLCE:endurant and Kyoto#organization are returned on the same level of hierarchy when querying the resource for the input word company. While for some approaches highly abstract concepts, such as DOLCE:endurant, are very useful for others, such as ours, they are too high-level. This also applies to the number of concepts and relations that are retrieved for each sense in WordNet. When considering two levels of hierarchy for an already disambiguated sense, the number of concepts can easily reach 14. Since a full evaluation of all relations and concepts retrieved goes beyond the scope of this paper, we decided to focus on more concrete levels of granularity, i.e., disregard upper level ontologies such as DOLCE for this approach, and only subclass relations. Kyoto returns a number of non-hierarchical relations, however, their evaluation goes beyond the scope of this paper. One alternative approach to handling this wealth of information might be a classification of concepts and relations based on machine learning, as implemented and proposed by [16]. Alternatively, crowdsourcing could also be applied to this step.

The comparison of the categories retrieved from Word Reference with the ones in the taxonomy built from Wikipedia shows that in most cases the labels match. Some of the ones that do not match directly are subcategories of Wikipedia labels, which seems to show that creating an organized taxonomy using the Word Reference taxonomies as seeds would be a promising direction. In other cases, the categories match with others in the Wikipedia taxonomies for *country* or for *region*, this should be taken into account when using this kind of approaches, since players tend to describe instances not only with their properties but also with properties from their parent categories.

A comparison of the two data sets resulted in a surprisingly high level of overlap of semantic categories given that they differ in size and the data of the first mechanized labor-based approach cannot be provided as data of the second approach since they represent the taboo words in the game. Thus, we concluded that the results of the two approaches are comparable, even if the taboo words lead to a higher level of granularity in the conceptualization of city descriptions. We found in the word sense disambiguation results that more specific descriptions provided more interesting hierarchies for the characterization of a city. For instance, *food* quickly became *substance* up the hierarchical classification ladder while *kiwi* mapped to *vine* and then *plant/flora* and *barbecue* to *nutriment* and then *food*.

6 CONCLUSION AND FUTURE WORK

This paper addresses ways to benefit from the diversity of people in the world by utilizing two distinct crowdsourcing techniques to gather data for ontology building. It further utilizes a third crowdsourcing platform, namely Wikipedia, to build an evaluation resource for the results obtained from the first two. The two word sense disambiguation methods used herein provide promising results for automating the step of concept formation and categorization of city descriptions. We also semi-automatically built a hierarchical backbone to the retrieved categories in order to facilitate their comparison with a manually created ontology for city descriptions based on the crowdsourcing platform Wikipedia. The results thereof show that the mechanized labor-based technique returns more specific categoriza-

tions and a more refined level of hierarchy. Nevertheless, the game-based approach returns very promising results and we believe that it is a more interesting way for the crowd to engage in a knowledge production tasks.

There are many directions of research that are derived naturally from this work. Some of the technical ones were pointed out in the previous section, while here we discuss more general questions that should be addressed.

First, the relation extraction part should be developed. Although there exist approaches that tackle this problem in particular, both with automatic and crowdsourcing techniques, their adequacy to our problem should be analyzed, since they are not particularly designed to identify relations between a concept and its attributes. The classification part, for which we provide automated methods here, could also be crowdsourced.

Second, the choice of using an implicit crowdsourcing method could be justified empirically by comparing it with an explicit technique for the same task, something that we kept for future experiments for now. To this end, we should perform a third experiment in which users are asked directly to name properties of the general concept city.

Finally, the true diversity of the obtained domain knowledge could be further explored by building clusters based on common traits of participants, such as country of origin or age, and comparing the results of individual clusters to each other. This also provides a large number of individualized domain ontologies that are highly comparable and might provide some insights into the diversity of knowledge production. Furthermore, conducting comparable experiments with non-English speaking crowds and comparing the results obtained from multilingual corpora obtained from crowdsourcing could be an interesting direction for further research.

ACKNOWLEDGEMENTS

This research has been funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 567652 /ESSENCE: Evolution of Shared Semantics in Computational Environments/.

References

[1] P. Basile, A. Caputo, G. Semeraro, and F. Narducci, 'Uniba: Exploiting a distributional semantic model for disambiguating and linking entities in tweets', *CEUR Workshop Proceedings*, **1395**, 62–63, (2015).

[2] Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta, 'Revising the wordnet domains hierarchy: semantics, coverage and balancing', in *Proceedings of the Workshop on Multilingual Linguistic Ressources*, pp. 101–108. Association for Computational Linguistics, (2004).

[3] Hafedh Chourabi, Taewoo Nam, Shawn Walker, J Ramon Gil-Garcia, Sehl Mellouli, Karine Nahon, Theresa A Pardo, and Hans Jochen Scholl, 'Understanding smart cities: An integrative framework', in *Proceedings of the 45th Hawaii International Conference on System Science (HICSS)*, pp. 2289–2297. IEEE, (2012).

[4] Jia Deng, Jonathan Krause, Michael Stark, and Li Fei-Fei, 'Leveraging the Wisdom of the Crowd for Fine-Grained Recognition', *Ieee Transactions on Pattern Analysis and Machine Intelligence*, **38**(4), 666–676, (April 2016).

[5] Anca Dumitrache, Lora Aroyo, Chris Welty, Robert-Jan Sips, and Anthony Levas, "'dr. detective": Combining gamication techniques and crowdsourcing to create a gold standard in medical text', in *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web - Volume 1030*, CrowdSem'13, pp. 16–31, (2013).

[6] Kai Eckert, Mathias Niepert, Christof Niemann, Cameron Buckner, Colin Allen, and Heiner Stuckenschmidt, 'Crowdsourcing the assembly of concept hierarchies', in *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, pp. 139–148, (2010).

[7] Florian Hanika, Gerhard Wohlgenannt, and Marta Sabou, *The uComp Protégé Plugin: Crowdsourcing Enabled Ontology Engineering*, 181–196, Springer International Publishing, 2014.

[8] Rubén Izquierdo Beviá, Armando Suárez Cueto, German Rigau Claramunt, et al., 'Word vs. class-based word sense disambiguation', *Journal of Artificial Intelligence Research*, **54**, 83–122, (2015).

[9] Lili Jiang, Yafang Wang, Johannes Hoffart, and Gerhard Weikum, 'Crowdsourced entity markup', in *Proceedings of the 1st International Conference on Crowdsourcing the Semantic Web-Volume 1030*, pp. 59–68, (2013).

[10] Vladimir I Levenshtein, 'Binary codes capable of correcting deletions, insertions, and reversals', in *Soviet physics doklady*, volume 10, pp. 707–710, (1966).

[11] Miguel Angel Luengo-Oroz, 'Crowdsourcing Malaria Parasite Quantification: An Online Game for Analyzing Images of Infected Thick Blood Smears', *Journal of Medical Internet Research*, **14**(6), e167, (2012).

[12] Alexander Maedche, *Ontology learning for the semantic web*, volume 665, Springer Science & Business Media, 2012.

[13] Emir Muñoz, Aidan Hogan, and Alessandra Mileo, 'Triplifying wikipedia's tables', in *Proceedings of the First International Conference on Linked Data for Information Extraction - Volume 1057*, LD4IE'13, pp. 26–37, (2013).

[14] P. Nasirifard, S. Grzonkowski, and V. Peristeras, 'Ontopair: Towards a collaborative game for building owl-based ontologies', volume 351, pp. 94–108, (2008).

[15] Natalya F Noy, Jonathan Mortensen, Mark A Musen, and Paul R Alexander, 'Mechanical turk as an ontology engineer?: using microtasks as a component of an ontology-engineering workflow', in *Proceedings of the 5th Annual ACM Web Science Conference*, pp. 262–271. ACM, (2013).

[16] Alina Petrova, Yue Ma, George Tsatsaronis, Maria Kissa, Felix Distel, Franz Baader, and Michael Schroeder, 'Formalizing biomedical concepts from textual definitions', *Journal of biomedical semantics*, **6**(1), 1, (2015).

[17] Cristina Sarasua, Elena Simperl, and Natalya F. Noy, *CrowdMap: Crowdsourcing Ontology Alignment with Microtasks*, 525–541, Springer Berlin Heidelberg, 2012.

[18] Neil Savage, 'Gaining Wisdom from Crowds', *Communications of the Acm*, **55**(3), 13–15, (March 2012).

[19] Katharina Siorpaes and Martin Hepp, 'Ontogame: Towards overcoming the incentive bottleneck in ontology building', in *Proceedings of the 2007 OTM Confederated International Conference on On the Move to Meaningful Internet Systems - Volume Part II*, OTM'07, pp. 1222–1232. Springer-Verlag, (2007).

[20] Stefan Thaler, Elena Simperl, and Katharina Siorpaes, 'SpotTheLink: A game for ontology alignment', in *Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI)*, volume P-182, pp. 246–253, (2011).

[21] Daniele Vannella, David Jurgens, Daniele Scarfini, Domenico Toscani, and Roberto Navigli, 'Validating and extending semantic knowledge bases using video games with a purpose.', in *ACL (1)*, pp. 1294–1304, (2014).

[22] Johanna Völker, Daniel Fleischhacker, and Heiner Stuckenschmidt, 'Automatic acquisition of class disjointness', *Web Semantics: Science, Services and Agents on the World Wide Web*, **35**, 124–139, (2015).

[23] Luis von Ahn, 'Duolingo: learn a language for free while helping to translate the web', in *Proceedings of the 2013 international conference on Intelligent user interfaces*, pp. 1–2. ACM, (2013).

[24] Luis Von Ahn and Laura Dabbish, 'Designing games with a purpose', *Communications of the ACM*, **51**(8), 58–67, (2008).

[25] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum, 'recaptcha: Human-based character recognition via web security measures', *Science*, **321**(5895), 1465–1468, (2008).

[26] Wilson Wong, Wei Liu, and Mohammed Bennamoun, 'Ontology learning from text: A look back and into the future', *ACM Computing Surveys (CSUR)*, **44**(4), 20, (2012).

[27] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung, 'A survey of crowdsourcing systems', in *Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT)*

and 2011 IEEE Third International Conference on Social Computing (SocialCom), pp. 766–773, (2011).

[28] Maayan Zhitomirsky-Geffet, Eden S Erez, and Bar-Ilan Judit, 'Toward multiviewpoint ontology construction by collaboration of non-experts and crowdsourcing: The case of the effect of diet on health', *Journal of the Association for Information Science and Technology*, (2016). Online version.

[29] James Y Zou, Kamalika Chaudhuri, and Adam Tauman Kalai, 'Crowdsourcing Feature Discovery via Adaptively Chosen Comparisons', in *Proceedings of the Conference on Human Computation and Crowdsourcing (HCOMP) 2015*, pp. 198–212, (2015).

A Methodology to Take Account of Diversity in Collective Adaptive Systems

Heather S. Packer and **Luc Moreau**[1]

Abstract. Collective Adaptive Systems (CASs) are comprised of a heterogeneous set of components often developed in a distributed manner. Their users are diverse with respect to their profiles, preferences, interests and goals, and hence, have different requirements. We propose a typology for the diversity of these components, users, and their requirements. We then present a methodology which provides steps to integrate features that record diversity to support accountability. The foundation of accountability is provided by provenance data, and a CAS vocabulary, these knowledge representation languages provide the core vocabulary that can be exploited by agents and services.

1 INTRODUCTION

Collective Adaptive Systems (CAS) are heterogeneous collections of autonomous task-oriented systems which contribute to a common goal, thus forming a collective system. The heterogeneous collections of systems means that CASs have diverse requirements because they have multiple stakeholders with different motivations, methods, tooling, profiles, and goals. There is also diversity in the way that each system processes the same data because of different perspectives and interpretations. It can be hard for participants to trust CASs because they are comprised of many systems which are often black boxes and strangers may be required to collaborate. Accountability in CASs enables its participants to build trust in the system and make informed decisions about other participants. In order to support the analysis of diversity in a CAS, it is important that their components adopt a standard model to express their properties and goals. CASs can also support diversity through the way that information is presented to different stakeholders, because they may require different types of information. For example, administrators might require statistics about usage, whereas a participant might require information about another participant to complete a task.

In CASs that rely on participants collaborating, reputation ratings and reviews are often used and can affect how members select or trust input from others. The algorithms and how participant use rating systems can vary greatly from CAS to CAS, therefore it can be hard to understand what ratings actual represent and mean to the community. Thus, it is important for a CAS's participants to understand how ratings are used and generated so that they can evaluate how to improve their rating or how much it should influence their selection process. It is also important that members can understand the potential utility of selecting others because strategically selecting members can maintain or improve their ratings. For example, some members may have high expectations and preserve high ratings for truly exceptional services, while others give high ratings more freely. Our focus is to provide end-users with an accountable CAS that instills trust in it and its diverse community.

Provenance is increasingly used for making systems accountable through exposing how information flows through a system and helping users to decide whether the resulting information can be trusted. The recent standard PROV [29] of the World Wide Web Consortium defines provenance as "a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing." PROV is a conceptual data model (PROV-DM [29]), which can be mapped and serialised to different technologies.

In order to provide accountability detailing diversity in CASs, we identify a typology for diversity in CASs and present a methodology to accommodate it. We use PROV which takes a Linked Data approach and benefits from its principles, namely through the use of Uniform Resource Identifiers (URIs), and the use of URIs to denote types for identifying resources. Diversity is supported through URIs, where individual and collectives are resources with properties, and those elements and properties are typed. In this paper we contribute:

1. A typology identifying diversity in CASs;
2. The CAS Vocabulary, which provides types for individuals and collectives;
3. A methodology for authoring provenance in CASs;
4. An approach that allows for diversity in an accountable way.

The rest of the paper as organised as follows. Section 2 describes our typology for diversity in CASs. In Section 3, we present the architecture of the platform to which we add accountability to diversity. Then, in Section 4 we describe an example application using the previously presented architecture. In Section 5 we detail our methodology. Then, in Section 6 we introduce the CAS vocabulary. In Section 7 we describe the diversity in applications. Following that in Section 8 we present how we use provenance and the CAS vocabulary to represent collectives and agents with different roles. In Section 9, we describe how queries can support different a range of requirements. Then in Section 10 we describe the reputation system and how we use the provenance data to describe diversity and present that data to a diverse set of end users. In Section 11, we summarised the features presented in the paper that support the diversity identified in Section 2 and discuss privacy and accountability. In section 12 we present related work to provenance and accountability. Finally, in Section 13 we conclude.

2 A TYPOLOGY OF DIVERSITY IN CASs

CAS are inherently diverse due to their human peers, components, stakeholders and goals. In order to support this diversity, we first

[1] University of Southampton, UK, email: {hp3, l.moreau}@ecs.soton.ac.uk

identify possible diversity in CASs:

1. Diversity in participants;
   (a) Human participants:
      i. The members of a CASs aim to achieve a common goal. However, each person has their own attributes, preferences, and perspectives.
      ii. People may opt to form a collective, where they formally aim to achieve a collective goal regardless of their differences.
      iii. People may be placed into a collective with or without their knowledge by a CAS, based on certain attributes which may including their actions or roles within a CAS.
      iv. The developers and designers of the CAS have a different perspective and different goals to the users of a CAS. They may consume different types of data to the end-users.
   (b) System components in a CAS have different responsibilities and roles within a system. They can be developed and hosted on different stacks and servers.
   (c) There are other types of participants, such as hardware agents using the CASs which may or may not aline with the goals of a CAS.
2. Diversity in interest:
   (a) While the members of a CAS work together to achieve a common goal, they can desire different outcomes based on their role and perspective. In a ride sharing example, one user is a driver and the other is a commuter, the driver main aim is to reduce the cost of travel, while the commuter requires transport.
   (b) People may require different information from the CAS. For example, some require information to support decisions or analyse the CAS.
   (c) People may also want information to be presented in different ways.
3. Diversity in roles and involvement in activities. While a community that uses a CAS might have common goals, the members may play different roles to achieve those goals. There is also diversity in the roles of data ownership, data stewardship, and data attribution.

Furthermore, these facets of diversity may change over time. This temporal dimension may affect the algorithm or components used within the CAS, interests may evolve over time, or the role of a CAS might change. This evolution may be unforeseen during design time, and thus, the design should cater for these evolving facets.

3 ARCHITECTURAL OVERVIEW OF AN ACCOUNTABLE CAS

In this section, we present the SmartSociety platform to situate how we provision for the accountability of diversity in the rest of the paper. The platform supports multiple CAS applications. Concretely, the core components of the platform are:

Peer Manager - This component manages the profiles of the platforms end-users. It is also an authentication service.
Application - An application consists of a group of components working together to support a common goal.
Component - Components in an application serve different purposes and may be developed by different developers.
Orchestration Manager - Handles the sequence of processes run by the components in an application.
Mobile Application - Mobile applications can be developed to allow end-users to interact with an application via the REST API.
Reputation Service - The reputation service manages feedback reports and generates reputation ratings for end-users.
Provenance Service - Stores provenance documents generated by the mobile applications, applications, reputation service, and orchestration manager.

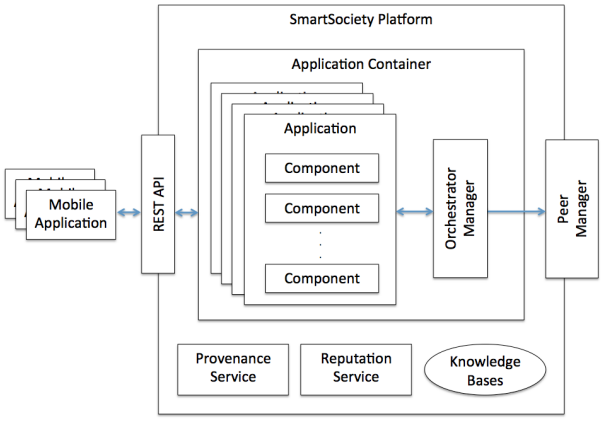


Figure 1. An overview of the architecture

4 RIDE SHARE

SmartShare is a car pooling application that allows drivers and commuters to offer and request rides. Ride offers and requests include details about required travels, timing, locations, capacity, prices, and other details relevant to car sharing. Specifically, this application is comprised of three core components, a Mobile Application, Orchestrator Manager and Reputation Service (see Figure 2). The application's orchestrator requests for a set of potential rides, which consists of a driver and commuters, from the Matcher. These potential rides are then agreed or rejected by its participants, this is handled by the Negotiator. Once a ride has been fulfilled, the drivers and commuters can leave each other feedback. This application is designed to be used in a diverse community, where its users range from office workers to tourists. The size and diversity of the community enables the application to be populated with many ride options, however, this diversity can cause problems in the application's adoption. Potential participants might be concerned with their safety with sharing rides with strangers from a diverse background.

SmartShare is provenance-enabled, capturing the provenance of any user decision, matching or rating managed by the system. The components in the architecture that record provenance are shaded in Figure 2. Specifically, the SmartShare application captures 10 processes that occur when:

1. A user logs into the mobile application;
2. A user changes a page on the mobile application;

3. The mobile application requests a resource from another service;
4. The mobile application submits a ride request to the orchestrator;
5. A composition of a ride is made by the orchestrator;
6. A ride is agreed on;
7. A ride is disagreed on;
8. A ride is agreed on all by all parties involved;
9. A reputation is generated;
10. A request is made via the reputation's API.

The provenance records a user's actions and how outcomes are generated, such the classification of a star or reputation ratings. The purpose of capturing provenance in SmartShare is to make the application accountable, in particular, by providing explanations about all decisions made. Its is required to be transparent and accountable to both the developers, and its end-users.

5 A METHODOLOGY FOR ACCOUNTABILITY IN CASs

In this section, we present our methodology which provide steps to integrate features that record diversity to support accountability. The steps in the methodology are iterated over to provide refinement through observations and changing requirements. The individual steps in this methodology need to be performed in a domain-specific way for each individual CAS. At a general level, our methodology consists of the following steps (see Figure 3):

1. CAS Vocabulary Development - Build a vocabulary to define types for agents, entities, and activities specific to a CAS. This step defines the conceptual space within which the system will operate, and specifies what processes fall within the boundaries of the system.
2. Component Design and Implementation - Design the interaction model(s) underlying the social computations that should be supported by the CAS. In this step, the protocols that will govern interactions are specified in terms of communication between interacting peers, the control flow of the collective coordination procedure, data access and synchronisation through shared state.
   During the second iteration of this step, the logging of the values for the variables defined in the template are generated.
3. Design Provenance Templates - Map the vocabulary and design the provenance that the system will capture. Provenance can capture the creation, modification and use of entities within the CAS. In order to model the provenance, we use PROV-Template[23], which is a declarative approach where the design of the provenance's semantics are separated from the logging of values recorded in the provenance, in templates and bindings, respectively. Provenance documents are generated by an expansion algorithm, which combines a template with a set of bindings. The granularity of the semantics captured in the provenance models may be modified through iterations of the methodology, to support different stakeholders requirements.
4. Define Requirements for Analysis and Social Computations - Define querying and summarisation functionalities for different stakeholders. These will produce the analysis facilities the system provides to human and machine peers for its analysis, and have to be adapted to the needs of the stakeholders involved, as well as to their interpretations (e.g. summaries for the platform operator might be different than for end users). The results from the queries can support approaches to express the information in provenance to end-users. In Figure 3, we show that the queries generated by this step use the PROV documents generated by PROV-Template's expansion algorithm, and the results are used to present information from the queries in different ways.

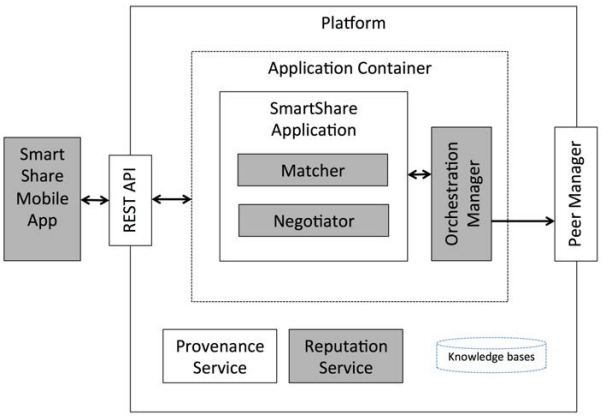


Figure 2. SmartShare Architecture

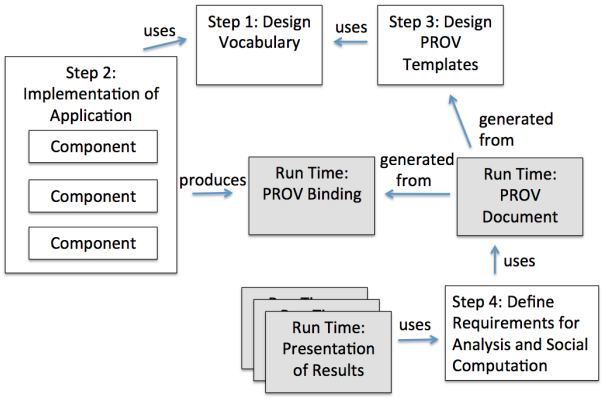


Figure 3. General Methodology

6 DIVERSITY IN VOCABULARY

In order for CASs to allow for diversity in their accounting, we require a way to differentiate between different facets of diversity identified in Section 2. Hence, in this section, we provide a vocabulary that defines types that can be used to differentiate between these facets. The diversity in a CAS may differ depending on its purpose and participants, therefore, we have designed an upper level vocabulary, which is designed to be extended to support CASs.

The core CAS vocabulary defines a hierarchy of sub-types branching from three key elements, agents, entities, and activities (see Figure 4). Specifically, the vocabulary focuses on describing three components: (1) agents within CASs, including users, peers, and collectives; (2) activities; and, (3) entities describing: outcomes of activities; and attributes of agents including preferences, capabilities, and goals.

Concretely, the vocabulary supports diversity:

1. In participants by defining (i) **cas:Peer** for CASs components

(ii) **cas:User** of the CASs (iii) **cas:Agent** are non-human agents that use CASs and (iv) **cas:Collective** that can be composed of **cas:Peer**, **cas:User** and **cas:Agent**.
2. In interests by defining **cas:Interest**.
3. In roles by defining **cas:Role** and **cas:Capability**.

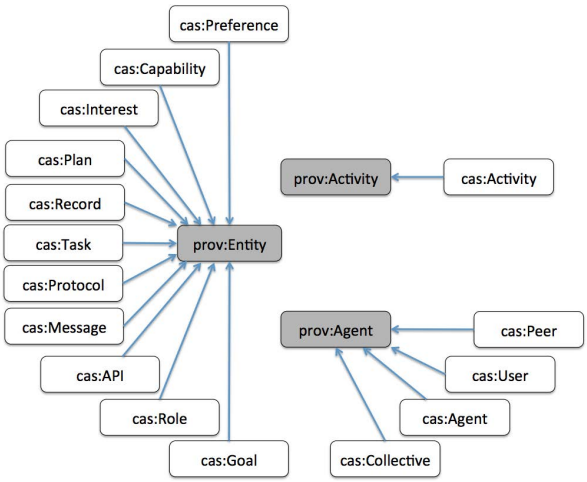


Figure 4. CAS Vocabulary

In the ride share example, the **cas:Peer** can be used to describe the orchestration manger, matcher, negotiator and reputation service. These programs play a distinct role with in the system, hence we can extend the vocabulary to include them using the following terms **cas:OrchestratorManager**, **cas:MatcherManager**, **cas:NegotiatorManager** and **cas:ReputationService**. The users play two distinct roles, driver and rider, we use the **cas:Role** to define **cas:Driver** and **cas:Commuter** (see Figure 5).

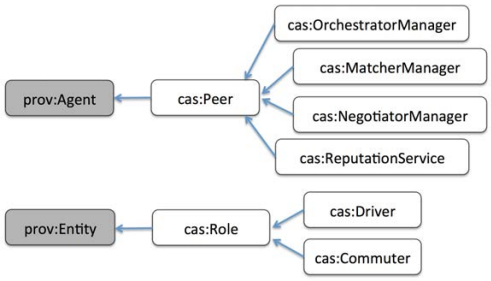


Figure 5. CAS Vocabulary Extension of Roles and Peer

7 DIVERSITY IN APPLICATION

In order to account for diversity in an application its resources are required to be described with URIs and typed with the CAS vocabulary. The URIs provide links to resources that were created, modified and used in the system, and provide context about the state of **cas:User**, **cas:Agent**, **cas:Peer** and **cas:Collective**.

An application's purpose may change during its use, therefore they need to allow for new diverse facets to be supported. For example, this supports the merging of two CASs or new types of participants to be added to the application. These changes will be required to be reflected in the application's vocabulary.

The SmartSociety platform, presented in Section 3, caters for a wide range of applications. The applications are contained in the Application Container, where each application's components are managed by its own orchestrator. The applications can interact with the reputation service and store provenance documents in the provenance service.

8 DIVERSITY IN PROVENANCE

Provenance templates are used to describe patterns to be captured by a system. In order to model the diversity, we describe how the features of PROV and the CAS Vocabulary can be used. PROV is a recent set of recommendations of the W3C for representing provenance on the web (see Figure 6). PROV is a conceptual data model (PROV-DM [29]), which can be mapped and serialized to different technologies. There is an OWL2 ontology for PROV (PROV-O [20]), allowing mapping to RDF, an XML schema for provenance [15], and a textual representation for PROV (PROV-N [30]). Provenance templates allow for diverse levels of logging.

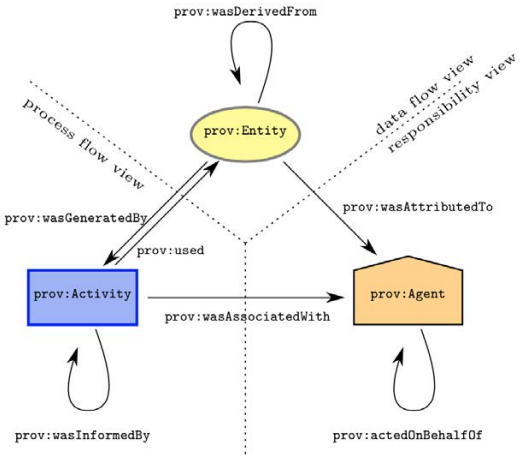


Figure 6. Three Different Views of the Core of PROV. The figure adopts the PROV layout conventions: an entity is represented by a yellow ellipsis, an activity by a blue rectangle, and an agent by an orange pentagon. We note here that the diagram is a "class diagram" illustrating the classes that occur as domain and range of properties. Taken from[27].

A user's diversity can be expressing using entities that are attributed to a prov:Agent. Specifically, **cas:Profile**, **cas:Preference**, **cas:Capability** can be used to type these entities (see Figure 7).

Collectives can be formed using **prov:Agent**s of type **cas:Collective** (see Figure 8). Collectives can also be organised into groups by their attributes. For example, Figure 9 shows an example where Alice and Bob are in a collective based on their capability of being able to drive.

A component or participant may play different roles within a CAS. A role can be express using the **cas:Role** and its type with **prov:type**. For example, Figures 10, 11 and 12 shows specialisations of a user with the roles of Driver, Commuter, and as a Commuter in a collective, respectively.

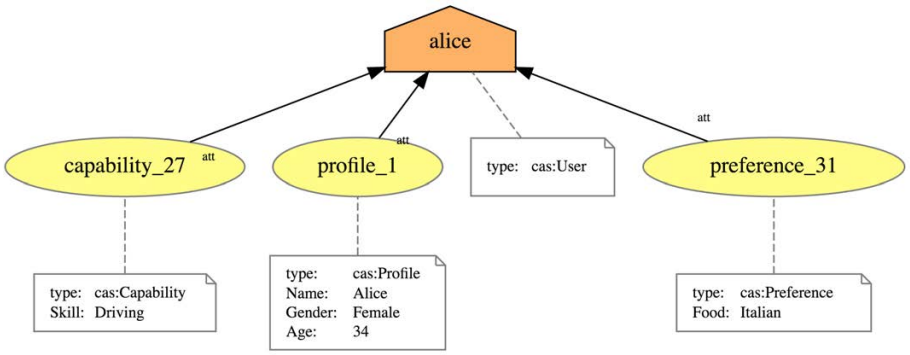


Figure 7. Alice's diversity show in her profile, preferences and capabilities

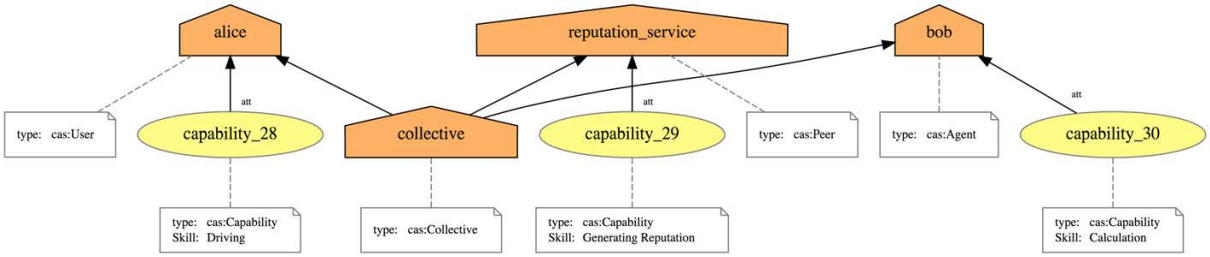


Figure 8. A collective with three different types of agent

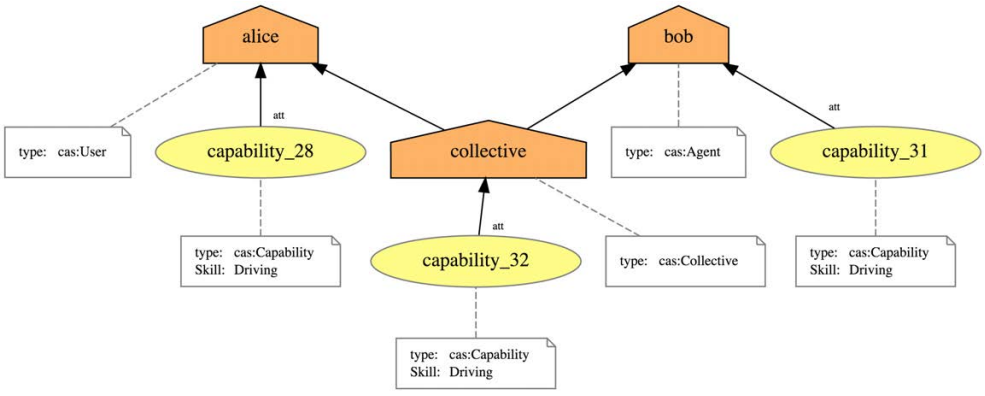


Figure 9. An ad-hoc collective connected by their capability to drive

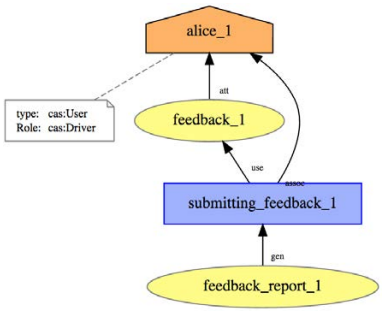


Figure 10. Alice in the role driver

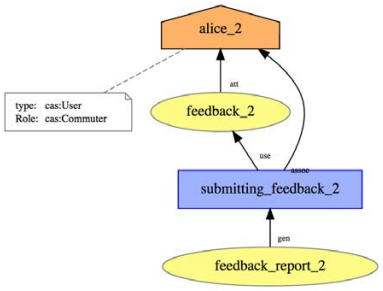


Figure 11. Alice in the role commuter

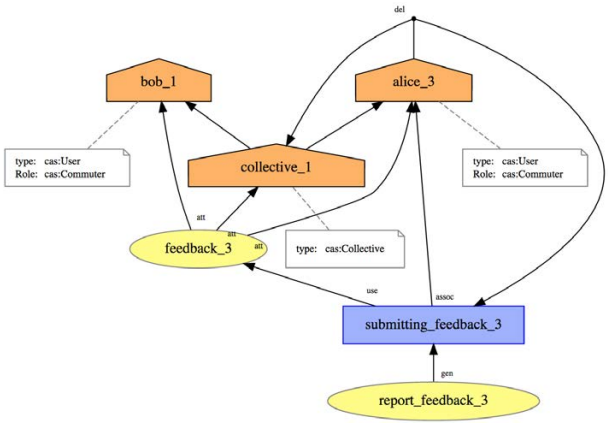


Figure 12. Alice in the role commuter submitting a feedback report, on behalf of a collective whose members were all connected to the submitted feedback report.

In Figure 13, the specialisations of Alice are linked to an Alice's generalisation, the model also captures state changes of Alice using derivations. Modelling an agent in this way enables them to act in clearly defined roles, and using PROV to define generalisations of an agent provides a clear hierarchical structure.

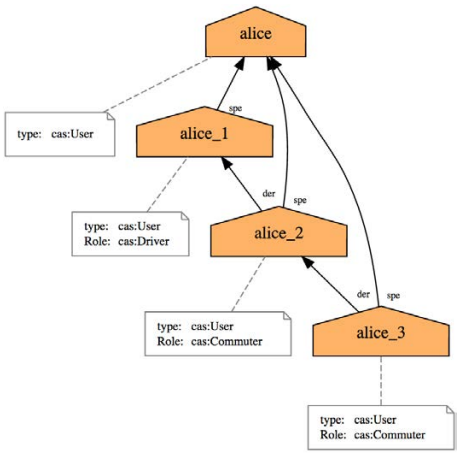


Figure 13. Provenance of linking of the specialisations of Alice to her generalisation

9 DIVERSITY IN QUERIES

Provenance graphs can be queried to support a diverse range of interests from:

End-users - Who may require specific information about themselves, others in their collective, or how a certain entity was generated so they can make informed decisions. End-users may find it easier to trust or act in transactions or collectives, if they know that they are collaborating with like minded people or in a collective with a particular range of skills.

Developers - Who may require statistics that provide them information about how end-users use the CAS, so that they can make improvements to popular features or stop supporting those that are unused.

Administrators - Who may require summative information about usage statistics and verify that entities are created following the CASs protocols.

Software Components - May require an aggregation to provide an input to a function so that the CAS can adapt to how it is being used.

Provenance expressed an RDF syntax can be queried with SPARQL. For example, the following SPARQL query returns all the users that are associated with collectives:

```
SELECT ?user WHERE {
  ?user prov:hadMember ?collective
  ?collective prov:type cas:Collective
}
```

The following SPARQL query returns all collectives that contain one or more drivers:

```
SELECT ?collective WHERE {
  ?user prov:hadMember ?collective
  ?collective prov:type cas:Collective
  ?user cas:Role driver
}
```

These queries can be used to support administrators or software components. Using this linked data approach supports different perspectives on the data.

10 DIVERSITY IN EXPLANATIONS

It is often necessary for CASs and their users to make decisions based on provenance data. Therefore, it is important to communicate that data accessibly to both machines to support adaptability in their algorithms and human users to facilitate transparency and accountability. Provenance data is machine readable, the largest challenge is communicating data with humans in a diverse way. It is possible to use graphical approaches for this, but, in many cases, it is more natural or appropriate to communicate this data either textually or verbally. We identified in [32] that the largest challenges related to utilising the structure and elements in provenance data were:

1. Identifying and presenting interesting facts to support a particular use case. For example, a user can view an explanation about someone else to aid in their decision to share a ride with them, a narrative can provide evidence of reliability from the provenance based on features such as the number of feedback reports left by a user, or the number of times this user has interacted with the application;
2. How to describe PROV elements without referring to long complicated URIs, while providing meaningful explanations. Long URIs break up the fluidity of sentences making them hard for humans to parse.

In order to mitigate these issues, we have created an approach to convert provenance data into a linear, textual form. In more detail, the steps of this narrative approach are:

1. Identify information is relevant to the target audiences. This step should involve an exploratory study involving potential users from a diverse set of backgrounds, profiles, and roles, which investigates the information users require to support their decision making;
2. Author queries to extract the identified information. These queries can extract direct values or aggregations from the data. For example, provenance can be queried for an instance of a particular type or provenance can be quiet for the number of instances of a particular type;
3. Author sentences templates for the target audiences appropriate for those identified in Step 1. in both first and third-person perspectives (see Table 1 for examples of sentence template). The sentence templates are authored in HTML so that they can benefit from hyperlinks. The sentence templates utilise the CAS types to refer to a resource so that we can reduce the number of URIs in the narrative, however, these URIs are preserved in the narrative through hyperlinks;
4. Execute the queries and identify which templates can be fulfilled based on the queries' results. Using those identified queries, replace the variables with the values from the queries.

These descriptions can be embedded into CASs to show in a transparent manner how resources are used and generated by the system. They can be used to describe user behaviour, which helps increase users' awareness of others and their actions, thus supporting accountability.

11 DISCUSSION

Our methodology aims to aid in the design of CASs, models of diversity, support diverse analysis requirements and provide accountability to continue to encourage a diverse community. Recording diversity in CASs means that it can be analysed throughout its components and participants. The transparency of diversity in CASs can enable developers to formulate approaches to support diversity. For example, developers could develop an incentive for participants to complete tasks with others that have different skills to them. Transparency of diversity also promotes trust within its community. For example, participants who share accommodation may want to share with others that have similar preferences and hobbies, people tend to trust others how are similar to themselves.

In the following Table 2, we discuss how we address the diversity typology that we identified in Section 2. Temporal aspects of diversity mentioned in Section 2 have been catered for by using the **prov:wasDerivedBy** relationship typed with **prov:Revision**. This enables the state changes to be modelled, and can show how profiles, preference, capabilities, goals and roles evolve (see Figure 13).

While accountability affords many benefits, it may, however, lead to breaches in privacy. Diversity may be expressed in PROV using types and properties, which are regarded as private data. Semantic inference may lead to exposing private diversity information. For example, the following statements expose X and Y's sexual preferences.

```
X and Y are involved in a marriage activity
    -> X and Y may or may not be of the same gender

U and V are involved in a civil partnership
    -> U and V are same sex couple
```

It is, therefore, important that a Privacy Impact Assessment (PIA) should be completed during each iteration of designing provenance templates in our methodology (Step 3 in Section 5). Specifically, PIA is a tool that you can use to identify and reduce the privacy risks. A PIA can reduce the risks of harm to individuals through the misuse of their personal information. It can also help you to design more efficient and effective processes for handling personal data.

In order to show the diversity in the SmartShare application, we describe how diversity is allowed for in the application in Table 3.

12 RELATED WORK

The diversity of people in online systems has long been recognised [37, 7, 33]. There have been numerous studies evaluating cultural differences in online systems [4, 21, 1, 10], which identify that diversity plays a big in the outcome of these systems. These types of evaluations often heavy rely on user studies and provide no standard models which can be used to provide comparisons between different systems.

There has been a large body of work advocating accountability in distributed systems [19, 2, 34], which handle a diverse set of system component. The work presented in [2] describes the role of accountability in distributed systems. They identify that accountability makes it "possible to tolerate, detect, isolate, discourage, and remove misbehaving components". In CASs accountability can play

Description	First Person	Second Person
Collective	You and {list_of_users} took part in a ride as {role}.	{list_of_users} took part in a ride as {role}.
Reputation Report	You have an average overall rating of {average_rating} from {no_feedback_reports} feedback reports left by {no_authors} authors. Out of the {total_feedback_reports} feedback reports written by you, only {no_feedback_reports} were used to generate your rating. The feedback reports used to generate your rating were authored in the last {no_days} day/s.	This user has an average overall rating of {average_rating} from {no_feedback_reports} feedback reports left by {no_authors} authors. Out of the {total_feedback_reports} feedback reports written about them, only {no_feedback_reports} were used to generate the rating. The feedback reports used to generate that rating were authored in the last {no_days} day/s.
Users Behaviour	You have left an average feedback of {average_feedback}, and have written {no_authored_reports} feedback reports. You have left feedback for {no_unique_users} different users, and it agrees with {agreement_percentage}% of the other raters. Your feedback that does not agree with other raters was {disagreement_percentage}% higher.	This user left an average feedback of {average_feedback}, and has written {no_authored_reports} feedback reports. They have left feedback for {no_unique_users} different users, and it agrees with {agreement_percentage}% of the other raters. The feedback that does not agree with other raters was {disagreement_percentage}% higher.

Table 1. Sentence templates supporting both first and second person perspective, elements surrounded by {} are variables.

the same role, where components and users can detect misbehaving components or users.

Provenance can be used to describe the flow of information and human participation in activities. Applications that record provenance and provenance use cases are well documented [3, 24, 25, 13]. Moreover, the use cases include support for: making social computations accountable and transparent [36, 31]; determining whether data or users can be trusted [16]; and ensuring reproducibility [26] of computations; auditability and accountability [36]; deriving trust and classification [17]; asserting attribution and generating acknowledgements [27]; and traceability [8]. To enable such a powerful functionality, however, one needs to adapt or write applications, so that they generate provenance information, which can then be exploited to offer new benefits to their users. Provenance can be generated during runtime [11, 14, 28], compile time [6, 5], and reconstructed retrospectively [22, 9].

Previously, Semantic Web technologies have been used to generate narratives [35, 18, 12]. In more detail, Tuffield et al. [35] and Jewell et al. [18] describe the OntoMedia ontology, which supports the generation of narratives. Tuffield et al. [35] discuss approaches to generate narratives from a vocabulary, the approaches included are based on character, plot and user modelling. Jewell et al. [18] describes how OntoMedia is used to annotate the vast collection of heterogeneous media. Geurts et al. [12] use ontological domain knowledge to select and organise a narrative discourse on a topic of interest to a user.

13 CONCLUDING REMARKS

In this paper, we present a typology for diversity in CASs and a methodology to aid in the design of CASs, models of diversity, support diverse analysis requirements and provide accountability to continue to encourage a diverse community. Supporting diverse analysis requirements promotes trust and familiarity in the CAS and its participants. This transparency allows its participants to view how components and others behave. Thus, it enables its participants more information when to support their choice to contributing to a transaction and or joining collectives. The methodology presented in this paper draws on linked data principles to provide the basis of an information model that is diversity aware and supports reuse. PROV and the CAS vocabulary allow the actions of CASs peers to be modelled, this model can be exploited by other services to support end-users or adaptive algorithms.

The narrative approach is one such example of how to convey and support diversity, by enabling provenance information about reputation to be consumed easily by humans with different perspectives. We have planned an evaluation, to evaluate explanations from the provenance generated by the reputation service that will enable users to understand (1) how their reputation is generated, which takes into account the decay of feedback reports; (2) recommendations of which subject to choose, which are motivated by whether a subject routinely leaves feedback and whether they rate highly, which contrasts to using just a reputation rating to support decision-making; and (3) how they are perceived by others, which aims to increase their awareness that their actions within an CAS have consequences.

ACKNOWLEDGEMENTS

The research leading to these results has received partially funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement n. 600854 Smart Society: hybrid and diversity-aware collective adaptive systems: where people meet machines to build smarter societies http://www.smart-society-project.eu/.

REFERENCES

[1] Alexandre Ardichvili, Martin Maurer, Wei Li, Tim Wentling, and Reed Stuedemann, 'Cultural influences on knowledge sharing throughtr online communities of practice', *Journal of knowledge management*, **10**(1), 94–107, (2006).

[2] Elisa Bertino, Wonjun Lee, Anna C Squicciarini, and Bhavani Thuraisingham, 'End-to-end accountability in grid computing systems for coalition information sharing', in *Proceedings of the 4th annual workshop on Cyber security and information intelligence research: developing strategies to meet the cyber security and information intelligence challenges ahead*, p. 29. ACM, (2008).

[3] Rajendra Bose and James Frew, 'Lineage retrieval for scientific data processing: A survey', *ACM Computing Surveys*, **37**(1), 1–28, (March 2005).

Diversity	Support
1. Diversity in Participants	The participants in a CAS are typed with labels stemming from prov:Agent
(a) Human Participants	In the CAS Vocabulary human participants are described as **cas:User**. A **cas:User** can have attributes of **cas:Profile**, **cas:Preference** and **cas:Capability** (see Figure 7). In more detail: **cas:Profile** resources provide additional information about a user which may include details about their gender, age or allergies. **cas:Preference** resouces describe a users preferences, which may include a user's homepage, contact hours or type of food. **cas:Capability** resources describe a user's skill, which may include recognising star constellations, driving, or taking photographs. These types are assigned to PROV elements during the Step 3 of our methodology presented in Section 5. Any additional types required to define an agent's profile, preferences and capabilities for a specific purpose would be defined in Step 1 (see Section 5).
i. Individual Users	See the description for 1. (a)
ii. Collectives	An agent of type **cas:Collective** can act on behalf of others with the type **cas:User**, **cas:Peer**, **cas:Agent** (see Figure 8). Each of these agents can be resources of type **cas:Profile**, **cas:Preference** and **cas:Capability**. These types are assigned during the Step 3 of our methodology (see Section 5).
iii. Ad-Hoc Collectives	In order to differentiate between collectives that have been created by their users the collective can be attributed to entities of type **cas:Profile**, **cas:Preference** and or **cas:Capability** which describe the connection between the entities (see Figure 9). These types are assigned during the Step 3 of our methodology (see Section 5).
iv. Developers and Designers	Can be described by using the type **cas:User**. Their roles, profiles and capabilities can be typed with **cas:Role**, **cas:Profile** and **cas:Capability**, respectively. These types are assigned to PROV elements during the Step 3 of our methodology (see Section 5).
(b) System Components	Have type **cas:Peer**, which can be described by resources of type **cas:Profile** and **cas:Capability**. These types are assigned during the Step 3 of our methodology (see Section 5).
(c) Other	Have type **cas:Agent**, which can be described by resource of type **cas:Profile** and **cas:Capability**. These types are assigned during the Step 3 of our methodology (see Section 5).
2. Diversity in Interest	Can be described in the provenance captured from the CAS, this diversity can be queried using SPARQL, and the results from queries can be presented in different ways.
(a) Different Goals	Can be captured in the provenance by using the type **cas:Goal**. This type is assigned to PROV elements during the Step 3 of our methodology (see Section 5).
(b) View different Information	A wide range of queries can be authored to retrieve information from the provenance. These queries are designed during the Step 4 of our methodology (see Section 5).
(c) Alternative Presentation	The information from queries can be presented in different ways, see Section 10 for an example of sentence templates which can use a first person or third person perspective to describe the data. These sentence templates would be designed during the Step 4 of our methodology (see Section 5).
3. Diversity in Roles	Can be captured in the provenance by using the type **cas:Role**. This type is assigned to PROV elements during the Step 3 of our methodology (see Section 5).

Table 2. Table describing how our approach supports the diversity typology in Section 2.

[4] Patrick YK Chau, Melissa Cole, Anne P Massey, Mitzi Montoya-Weiss, and Robert M O'Keefe, 'Cultural differences in the online behavior of consumers', *Communications of the ACM*, **45**(10), 138–143, (2002).

[5] James Cheney, 'Program slicing and data provenance', *IEEE Data Engineering Bulletin*, 22–28, (December 2007).

[6] James Cheney, Amal Ahmed, and Umut A. Acar, 'Provenance as dependency analysis', *Mathematical Structures in Computer Science*, **21**(6), 1301–1337, (2011).

[7] BA Collis and Elka Remmers, 'The www in education: issues related to cross-cultural communication and interaction', (1997).

[8] Francisco Curbera, Yurdaer Doganata, Axel Martens, Nirmal K. Mukhi, and Aleksander Slominski, 'Business provenance – a technology to increase traceability of end-to-end operations', in *OTM 2008 Confederated International Conferences*, eds., Robert Meersman and Zahir Tari, pp. 100–119. Springer, (2008).

[9] Tom De Nies, Io Taxidou, Anastasia Dimou, Ruben Verborgh, Peter M. Fischer, Erik Mannens, and Rik Van de Walle, 'Towards multi-level provenance reconstruction of information diffusion on social media', in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pp. 1823–1826, New York, NY, USA, (2015). ACM.

[10] Casey Dugan, Werner Geyer, Michael Muller, Joan DiMicco, Beth Brownholtz, and David R Millen, 'It's all'about you': diversity in online profiles', in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pp. 703–706. ACM, (2008).

[11] James Frew, Dominic Metzger, and Peter Slaughter, 'Automatic capture and reconstruction of computational provenance', *Concurrency and Computation: Practice and Experience*, **20**(5), 485–496, (2008).

[12] Joost Geurts, Stefano Bocconi, Jacco Van Ossenbruggen, and Lynda Hardman, *Towards ontology-driven discourse: From semantic graphs*

Diversity	Support
1. Diversity in Participants	
(a) Human Participants	
i. Individual Users	SmartShare users are typed with **cas:User**, and a driver has the skill driver.
ii. Collectives	In SmartShare individuals can select to be in the same collective to share a ride after they have been matched because they share the same departure, destination and time. This is modelled in the provenance captured.
iii. Ad-Hoc Collectives	In SmartShare individuals are placed into ad-hoc collectives by the matching service based on their departure and destination locations and times. The matching service then proposes these matches to the matched users. This is modelled in the provenance captured.
iv. Developers and Designers	The SmartShare application has different components each component has a different team of developers.
(b) System Components	SmartShare is comprised of Orchestrator, Negotiation, Matcher, and Reputation services, and a mobile application. Each of these services have their own type **cas:Orchestrator**, **cas:NegotiationManager**, **cas:MatchingManager**, **cas:ReputationManager**, and **cas:MobileApplication**, respectively, these types stem from **cas:Peer**.
(c) Other	For now, there are no other types of participants. This could change if driverless cars were part of the system.
2. Diversity in Interest	
(a) Different Goals	The end-users have different goals, some want to save money on their commute while others want a transportation. This is represented in the provenance using the **cas:Goal** type. The developer for each component also have different goals, for example, the orchestrator aims to organises tasks as efficiency as possible, while the reputation services aims to generate reputation reports.
(b) View Different Information	User's of the system want to see which ride match their requirements and reputation information about each ride participant. This information is modelled in the provenance. Administrators want an overview of the activity on SmartShare for example how many rides have been completed, where the most rides have taken place, and how many new users there are. These figures can are obtained from the provenance.
(c) Alternative Presentation	Provenance information can be displayed as graphs to the developers and textual information using sentence templates to end-users. The reputation service generates text to describe how reputation is generated and statistics about how a user interacts with the reputation service.
3. Diversity in Roles	The end-users can have two types of role, **cas:Driver** and or **cas:Commuter**.

Table 3. Table describing how our approach supports the diversity typology in Section 2.

to multimedia presentations, Springer, 2003.

[13] Yolanda Gil, James Cheney, Paul Groth, Olaf Hartig, Simon Miles, Luc Moreau, and Paulo Pinheiro da Silva, 'Provenance xg final report', Technical report, World Wide Web Consortium, (2010).

[14] David A. Holland, Margo Seltzer, Uri Braun, and Kiran-Kumar Muniswamy-Reddy, 'Pass-ing the provenance challenge', *Concurrency and Computation: Practice and Experience*, **20**(5), (2008).

[15] Hook Hua, Curt Tilmes, Stephan Zednik (eds.), and Luc Moreau, 'PROV-XML: The PROV XML Schema', W3C Working Group Note NOTE-prov-xml-20130430, World Wide Web Consortium, (April 2013).

[16] Trung Dong Huynh, *Trust and reputation in open multi-agent systems*, Ph.D. dissertation, University of Southampton, 2006.

[17] Trung Dong Huynh, Mark Ebden, Matteo Venanzi, Sarvapali Ramchurn, Stephen Roberts, and Luc Moreau, 'Interpretation of crowdsourced activities using provenance network analysis', in *Conference on Human Computation and Crowdsourcing (HCOMP'13)*, (November 2013).

[18] Michael O Jewell, K Faith Lawrence, Mischa M Tuffield, Adam Prugel-Bennett, David E Millard, Mark S Nixon, Nigel R Shadbolt, et al., 'Ontomedia: An ontology for the representation of heterogeneous media', in *In Proceeding of SIGIR workshop on Mutlimedia Information Retrieval*. ACM SIGIR, (2005).

[19] Ryan KL Ko, Peter Jagadpramana, Miranda Mowbray, Siani Pearson, Markus Kirchberg, Qianhui Liang, and Bu Sung Lee, 'Trustcloud: A framework for accountability and trust in cloud computing', in *Services (SERVICES), 2011 IEEE World Congress on*, pp. 584–588. IEEE, (2011).

[20] Timothy Lebo, Satya Sahoo, Deborah McGuinness (eds.), Khalid Behajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao, 'PROV-O: The PROV Ontology', W3C Recommendation REC-prov-o-20130430, World Wide Web Consortium, (October 2013).

[21] Doo H Lim, 'Cross cultural differences in online learning motivation', *Educational Media International*, **41**(2), 163–175, (2004).

[22] Sara Magliacane, 'Reconstructing provenance', in *The Semantic Web — ISWC 2012*, volume 7650 of *Lecture Notes in Computer Science*, pp. 399–406. Springer, (2012).

[23] Danius Michaelides, Trung Dong Huynh, and Luc Moreau. Provtemplate: A template system for prov documents, June 2014. Technical Note.

[24] Simon Miles, Paul Groth, Miguel Branco, and Luc Moreau, 'The requirements of recording and using provenance in e-science experiments', *Journal of Grid Computing*, **5**(1), 1–25, (2007).

[25] Luc Moreau, 'The foundations for provenance on the web', *Foundations and Trends in Web Science*, **2**(2–3), 99–241, (November 2010).

[26] Luc Moreau, 'Provenance-based reproducibility in the semantic web', *Web Semantics: Science Services and Agents on the World Wide Web*, **9**(2), 202–221, (July 2011).

[27] Luc Moreau and Paul Groth, *Provenance: An Introduction to PROV*, Morgan and Claypool, September 2013.

[28] Luc Moreau and Paul Groth, 'Provenance of publications: A prov style for latex', in *Seventh USENIX Workshop on the Theory and Practice of Provenance (TAPP'15)*, Edinburgh, Scotland, (July 2015). USENIX.

[29] Luc Moreau, Paolo Missier (eds.), Khalid Belhajjame, Reza B'Far, James Cheney, Sam Coppens, Stephen Cresswell, Yolanda Gil, Paul Groth, Graham Klyne, Timothy Lebo, Jim McCusker, Simon Miles, James Myers, Satya Sahoo, and Curt Tilmes, 'PROV-DM: The PROV Data Model', W3C Recommendation REC-prov-dm-20130430, World Wide Web Consortium, (October 2013).
[30] Luc Moreau, Paolo Missier (eds.), James Cheney, and Stian Soiland-Reyes, 'PROV-N: The Provenance Notation', W3C Recommendation REC-prov-n-20130430, World Wide Web Consortium, (October 2013).
[31] Heather S Packer, Laura Drăgan, and Luc Moreau, 'An auditable reputation service for collective adaptive systems', in *Social Collective Intelligence*, 159–184, Springer, (2014).
[32] Heather S Packer and Luc Moreau, 'Generating narratives from provenance relationship chains', in *Proceedings of the 2015 Workshop on Narrative & Hypertext*, pp. 37–41. ACM, (2015).
[33] John P Robinson, Alan Neustadtl, and Meyer Kestnbaum, 'The online "diversity divide": Public opinion differences among internet users and nonusers', *IT & Society*, **1**(1), 284–302, (2002).
[34] Thomas Ruebsamen and Christoph Reich, 'Supporting cloud accountability by collecting evidence using audit agents', in *Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on*, volume 1, pp. 185–190. IEEE, (2013).
[35] Mischa M Tuffield, Nigel R Shadbolt, and David E Millard, 'Narrative as a form of knowledge transfer: Narrative theory and semantics', in *In Proceedings of the First AKT DTA Colloquium*, (2005).
[36] Daniel J. Weitzner, Harold Abelson, Tim Berners-Lee, Joan Feigenbaum, James Hendler, and Geral Jay Sussman, 'Information accountability', *Commun. ACM*, **51**(6), 81–87, (June 2008).
[37] Peter Woolliams and David Gee, 'Accounting for user diversity in configuring online systems', *Online Review*, **16**(5), 303–311, (1992).

Diversity-Aware Recommendation for Human Collectives

Pavlos Andreadis and **Sofia Ceppi** and **Michael Rovatsos** and **Subramanian Ramamoorthy** [1]

Abstract.

Sharing economy applications need to coordinate humans, each of whom may have different preferences over the provided service. Traditional approaches model this as a resource allocation problem and solve it by identifying matches between users and resources. These require knowledge of user preferences and, crucially, assume that they act deterministically or, equivalently, that each of them is expected to accept the proposed match. This assumption is unrealistic for applications like ridesharing and house sharing (like airbnb), where user coordination requires handling of the diversity and uncertainty in human behaviour.

We address this shortcoming by proposing a diversity-aware recommender system that leaves the decision-power to users but still assists them in coordinating their activities. We achieve this through taxation, which indirectly modifies users' preferences over options by imposing a penalty on them. This is applied on options that, if selected, are expected to lead to less favourable outcomes, from the perspective of the collective. The framework we used to identify the options to recommend is composed by three optimisation steps, each of which has a mixed integer linear program at its core. Using a combination of these three programs, we are also able to compute solutions that permit a good trade-off between satisfying the global goals of the collective and the individual users' interests. We demonstrate the effectiveness of our approach with two experiments in a simulated ridesharing scenario, showing: (a) significantly better coordination results with the approach we propose, than with a set of recommendations in which taxation is not applied and each solution maximises the goal of the collective, (b) that we can propose a recommendation set to users instead of imposing them a single allocation at no loss to the collective, and (c) that our system allows for an adaptive trade-off between conflicting criteria.

1 Introduction

Sharing economy applications constitute an interesting domain for multi-agent resource allocation and coalition formation. In these applications, users act as producers and consumers of resources, aiming to find peers to share the resources with, while a platform supports them during peer discovery and resource sharing. These fundamental aspects of sharing applications highlight how the decisions of the *collective* of users lead to a globally desirable outcome, while the choices of a single user alone have no such power. However, the services the sharing applications provide should leave the decision-making power to each user in order to allow her to express her preferences and satisfy her individual needs. Consequently, instead of facing the problem of identifying a solution for the collective of users, the platform needs to help them in coordinating their individual choices in such a way that the goal of the collective can still be achieved. In this work, we tackle this issue by providing a recommender system that accounts for user preferences and facilitates the coordination among users, in scenarios where users perform *joint* tasks in subgroups consisting of the members of a larger collective. In the example of a ridesharing application, each user could be aiming to achieve the best fit between his schedule and the planned ride. However, since rides cannot be achieved without the collaboration of multiple users, the collective goal of facilitating as many users as possible will come into conflict with this individual preference.

Many multi-agent applications face the problem of coordinating autonomous agents that aim to share resources, which can be seen as a resource allocation problem. Traditional approaches to this problem typically express a degree of centralised control in order to provide functional, viable solutions to most, if not all, participating users. In particular, several algorithms have been designed that identify stable matches between users and resources [14, 18, 9]. However, users cannot affect the algorithm. The most flexible approaches proposed in the literature make use of sequential mechanisms that allow users to accept or reject the solution currently proposed to them [11, 1]. Finally, some of the existing approaches assume that the system knows the complete preference ordering of users over, e.g., other users [8]. The crucial drawback of this type of work is that it focuses on problems like how to assign children to schools, how to allocate students to shared rooms, and how to match donors with patients in the kidney exchange market. In these scenarios, the possibility that users might prefer not to be allocated, rather than be to allocated as prescribed by the algorithm, is not considered. However, this assumption makes the adoption of such algorithms unrealistic for several sharing applications, e.g. ridesharing and joint event planning.

Indeed, these approaches lack a crucial characteristic that systems that mediate between humans should have: the ability to model human diversity and consider the uncertainty of human behaviour. Indeed, human decision making is affected by multiple factors: social, cultural, psychological, personal, and available information [16], that are unique for each individual. These create variations *among* individuals in terms of preferences over given characteristics of the peers and resources, leading to diversity across users. Moreover, the variability of these factors adds complexity to the decision-making process *within* each individual, to the extent that near identical situations may lead to significantly different behaviour. This leads to uncertainty regarding user behaviour. Crucially, a sharing application that does not account for diversity and the uncertainty of human behaviour is likely to fail in supporting large-scale coordination within human collectives

[1] School of Informatics, University of Edinburgh, United Kingdom, email: p.andreadis@sms.ed.ac.uk, sceppi@inf.ed.ac.uk, mrovatso@inf.ed.ac.uk, sramamoo@staffmail.ed.ac.uk

effectively.

Ideally, a system could address user diversity and provide a very personalised service by eliciting information from users and understanding their general preferences over some defined characteristics of the services. In the literature, there is ample work providing techniques for learning user preferences in an accurate way [17, 7, 4, 12, 10, 5, 13, 21, 15] and that focuses on delivering personalised services [2, 10, 19, 6]. However, apart from the tricky task of eliciting information from users and understanding how any given factors affect user preferences, a system has to deal with the problem of understanding which factors affect human behaviour. This is a currently open problem that is attracting attention from researchers interested in, e.g., social computation and psychology. Given this, a sharing application should account for uncertainty in human behaviour. In particular, in designing such an application, the designer has to (i) pay attention to the type of interaction between the system and the users (both individually as a collective) and (ii) allow for flexibility such that it can adapt to unforeseen behaviour. In this work, in order to provide such flexibility, instead of offering users a single option computed with the techniques discussed above, we focus on the problem of recommending multiple options to users.

The allocation problem faced by sharing applications, whose users aim to find peers to share a resource or a task with, is of a combinatorial nature. As such, when a system offers multiple options, all the users assigned to a task have to agree to it, i.e. they have to choose the task for it to happen. Since there is no guarantee that users' independent choices are consistent with one another, the system has to provide a coordination mechanism. This problem can be seen as a coalition formation problem [20] in which incentives to stay in a suggested coalition may be provided to users who would otherwise reject it. Cost of stability [3] and taxation [23] are two techniques proposed to provide such incentives and achieve the desired effect by artificially modifying users' preferences. In this work, instead of using explicit coordination techniques that require communication with users, we provide an indirect coordination mechanism, based on the techniques used in coalition formation problems. More specifically, we introduce a taxation mechanism in the options computation process.

Note that a sharing application that aims to adapt to the user collective but also wants to account for the interests of individual users, faces a multi-criteria optimisation problem [17]. Indeed, the interest of the collective (that requires users' collaboration) is in conflict with the interest of individual users whose aim is to obtain what is the best option for themselves. The approach we propose allows the sharing application to specify to which extent it wants to account for individual users' interests and identify options that achieve the desired trade-off between conflicting interests.

The three main contribution of this work are:

- A formulation of the user coordination problem faced by sharing economy applications, in such a way that it allows for the explicit representation of the diversity and uncertainty in human behaviour;
- A diversity-aware system for the coordination of users in sharing economy applications that does not require communication between users;
- Experimental evidence for the necessity of taking human diversity and uncertainty into account when coordinating such applications. Specifically, we demonstrate that we can replace direct allocations with recommendation sets at no cost, while also allowing for adaptively trading-off between various criteria of optimality.

The remainder of this paper is organised as follows. In Section 2, we provide a formal description of the allocation problem that characterises sharing applications, propose a diversity-aware approach that aims to account for diversity and uncertainty of human behaviour, and describe our framework. Section 4 proposes a detailed description and formulation of the mixed integer linear programs used in our optimisation framework. The experimental evaluation and obtained results are described in Section 5. Finally, Section 6 concludes the paper.

2 Formal Model

In this section, we formalise the resource allocation problem that characterises sharing economy applications.

Consider a set of tasks $J = \{1, \dots, |J|\}$. Each task $j \in J$ is associated with one and only one user who owns the task, for example, the user is the owner of the resource that will be shared, or whoever initiated a concrete sharing task. Since we assume that each owner has one and only one task associated with her, for the sake of simplicity, we interchangeably refer to $j \in J$ as both the task and its owner. Let $I = \{1, \dots, |I|\}$ be the set of users who do not own a task, and K the set of all users, i.e. $K = I \cup J$. Each owner $j \in J$ aims to find non-owners to share her task with and each non-owner $i \in I$ aims to find one task to join. All users have requirements and preferences about tasks and users to share a task with.

Let $\mathbf{x} = \{x_{1,1}, \dots, x_{1,|J|}, \dots, x_{|I|,1}, \dots x_{|I|,|J|}\} \in \mathbf{X}$ define which non-owner joins each task, i.e., $\mathbf{x}$ is an allocation of non-owners to tasks, where $\mathbf{X}$ is the set of allocations. In particular, if i is allocated to task j then $x_{i,j} = 1$, otherwise $x_{i,j} = 0$.

The preferences of each user $k \in K$ are represented by a utility function $u_k : \mathbf{X} \to \mathbb{R}$ which provides a complete ranking over potential allocations $\mathbf{x} \in \mathbf{X}$. Similarly, the system-level utility function is defined as $U_s : \mathbf{X} \to \mathbb{R}$.

Crucially, in sharing economy applications, single users and the collective of users have conflicting interests. While a user i aims to maximise her utility $u_i(\cdot)$, the interest of the collective of users, represented by the system-level utility function $U_s(\cdot)$, is related to the overall benefit the users can achieve. For example, $U_s(\cdot)$ may consider the sum of the user utility or the number of users that are allocated to tasks. Given this, it is obvious that the maximisation of $U_s(\cdot)$ provides no guarantee to individual users in terms of achieved utility. In order to provide such a guarantee, the application designer should, e.g., maximise the fairness of the solutions (i.e., minimise the difference between the utility achieved by every user) or maximise the minimum single user utility. However, in this case, no guarantee is given in terms of system-level utility.

The aim of the application is to aid users in finding compatible peers by suggesting allocations while accounting for this conflict of interest. However, it is fundamental to highlight that not all allocations are guaranteed to occur. Indeed, each user $k \in K$ selects an allocation from a set R or recommended solutions independently and without direct coordination with other users, according to a *user response model*. The three user response models typically used in the literature are [21]:

- *noiseless* response model: each user acts deterministically and always selects the solution that would maximise her utility. Formally, if $\mathbf{x} \in R$ is such that $u_k(\mathbf{x}) \geq u_k(\mathbf{x}'), \forall \mathbf{x}' \in R$ then $p_k(\mathbf{x}) = 1$ otherwise $p_k(\mathbf{x}) = 0$ for all $k \in K$, where $p_k(\mathbf{x})$ is the probability with which user $k \in K$ chooses solution $\mathbf{x}$.
- *constant noise* response model: each user selects the solution that would maximise her utility the majority of the time irrespective of the utility of other solutions. Each of the remaining solutions

is chosen with a equal small probability. Formally, if $\mathbf{x} \in R$ is such that $u_k(\mathbf{x}) \geq u_k(\mathbf{x}'), \forall \mathbf{x}' \in R$ then $p_k(\mathbf{x}) = \alpha$ otherwise $p_k(\mathbf{x}) = \beta$ with $\alpha >> \beta$ and $\sum_{\mathbf{x} \in R} p_k(\mathbf{x}) = 1$ for all $k \in K$.

- *logit* response model: each user selects an allocation from the set R proportionally to its utility value. Formally, $p_k(\mathbf{x}) = \frac{u_k(\mathbf{x})}{\sum_{\mathbf{x}' \in R} u_k(\mathbf{x}')}, \forall k \in K$.

Given that with each of these response models, every user selects a solution without reasoning about other users' choices but, rather, makes her decisions by exclusively considering her utility over each allocation, in order for a task to occur, all the users, owner and non-owners, allocated to that task in a given allocation $\mathbf{x}$ have to select $\mathbf{x}$. For example, if allocation $\mathbf{x}$ assigns to task j the subset of non-owners $\tilde{I} \subseteq I$, then, in order for task j to occur, all non-owners $i \in \tilde{I}$ must select allocation $\mathbf{x}$.

3 Diverse Aware Approach

The problem described in the previous section can be approached as a resource allocation problem, in which users are implicitly assumed to be compliant with any solution proposed to them, and are therefore not afforded any alternatives. The results of such an approach are constrained to that of a matching between users and resources. Consequently, there is no consideration of the inherent uncertainty in user behaviour, or the fact that users could simply refuse to participate in systems that do not satisfy their needs. A system that realistically addresses human diversity and the uncertainty in human behaviour requires an explicit representation of user preferences and their responses to different decision scenarios. Furthermore, the decision scenario needs to be formulated so that it allows for the recommendation of solutions to users, while accounting for their possible deviations from expected behaviour.

In this section, we will provide a detailed framework for the representation of the diversity and uncertainty in user behaviour, and outline our approach that focuses on the problems of *recommending* alternatives and facilitating the coordination of users. In particular, our system intends to present a set of allocations $R = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{|R|}\}$ of fixed size $|R|$.

To achieve this goal we need to deal with two issues. The first of these is the multi-criteria optimisation problem in which we have to balance the conflicting interests of each single user (represented by her utility function $u_i(\cdot)$) and the interest of the collective of users (represented by the system-level utility function $U_s(\cdot)$). We overcome this problem by computing solutions that guarantee a minimum level of system utility and maximise, e.g. the fairness of the solution with respect to the allocated resources. The second is that the uncoordinated selection of allocations done by the users, along with their diversity in preferences, makes it unlikely that users will select allocations such that the task can actually occur. In order to help the system in the process of coordinating users' selections, we introduce a *taxation* mechanism, so as to influence user selection behaviour by artificially modifying the utility they have for the recommended solutions, i.e. by modifying their preferences. Effectively, taxation allows the system to impose a penalty on allocations users are better of not selecting. Generally, the tax imposed is different for each user and for each allocation, and must guarantee that users still have multiple options (e.g. the system cannot impose an infinite tax). We develop a different taxation mechanism for each of the user response models described in Section 2.

Crucially, these two problems must be tackled simultaneously, otherwise properties that are satisfied when the first issue is solved, e.g. fairness, may not hold anymore if taxation is applied in a separate step. The reason behind this is that both the problems and the solution to these problems are related to the function, i.e. user utility.

Now, we describe our approach that aims to optimise the recommendation set R while simultaneously dealing with the problems due to user's and collective's conflicting interests, and the lack of coordination in user selection. In order to handle this problem, we iteratively construct the recommendation set by sequentially executing three Mixed Integer Linear Programs (MILPs), each of them guaranteeing different solution properties. In this way we can deal with both the multi-criteria optimisation and the computational complexity of finding an exact solution.

In particular, in order to account for this conflict of interest, we initially construct a program called $MILP^{system}$ that aims to maximise the system utility $U_s(\cdot)$ and thus find a solution with the highest utility V^* that the system can achieve. A second program called $MILP^{first}$ takes V^* as input and guarantees that the computed solution achieves at least a give percentage of V^* in terms of system utility, while the objective function of the program is focused on maximising a different property, e.g. fairness. The advantage of this approach is that the application controls exactly to which extent the maximisation of the system-level utility and the fairness are satisfied. The alternative would have been to use a single MILP whose objective function accounts for both the system utility and the fairness. However, in this case *(i)* the best trade-off between the two factors would have been decided by the program and not by the application designer, and *(ii)* it would be possible to obtain solutions completely unbalanced towards one of the two factors.

To face the lack of coordination among users, we aim to modify their utility for the recommended allocations such that they all prefer the same solution, termed *sponsored* solution. This solution is the one computed by $MILP^{first}$ and, since is the one we want to sponsor, we do not alter the utility the users have for it. Instead, we apply taxation on all other recommended solutions. Since we need to solve the problem of identifying the solutions with the desired properties and apply taxation simultaneously (as explained in the previous section), we design a third program, $MILP^{others}$, dedicated to this. In particular, a solution obtained with this program aims to be similar to the one of $MILP^{first}$ in terms of users' utility, guarantees a minimum level of system utility, and has taxes computed on the basis of the specific user response model considered.

Given this, we can view our framework as composed of three steps. In the first one, $MILP^{system}$ is executed in order to identify the highest possible system utility achievable, in the second step $MILP^{first}$ is used to identify the sponsored solution, and in the third step all the remaining non-sponsored $|R| - 1$ solutions are computed by executing $MILP^{other}$ $|R| - 1$ times.

Note that, users may have other requirements that the system should satisfy, for example, they may have constraints regarding the characteristics of users they are willing to share a task with. Thus, all the MILPs must satisfy these requirements in order to compute a feasible solution. In the next section, we provide a description of the constraints needed to satisfy the properties our framework necessitates and the different type of user requirements.

4 Optimisation Problem Formulation

In this section we present the details of the Mixed Integer Linear Programs (MILPs) that compose the framework described in the previous section. For the sake of clarity and without loss of generality, we present the MILPs for the ridesharing scenario.

Ridesharing is a sharing application that can be modelled as specified above. Indeed, a set of passengers I and a set of drivers J aim to find people to share a ride with. Each driver is the owner of a task, i.e., a car, and passengers aim to join one car each. We assume that drivers impose their pick-up point, drop-off point, and time of pick-up on passengers they are sharing the ride with. Passengers, in turn, have preferences over the pick-up point and drop-off point, and their utility decreases with the distance between their preferences and what the driver they are assigned to imposes on them. The pick-up time does not affect users' utility, however the system imposes a threshold on the maximum difference between the pick-up time desired by a passenger and the one specified by the driver she is assigned to. Note that this is a hard requirement and thus no allocation that violates this can be recommended. Moreover, both passengers and drivers may require to be in a car without smokers, Finally, they may also require to either share the ride, or not to be in the same car as another specific user.

The requirements just described for the ridesharing scenario are examples of three different types of constraints that users of sharing applications may have. Thus, even if in the following formulations we focus on ridesharing, the constraints presented can be used for a wide range of applications. In order to illustrate the expressiveness of the MILP formulations and the requirements that can be captured, in what follows we describe the characteristic of each possible type of constraint and show how to formulate it by using an example.

4.1 Maximising collective-level objectives

The MILP presented in this section is used in the first step of our framework and aims to compute the maximum system utility achievable without violating any requirements.

We start by defining the utility function $u_i(\mathbf{x})$ of a passenger $i \in I$. Her utility function is affected by how much the allocation $\mathbf{x}$ satisfies her preferences. In particular, here we assume that users have preferences over two aspects that characterise a task: the pick-up point and the drop-off point.

Without loss of generality, assume the utility function $u_i(\mathbf{x})$ of each passengers $i \in I$ is a sum of partial utility functions $\alpha_i(\mathbf{x})$ and $\beta_i(\mathbf{x})$ as shown in Equation 1. In particular, $\alpha_i(\mathbf{x})$ is the contribution to the utility of agent i that depends on the difference between the pick-up point of i and the one of the driver assigned to her by allocation $\mathbf{x}$. Similarly, $\beta_i(\mathbf{x})$ depends on the difference between their drop-off point. We assume that these differences are divided into intervals and that all differences in the same interval affect the user's utility in the same way.

$$u_i(\mathbf{x}) = \alpha_i(\mathbf{x}) + \beta_i(\mathbf{x}) \tag{1}$$

The utility function of the system is a linear weighted combination as shown by Equation 2. In this specific case, w_1, w_2, and w_3 are the weights. The first weight multiplies the sum of passengers utility, i.e., the social welfare, the second the number of passengers that are allocated to a car, and the third the number of drivers. The idea is that the system cares about the sum of the utility achieved by passengers but also the number of users that have the possibility to get a ride.

$$U_s(\mathbf{x}) = w_1 \sum_{i \in I} u_i(\mathbf{x}) + w_2 \sum_{i \in I} \sum_{j \in J} x_{i,j} + w_3 \sum_{j \in J} 1 \tag{2}$$

We are now ready to define the objective function of $MILP^{system}$ that is to maximise the system's utility (Equation 3).

$$obj \quad max_{\mathbf{x} \in \mathbf{X}} U_s(\mathbf{x}) \tag{3}$$

We start the description of the constraints by focusing on the allocation variables $x_{i,j} \in [0,1], \forall i \in I, \forall j \in J$. Since each car $j \in J$ has a capacity c_j, we need to guarantee that no more than c_j passengers are allocated to j (Constraint 4). Finally, we need to guarantee that each passenger is allocated to at most one car (Constraint 5).

$$\sum_{i \in I} x_{i,j} \leq c_j, \forall j \in J \tag{4}$$

$$\sum_{j \in J} x_{i,j} \leq 1, \forall i \in I \tag{5}$$

We introduce a second set of variables $h_{i,i',j}$, one for each passenger $i \in I$, passenger $i' \in I$, and driver $j \in J$. These are binary variables indicating if two passengers are sharing the same car. In particular, Constraints 6 guarantee that $h_{i,i',j} = 1$ if passengers i and i' are both allocated to car j, and $h_{i,i',j} = 0$ otherwise.

$$\begin{aligned} h_{i,i',j} \leq & x_{i,j}, \forall i, i' \in I, \forall j \in J \\ h_{i,i',j} \leq & x_{i',j}, \forall i, i' \in I, \forall j \in J \\ h_{i,i',j} \geq & |x_{i,j} + x_{i',j}| - 1, \forall i, i' \in I, \forall j \in J \end{aligned} \tag{6}$$

We now move to describe the constraints needed to guarantee that the partial utility $\alpha_i(\mathbf{x})$ correctly reflects the distance between the pick-up preference of passenger i and what the driver she is assigned to imposes on her. Note that similar constraints are used to compute the partial utility $\beta_i(\mathbf{x})$. As mentioned before, we consider the possible pick-up distance as divided into $|T|$ intervals (where T is the set of intervals). For each interval $n \in T$, the parameter $\alpha_{i,n}$ indicates i's partial utility if the pick-up distance is in interval n, and parameters $\alpha_{n,lower}$ and $\alpha_{n,upper}$ denote the lower bound and the upper bound of interval n, respectively. A set of variables $k^{\alpha}_{i,n}$, one for each $n \in T$, is used to select the right interval. Note that $k^{\alpha}_{i,n}$ is a binary variable and that $k^{\alpha}_{i,n} = 0$ if the pick-up distance is in interval n and $k^{\alpha}_{i,n} = 1$ otherwise. Given this, the partial utility $\alpha_i(\mathbf{x})$ is given by Equation 7, and Constraints 8 guarantee the the only variable $k^{\alpha}_{i,n}$ that equals zero is the one of the interval n, to which the pick-up distance belongs to. Note that M is a very large number as typically used in the Big M method [22], while $\Delta^{\alpha}_i(\mathbf{x})$ measures the pick-up distance. In this particular case, we compute the distance by considering latitude ($p_{i,lat,pu}$ and $p_{j,lat,pu}$) and longitude ($p_{i,long,pu}$ and $p_{j,long,pu}$) of the pick-up points, and compute the Manhattan distance between them as shown in Equation 9. This equation is particularly interesting because it shows how we deal with imposing a zero partial utility to a passenger when she is not assigned to any car. In particular, in order to achieve this, a fictitious pick-up distance interval $n = |T|$ is introduced in the set T such that the large number M (bigger than any pick-up distance) is in interval $n = |T|$, $alpha_{i,|T|} = 0$, and $\Delta^{\alpha}_i(\mathbf{x}) = M$ if i is not allocated to any $j \in J$.

$$\alpha_i(\mathbf{x}) = \sum_{n \in T} \alpha_{i,n} \cdot (1 - k^{\alpha}_{i,n}), \forall i \in I \tag{7}$$

$$\begin{aligned} M \cdot k^{\alpha}_{i,n} + (\alpha_{n,upper} - \Delta^{\alpha}_i(\mathbf{x})) \geq & 0, \forall i \in I, \forall n \in T \\ M \cdot k^{\alpha}_{i,n} + (\Delta_i(\mathbf{x}) - \alpha_{n,lower}) \geq & 0, \forall i \in I, \forall n \in T \\ \sum_{n \in T} k^{\alpha}_{i,n} \leq |T| - 1, & \forall i \in I \end{aligned} \tag{8}$$

$$\Delta_i^{\alpha}(\mathbf{x}) = \sum_{j \in J} x_{i,j}(|p_{i,lat,pu} - p_{j,lat,pu}| + |p_{i,long,pu} - p_{j,long,pu}|) + (1 - \sum_{j \in J} x_{i,j}) \cdot M, \forall i \in I \quad (9)$$

In defining constraints due to requirements, we differentiate between *strict* constraints, *non-strict* constraints, and *potential* constraints.

Strict constraints impose that every group of users sharing a car must have the same value for a given user's parameter. For example, in the ridesharing scenario, a strict constraint is imposed on the day of the ride and, thus, all users allocated to the same car must have the same value for the parameter p_i^{day}. In particular, if the two users are passengers, then the strict constraint is Constraint 10, while if the two users are a passenger and a driver, then Constraint 11 must be imposed.

$$|h_{i,i',j} p_i^{day} - h_{i,i',j} p_{i'}^{day}| \leq 0, \forall i, i' \in I, \forall j \in J \quad (10)$$

$$|x_{i,j} p_i^{day} - x_{i,j} p_j^{day}| \leq 0, \forall i, \in I, \forall j \in J \quad (11)$$

Non-strict constraints impose a threshold on how a passenger's requirement is satisfied. In the ridesharing scenario, this type of constraint is applied to the difference between the time of pick-up specified by the passenger, and the one of the driver she is sharing the car with. We formulate this type of constraint as shown by Constraint 12, where p_i^{time} and p_j^{time} are parameters that indicate the pick-up time of passenger i and driver j, respectively, and $p_{threshold}^{time}$ is the threshold.

$$\sum_{j \in J} x_{i,j}(|p_i^{time} - p_j^{time}|) \leq p_{threshold}^{time} \quad (12)$$

Finally, we discuss potential constraints. Constraints of this type do not always impose a condition that must be satisfied by a solution. Indeed, a user may have specific requirements about a characteristic of the users she is sharing the task with or she may be indifferent with respect to this characteristic. For example, in the ridesharing scenario, a user may require to be in a car without smokers, while another user, even if she is not a smoker, may not have such a requirement. In order to formulate the constraint that guarantees these requirements, we need to introduce a new binary variable v_j^{smoker}, one for each car $j \in J$. Constraints 13 impose that $v_j^{smoker} = 1$, if at least one user among the passengers and the driver sharing car j requires to be in a car without smokers. Parameter $p_i^{reqNoSmoke}$ indicates the "no smoker request" of a passenger $i \in I$. In particular if $p_i^{reqNoSmoke} = 1$ the passenger requires to be in a car with no smokers, while if $p_i^{reqNoSmoke} = 0$ the passenger has no preferences. $p_j^{reqNoSmoke}$ is similarly defined for driver $j \in J$. Constraint 14 guarantees that if none of the users in a car j has a "no smokers" requirement then $v_j^{smoker} = 0$. Now, if variable $v_j^{smoker} = 1$, then we need to impose that all the users in car j do not want to smoke during the ride. This is achieved by Constraint 15, where parameter $p_i^{NoSmoke} = 0$ if passenger $i \in I$ wants to smoke in the car and $p_i^{NoSmoke} = 1$ otherwise. $p_j^{NoSmoke}$ is similarly defined for driver $j \in J$.

$$\begin{aligned} v_j^{smoker} &\geq x_{i,j} p_i^{reqNoSmoke}, \forall j \in J, \forall i \in I \\ v_j^{smoker} &\geq p_j^{reqNoSmoke}, \forall j \in J \end{aligned} \quad (13)$$

$$p_j^{reqNoSmoke} + \sum_{i \in I} x_{i,j} p_i^{reqNoSmoke} \geq v_j^{smoker}, \forall j \in J \quad (14)$$

$$p_j^{NoSmoke} + \sum_{i \in I} x_{i,j} p_i^{NoSmoke} \geq (c_j + 1) v_j^{smoker}, \forall j \in J \quad (15)$$

To conclude, we consider the case in which the system allows a user to specify if she wants/does not want to share a task with another specific user. The constraints used to guarantee these requirements are strict constraints and are formalised as follows. Constraints 16 and 17 guarantee that passengers i and i' and passenger i and driver j are allocated to the same car, respectively. While Constraints 18 and 19 guarantee the opposite.

$$\sum_{j \in J} h_{i,i',j} \leq 0 \quad (16)$$

$$x_{i,j} \leq 0 \quad (17)$$

$$\sum_{j \in J} h_{i,i'',j} \geq 1 \quad (18)$$

$$x_{i,j} \geq 1 \quad (19)$$

4.2 Maximising fairness among users

$MILP^{first}$ constitutes the second step of our framework. The aim for this program is to identify the first solution that will be presented to users. Note that this is the solution we would like all users to choose among the ones in the recommendation set. The solution the program provides guarantees a minimum level of system utility while focusing on an objective function that is oriented to being beneficial to the users. In the following formulation we assume, without loss of generality, that the program aims to maximise user fairness. This translates into an objective function that minimises the difference between the utility of every pair of passengers $i, i' \in I$ (Equation 20).

$$obj \quad min_{\mathbf{x} \in \mathbf{X}} \sum_{i \in I} \sum_{i' \in I | i' > i} |u_i(\mathbf{x}) - u_{i'}(\mathbf{x})| \quad (20)$$

As mentioned before, the solution provided by this program must guarantee a minimum level of system utility. Given the maximum utility V^* the system can achieve (computed by $MILP^{system}$) and the parameter $d \in [0, 1]$, the required guarantee can be obtained by imposing Constraint 21.

$$U_s(\mathbf{x}) \geq V^* \cdot d \quad (21)$$

In addition to this, all the constraints described for $MILP^{system}$ must also hold for $MILP^{first}$.

4.3 Maximising user coordination

The last step of our framework aims to compute the remaining $|R| - 1$ solutions (one has already been identified by $MILP^{first}$). To achieve this, $MILP^{others}$ is executed $|R| - 1$ times.

A solution identified by $MILP^{others}$ has the following three characteristics: it guarantees a minimum level of system utility (as $MILP^{first}$ does), computes a solution that is different from the ones previously chosen, and artificially modifies the utility each passenger $i \in \bar{I}$ has for this solution such that all of them would prefer the solution $\mathbf{x}^*$ identified by $MILP^{first}$. Note, that the set $\bar{I} \subseteq I$ is composed by all passengers $i \in I$ such that $\sum_{j \in J} x_{i,j}^* = 1$, i.e., only the utility of passengers who are assigned to a driver in solution $\mathbf{x}^*$ is

artificially modified. This last characteristic depends on the response model of the users, and is achieved by using taxation.

When considering the noiseless response model and constant noise model, in order for a passenger $i \in I$ to select/prefer the solution computed by $MILP^{first}$, we need to guarantee that her utility for the solution $\mathbf{x}$ currently computed is lower than her utility for allocation $\mathbf{x}^*$, i.e. the solution computed by $MILP^{first}$. We achieve this by imposing a tax $\tau_i(\mathbf{x})$ on passenger i for allocation $\mathbf{x}$ that decreases the utility i has for $\mathbf{x}$. Constraint 23 guarantees that $\mathbf{x}^*$ is the preferred solution of passenger i. In this constraint, ϵ is a very small number used to express the fact that $u_i(\mathbf{x}^*)$ must be strictly higher than $u_i(\mathbf{x}) - \tau_i(\mathbf{x})$. However, this constraint imposes a lower bound to the tax $\tau_i(\mathbf{x})$ but no upper bound. Thus, potentially, the program can assign an infinite value to $\tau_i(\mathbf{x})$. This is not desirable because no real options would be effectively given to the passengers if all but one could be infinitely taxed. Moreover, we also want to avoid the case in which $\tau_i(\mathbf{x})$ is higher than the minimum tax required as the system should not aim to unnecessarily extract excessive utility from its participants. Thus, the upper bound $\tau_i(\mathbf{x})$ should be equal to its lower bound. We obtain this by using the Big M method [22] that involves changing the objective function as shown by Equation 22.

$$min \sum_{i \in I} |u_i(\mathbf{x}^*) - u_i(\mathbf{x}) + \tau_i(\mathbf{x})| + M\Big(\sum_{i \in I} \big(u_i(\mathbf{x}^*) - \epsilon - u_i(\mathbf{x}) + \tau_i(\mathbf{x})\big)\Big) \quad (22)$$

$$u_i(\mathbf{x}^*) - \epsilon \geq u_i(\mathbf{x}) - \tau_i(\mathbf{x}) \quad (23)$$

Similarly to the constant response model case, $MILP^{others}$ needs to impose a lower bound on the tax also for the logit response model (Constraint 25), and modify the objective function (Equation 24) such that the tax is the lowest possible. However, since in this case the probability with which a passenger selects solution $\mathbf{x}$ is proportional to the utility that $\mathbf{x}$ represents for that passenger, imposing a lower bound to the tax means imposing a lower bound to the selection probability of $\mathbf{x}^*$ for each passenger $i \in \bar{I}$. In Constraint 25, ψ is the minimum selection probability required for $\mathbf{x}^*$. Note that this constraint is linear because everything but $u_i(\mathbf{x})$ and $\tau_i(\mathbf{x})$ are parameters given as input to the program.

$$min \sum_{i \in I} |u_i(\mathbf{x}^*) - u_i(\mathbf{x}) + \tau_i(\mathbf{x})| + M\Big(\sum_{i \in I} \big(u_i(\mathbf{x}^*) - \psi \cdot \Big(\sum_{\mathbf{x}' \in R} \big(u_i(\mathbf{x}') - \tau_i(\mathbf{x}')\big) + u_i(\mathbf{x}) - \tau_i(\mathbf{x})\Big)\Big)\Big) \quad (24)$$

$$\frac{u_i(\mathbf{x}^*)}{\sum_{\mathbf{x}' \in R} \big(u_i(\mathbf{x}') - \tau_i(\mathbf{x}')\big) + u_i(\mathbf{x}) - \tau_i(\mathbf{x})\}} \geq \psi \quad (25)$$

Finally, $MILP^{others}$ needs to guarantee that solution $\mathbf{x}$ is different from the ones previously computed. Depending on the required degree of difference between two solutions, we can formulate the $MILP$ constraints as follows. Constraint 26 guarantees that the solution $\mathbf{x}$ differs from the ones already computed, i.e. the ones in set R, at least for the allocation of one passenger. While Constraint 27 requires that each ride (except the one with no passenger allocated to it) of solution $\mathbf{x}$ differs from the corresponding ride in solution $\mathbf{x}' \in R$, at least for the allocation of one passenger.

$$\sum_{i \in I} \sum_{j \in J} |x_{i,j} - x'_{i,j}| > 1, \forall \mathbf{x}' \in R \quad (26)$$

$$\sum_{i \in I} |x_{i,j} - x'_{i,j}| > 1, \forall \mathbf{x}' \in R, \forall j \in J \quad (27)$$

In addition to these, all the constraints described for $MILP^{system}$ and $MILP^{first}$ must also hold for $MILP^{others}$.

5 Experimental Evaluation

In order to demonstrate the effectiveness of our approach, we run two sets of simulated experiments. In the first set, we compare the recommended set of solutions generated with our diversity-aware approach, with a set of solutions that maximise the system's utility and provide no support for user coordination. Essentially, the benchmark set is produced without considering the need for consistency across users' selections. In this way, we will demonstrate that our diversity-aware approach is *strictly better* for the generation of *recommendation sets*.

In the second set of experiments, we compare our approach to that of allocating a single solution that maximises system utility. We assume users are characterised by a utility threshold of acceptance, unknown to the system. This latter set of experiments, will show how recommending a set of solutions through our approach can produce *results that are equally good* to what a *direct allocation* would have produced.

5.1 Experiment design

For both types of experiments we consider different configurations, each of which is characterised by the population of users, specifically the number of users and the percentage of drivers among them, the value of the threshold d used in $MILP^{first}$ and $MILP^{others}$, the user response model, and, for the logit model, the probability ϵ with which the sponsored allocation $\mathbf{x}^*$ is selected. In particular, we run experiments for 10 and 20 users, for each percentage of drivers among 20, 30, and 40 percent. We vary the utility threshold d for $MILP^{first}$ and $MILP^{others}$ between the values 0.5, 0.75, and 1. The user models evaluated are the constant noise model and the logit model, and, for the latter, the probability ϵ of users selecting the sponsored recommendation is either 60 or 80 percent. Every configuration is repeated 100 times. We choose not to evaluate the noiseless response model because, by construction, it is a special case of the the constant noise model with the best performance, i.e. with the probability of the most preferred option set to 1. Without loss of generality, we set the weights $w_1 = w_2 = w_3 = 1$ in the system-level utility function.

The metrics used for each experiment are *system utility*, *fairness* (computed as described in the previous section), *number of drivers with allocated passengers*, and *number of allocated passengers*. Note, that all evaluations are performed *after* user selections have been performed. Thus, rides that have not been chosen by all the users allocated to them are not considered in the performance evaluation. Finally, we highlight that the metrics proposed account for the taxation imposed on the solutions. That is, each user, given his final allocation, has had any taxation imposed on him deducted from his effective utility, which in turn affects the evaluation of the system-level utility (Eq. 2) and fairness of allocation (Eq. 20).

The procedure used to obtain the experimental results is the following: First we generate the desired number of users, divided into drivers and passengers as prescribed by the configuration. For each user, we randomly generate the latitude and longitude of the pick-up point and drop-off point (we restrict the variability of this coordinate to 50), the pick-up time, whether she allows smokers in the car, and whether she

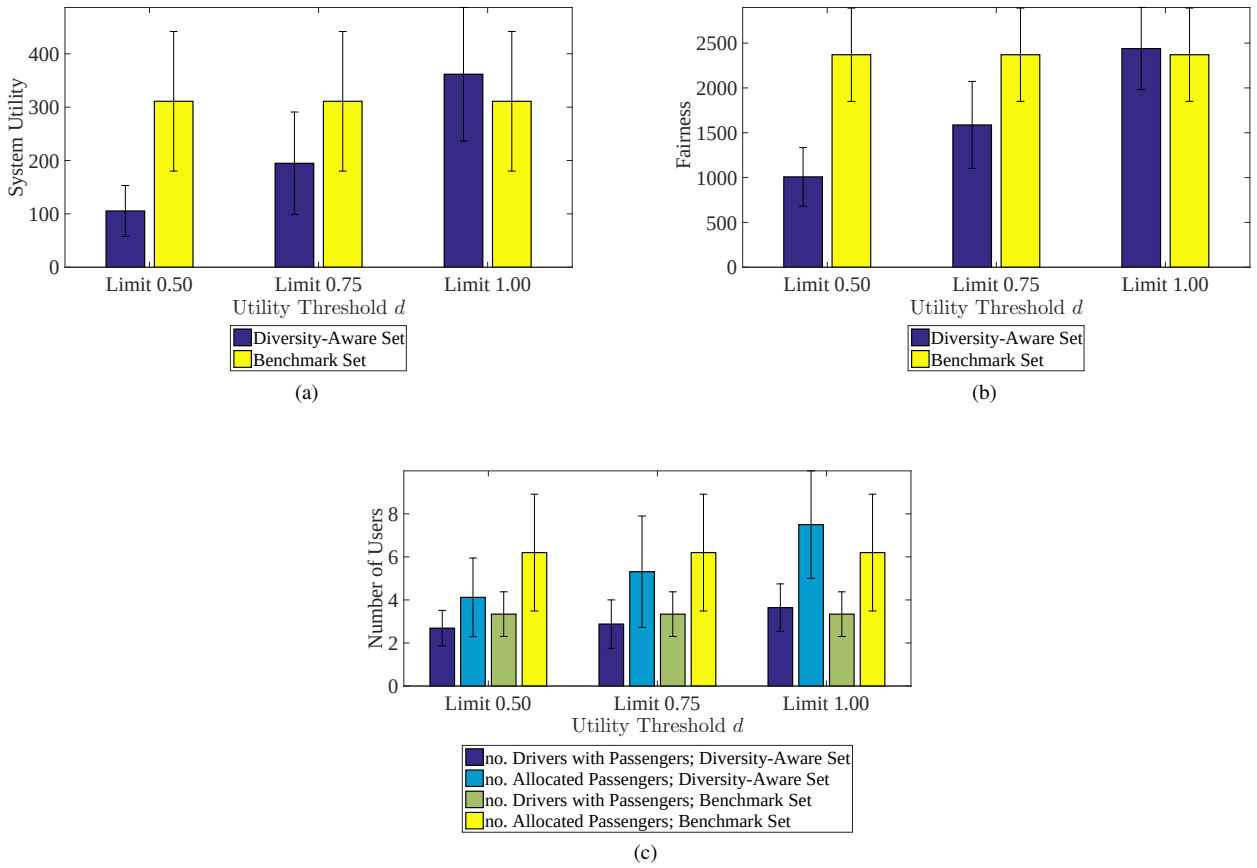


Figure 1: Results on *System Utility* (a), *Fairness* (b), and the number of *Allocated Passengers* and *Drivers with Passengers* (c) for the experiments with a recommendation set benchmark, with 20 users, 30% drivers, and constant noise.

wants to smoke. Then we generate $|R| = 7$ solutions (where possible) both for the case without coordination support in which the goal of the application is solely to maximise the system's utility, and following our approach. Finally, we simulate user behaviour according to the respective user response model and compute the metrics listed above.

This final step of our evaluation changes slightly depending on the benchmark we are comparing our approach to. As mentioned before, we consider two benchmarks: a benchmark with a recommendation set and a benchmark where a single allocation is proposed to the users. Note that in both cases, the solutions computed aim to maximise system-level utility.

When we consider the benchmark with a recommendation set, we compute the two sets of $|R|$ solutions computed as described above. Then, both for the benchmark case and for our approach, we recommend to each passenger the rides she is allocated to by these solutions. Each passenger then, independently, and without knowledge of other passengers' behaviour, selects a solution in accordance with her utility over each option, and her response model. Given these independent choices, we identify which rides have been selected by all the users allocated to them and, on the basis of this, we evaluate the performance of both approaches.

In the comparison with the benchmark with a single allocation, we assume that passengers select rides that satisfy a minimum level of user utility, i.e. there is a threshold over the user utility and solutions that do not satisfy this threshold cannot be selected. For example, in the case of ridesharing, we can assume that if the utility of a passenger for a ride is lower than her utility for taking the train, then she chooses not to join that ride. Given this, we consider the set $|R|$ of solutions computed using our diversity-aware approach and, for the benchmark, the solution computed by $MILP^{system}$. We assume that users apply this utility threshold and thus, all the rides that do not satisfy the threshold are removed from the ones that can be selected. This can be understood as users implicitly *rejecting* these solutions. We therefore make use of this label in corresponding figures. After this, each user selects one of the remaining options (note that in the benchmark case each user has either one ride or no option available). Now, as in the case of the previous benchmark, given these independent choices, we identify which rides have been selected by all users allocated to them and, on the basis of this, we evaluate the performance of both approaches.

A key point to remember, is that our diversity-aware approach influences users' utility over allocations through the use of taxation, and that this directly impacts on selection behaviour. We expect that this will aid in aligning user selections, and therefore lead to greater performance in terms of allocated passengers and drivers with passengers and, consequently, in terms of system-level utility.

5.2 Results

Below we present and discuss the results of our experiments for the case of 20 users with 30% drivers as representative of all experiments. The results on other population sizes and driver percentages were qualitatively equivalent. We analyse each set of experiments (recommendation set and single allocation) separately and present the

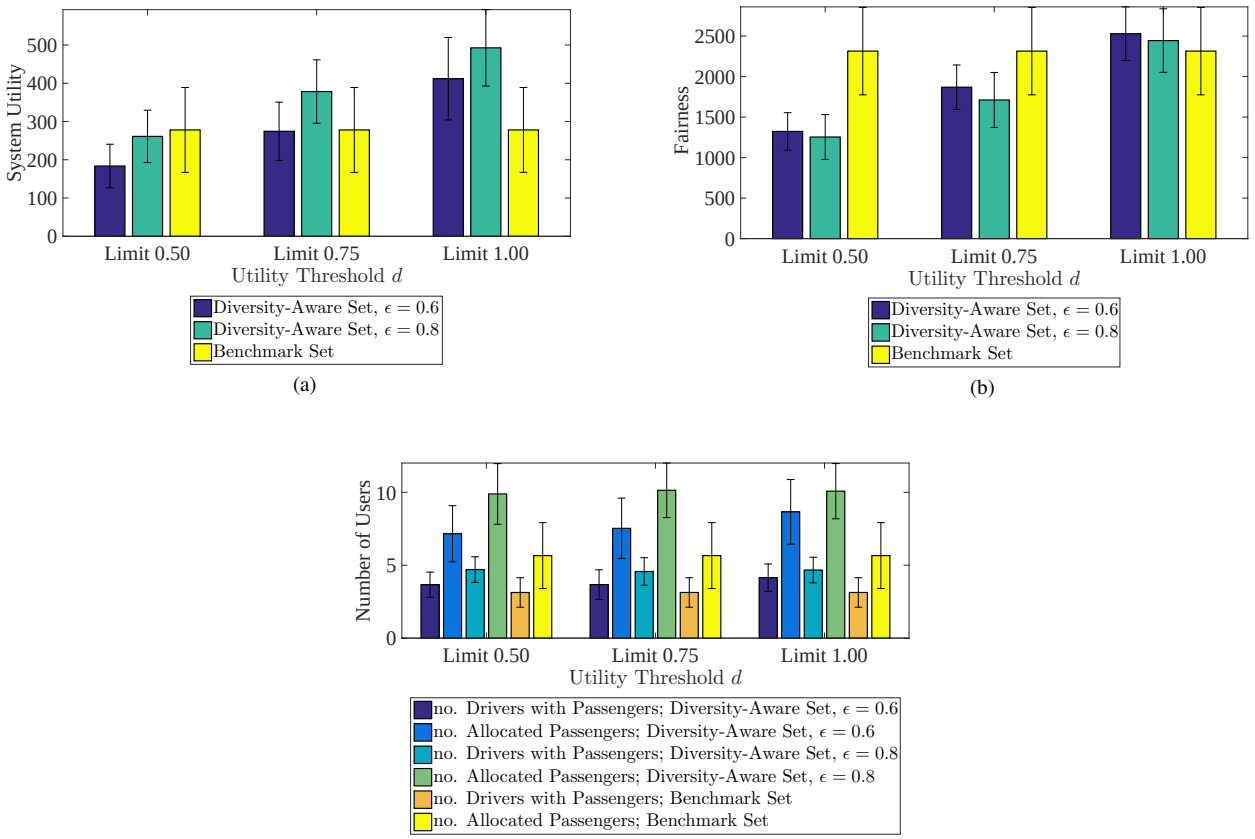


Figure 2: Results on *System Utility* (a), *Fairness* (b), and the number of *Allocated Passengers* and *Drivers with Passengers* (c) for the experiments with a recommendation set benchmark, with 20 users, 30% drivers, and logit noise.

constant and logit noise cases for each of them.

5.2.1 *Experiments with set recommendation benchmark*

This subsection discusses the experimental results from the set of experiments where the benchmark also presents the passengers with a set recommendation. Focusing first on the *constant noise* case (Fig. 1) we notice that there is a near linear trade-off between fairness and system utility (Fig. 1a, 1b). A trade-off that can be modulated by change of the utility threshold d. We also note that when $d = 1$, which forces MIP^{first} and MIP^{others} optimisations to prioritise system utility maximising solutions, the diversity-aware approach slightly outperforms the benchmark set, in terms of system utility, number of allocated passengers, and number of drivers with passengers (Fig. 1a, 1c). Reducing the value of parameter d offers better results on fairness (Fig. 1b), in comparison to the benchmark, but at a cost of reduced system utility (Fig. 1a).

Moving on to the *logit noise* case (Fig. 2), we notice once more the role of the utility threshold parameter d in trading off system utility and fairness (Fig. 2a, 2b). Further, the only scenario in which we perform slightly worse than the benchmark, in terms of user allocation and system utility, is for $\epsilon = 0.6$ and $d = 0.5$, i.e. when we optimise mostly for fairness and where we try not to influence user decisions too much. In terms of fairness, we only under-perform for $d = 1$, a result emerging from the large number of users with 0 utility, as results from the benchmark (Fig. 2a, 2c). Otherwise, we notice that the diversity-aware procedure significantly outperforms the set recommendation benchmark, in terms of system utility, number of allocated passengers, and number of drivers with passengers (Fig. 2a, 2c).

Summarising the results of this subsection, we note the significant improvement in performance afforded by our diversity-aware system. This signifies how important it is to be aware of the diversity among users when coordinating in sharing economy applications. These results increase in significance once we consider that there is no coordination present between users, and all the improvement is the result of implicit system interactions with each individual user; users that are free to choose amongst recommended alternatives. We conclude that making set recommendations by simply listing a set of system-optimal alternatives is a significantly sub-optimal procedure.

5.2.2 *Experiments with single allocation benchmark, and rejection*

This subsection discusses the experimental results from the set of experiments involving a diversity-aware set recommendation and a benchmark single allocation, while considering passengers that can reject rides. Focusing first on the *constant noise* case (Fig. 3), we notice that our system under-performs in terms of system utility (Fig. 3a), a loss it gains in its increased performance in terms of fairness (Fig. 3b). We further notice, as above, that there is a near linear trade-off between fairness and system utility, which can be modulated by change of the utility threshold d (Fig. 3a, 3b).

Finally, in the *logit noise* case for the second set of experiments (Fig. 4), we notice once more the role of the utility threshold parameter d

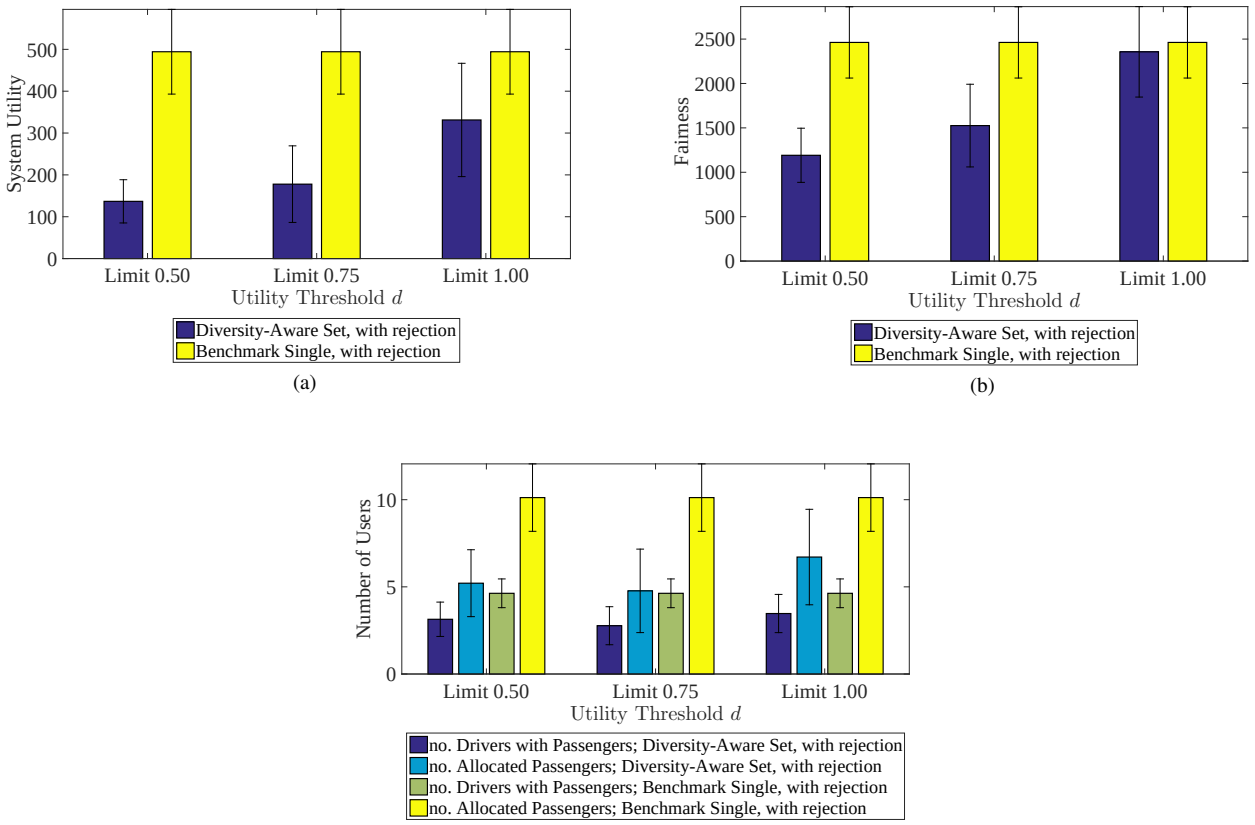


Figure 3: Results on *System Utility* (a), *Fairness* (b), and the number of *Allocated Passengers* and *Drivers with Passengers* (c) for the experiments with a single allocation benchmark, and rejection, with 20 users, 30% drivers, and constant noise.

in trading off system utility and fairness (Fig. 4a, 4b). Importantly, we note that our diversity-aware approach, which recommends a set, allowing for users to freely choose their preferred option, matches the performance of the benchmark, which allocates a single solution to each user (Fig. 4a, 4c). These results hold for when we do not wish to emphasise fairness, i.e in the $d = 1$ scenario.

The results of this subsection show that our diversity-aware set recommendation system, can consistently provide results that are equivalent to those of allocating a single item to each user. This shows, that providing users with options can be essentially *free*, in terms of system utility, even without considering any other beneficial effects that could result from allowing a system to recommend rather than allocate solutions to its user base. Moreover, we are afforded additional options in trading-off system utility with fairness.

6 Conclusion

We presented a methodology for the coordination of user collectives, in the absence of communication among agents. Our diversity-aware approach significantly outperforms the system utility maximising procedure, demonstrating that the recommendation of sets of solutions in sharing applications requires explicitly handling the uncertainty over user behaviour. Furthermore, we showed how our procedure can match the performance of a direct allocation of users to resources. This significant result demonstrates that we can allow users to *have a choice* in their alternatives, at no loss to the system. Lastly, our procedure allows for the adaptive trade-off between system-level utility and fairness of final allocation.

Future work will examine handling beliefs over user preferences in the context of recommending a set of options for sharing economies. Specifically, we are studying the inclusion of *active learning* procedures in the mixed integer linear program formulations. Further, we are interested in studying the robustness of our procedures to varying degrees of incorrect assumptions.

REFERENCES

[1] Haris Aziz, Markus Brill, Felix A. Fischer, Paul Harrenstein, Jérôme Lang, and Hans Georg Seedig, 'Possible and necessary winners of partial tournaments', *J. Artif. Intell. Res. (JAIR)*, **54**, 493–534, (2015).

[2] Yoram Bachrach, Sofia Ceppi, Ian A. Kash, Peter Key, Filip Radlinski, Ely Porat, Michael Armstrong, and Vijay Sharma, 'Building a personalized tourist attraction recommender system using crowdsourcing', in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, pp. 1631–1632, (2014).

[3] Yoram Bachrach, Edith Elkind, Reshef Meir, Dmitrii Pasechnik, Michael Zuckerman, Jörg Rothe, and Jeffrey S. Rosenschein, *The Cost of Stability in Coalitional Games*, 122–134, Springer Berlin Heidelberg, 2009.

[4] C. Boutilier, 'A POMDP formulation of preference elicitation problems', in *Proceedings of the 18th National Conference on Artificial Intelligence*, pp. 239–246, (2002).

[5] D. Braziunas, 'Computational approaches to preference elicitation', Technical report, University of Toronto, (2006).

[6] Sofia Ceppi and Ian Kash, 'Personalized payments for storage-as-a-service', *SIGMETRICS Perform. Eval. Rev.*, **43**(3), 83–86, (2015).

[7] Urszula Chajewska, Daphne Koller, and Ronald Parr, 'Making rational decisions using adaptive utility elicitation', in *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pp. 363–369. AAAI Press, (2000).

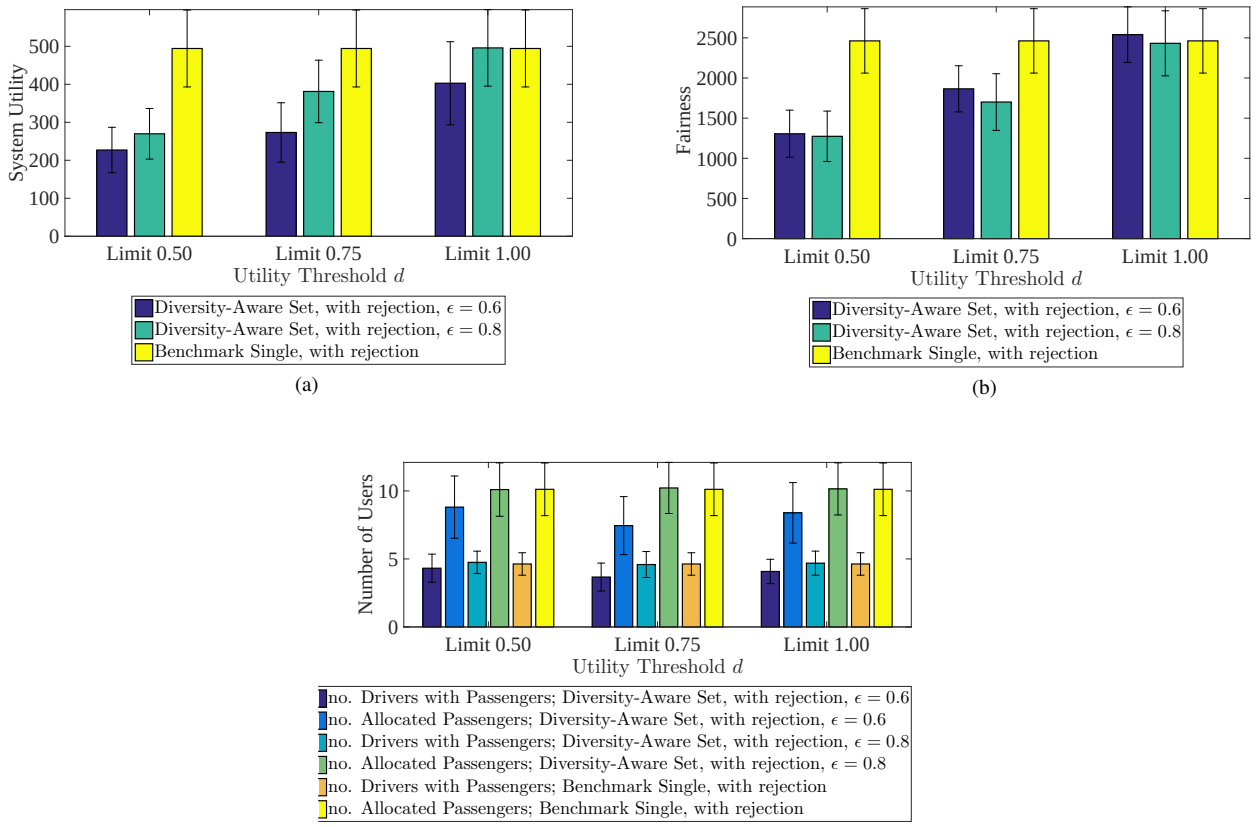


Figure 4: Results on *System Utility* (a), *Fairness* (b), and the number of *Allocated Passengers* and *Drivers with Passengers* (c) for the experiments with single allocation benchmark, and rejection, with 20 users, 30% drivers, and logit noise.

[8] Kim-Sau Chung, 'On the existence of stable roommate matchings', *Games and Economic Behavior*, **33**(2), 206 – 230, (2000).

[9] John P. Dickerson, Ariel D. Procaccia, and Tuomas Sandholm, 'Optimizing kidney exchange with transplant chains: theory and reality', in *International Conference on Autonomous Agents and Multiagent Systems, AAMAS*, pp. 711–718, (2012).

[10] Krzysztof Gajos and Daniel S. Weld, 'Preference elicitation for interface optimization', in *Proceedings of the 18th Annual ACM Symposium on User Interface Software and Technology*, UIST '05, pp. 173–182, New York, NY, USA, (2005). ACM.

[11] D. Gale and L. S. Shapley, 'College admissions and the stability of marriage', *The American Mathematical Monthly*, **69**(1), 9–15, (1962).

[12] Christophe Gonzales and Patrice Perny, 'GAI networks for utility elicitation.', *KR*, **4**, 224–234, (2004).

[13] Shengbo Guo and Scott Sanner, 'Real-time multiattribute bayesian preference elicitation with pairwise comparison queries', in *International Conference on Artificial Intelligence and Statistics*, pp. 289–296, (2010).

[14] Dan Gusfield and Robert W. Irving, *The Stable Marriage Problem: Structure and Algorithms*, MIT Press, Cambridge, MA, USA, 1989.

[15] Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Jose M Hernández-lobato, 'Collaborative Gaussian processes for preference learning', in *Advances in Neural Information Processing Systems*, pp. 2096–2104, (2012).

[16] Daniel Kahneman and Amos Tversky, 'Choices, values, and frames.', *American psychologist*, **39**(4), 341, (1984).

[17] R. L. Keeney and H. Raiffa, *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*, Cambridge University Press, 1993.

[18] Michel Balinski Mourad Baou, 'The stable allocation (or ordinal transportation) problem', *Mathematics of Operations Research*, (3), 485–503, (2002).

[19] Filip Radlinski and Susan Dumais, 'Improving personalized web search using result diversification', in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 691–692, (2006).

[20] Onn Shehory and Sarit Kraus, 'Methods for task allocation via agent coalition formation', *Artificial Intelligence*, **101**(1), 165 – 200, (1998).

[21] Paolo Viappiani and Craig Boutilier, 'Optimal bayesian recommendation sets and myopically optimal choice query sets', in *Advances in Neural Information Processing Systems*, pp. 2352–2360, (2010).

[22] Wayne L. Winston, *Introduction to Mathematical Programming: Applications and Algorithms*, Duxbury Resource Center, 2003.

[23] Yair Zick, Maria Polukarov, and Nicholas R. Jennings, 'Taxation and stability in cooperative games', in *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems*, pp. 523–530, (2013).

A Semantic Distance based Architecture for a Guesser Agent in ESSENCE's Location Taboo Challenge

Kemo Adrian [1], **Aysenur Bilgin** [2] and **Paul Van Eecke** [3]

Abstract. Taboo is a word-guessing game in which one player has to describe a target term to another player by giving hints that are neither the target term nor other terms specified in a predetermined list of taboo words. The Location Taboo (LT) Challenge, which has been proposed by the ESSENCE Marie Curie Initial Training Network, is a version of Taboo that only contains cities as target terms and is intended to be played by artificial guesser agents. The hints are extracted from games played by many different human players, whose associations of cities with specific terms are often based on past experiences and therefore very diverse. Modeling this diversity in word associations is one of the main difficulties in solving the LT Challenge. In this paper, we propose a semantic distance based architecture for a guesser agent for the LT Challenge. The proposed architecture employs a two-step approach that narrows down the geographical area of the guess first to the country and then to the city. For ranking countries and cities, different distance metrics are used. As these techniques can be used on web documents crafted by many different individuals, they are well suited to model the diversity in word associations. The results of our evaluation on the LTC test set show that the proposed guesser agent can guess the target city with up to 23.17% accuracy. For 68% of the correct guesses, the proposed agent guesses the target city faster than its human counterpart.

1 Introduction

Taboo is a word-guessing game in which one player has to describe a target term to another player by giving hints that are neither the target term nor other terms specified in a predetermined list of taboo words. For example, a player might have to describe *water* without using *sea*, *blue* or *beverage*. The Location Taboo Challenge (LTC), which has been proposed by the ESSENCE Marie Curie Initial Training Network [1], is a version of Taboo that only contains cities as target terms and is intended to be played by artificial guesser agents. In the LTC, the hints, which are words associated to the target city, are sequentially provided to the guesser agent, and the goal is to guess the target location as soon as possible.

The hints are extracted from games that were played by various human players having different backgrounds and demographics. The associations that individual players make with cities are often based on their own past experiences, and are therefore very diverse. For example, people that have visited Spain only once in their lives might associate *tapas* with *Madrid*, whereas others may think of *tapas* being typical for Andalusian cities and may not even consider it as a clue for *Madrid*. Modeling this diversity in word associations is one of the main difficulties in solving the ESSENCE LT Challenge.

In this paper, we propose a semantic distance based architecture for an LTC guesser agent. The proposed architecture employs a two-step approach that narrows down the geographical area of the guess first to the country and then to the city. For scoring the associative relevance of countries and cities with the given hints, the proposed architecture uses different distance measures. As these metrics are based on a large number of web documents crafted by various individuals, they considerably capture the diversity in word associations posed by human players.

The rest of this paper is structured as follows: Section 2 presents the game specification of the LT Challenge in more detail. Section 3 is dedicated to the background and previous work on modeling human behavior for word-guessing games. Section 4 presents the proposed architecture and the algorithms employed by our guesser agent. The experiments and the results are presented in Section 5. We provide a critical discussion in Section 6 followed by open research directions in Section 7. Finally, we draw conclusions in Section 8.

2 Game Specification

In this section, we introduce the most important aspects of the Location Taboo Challenge. A complete specification of the challenge can be found in [1].

An LTC game is played by two agents, the *describer* and the *guesser*. The game starts with the describer, providing a hint about a particular city anywhere in the world. Based on this hint, the guesser tries to guess the city that is being described. There are two possible outcomes after a guess has been made. For the outcome where the guess is correct, the game is considered to be successful. However, for the outcome where the guess is incorrect, the describer provides another hint and the game continues until the describer has consumed all the hints. The LT Challenge consists of implementing a guesser agent that can guess the correct city using the fewest number of guesses possible and before the describer runs out of hints. In the case where the describer runs out of hints and the correct guess has not yet been made, the game is considered to have failed.

For the LTC, the describer agent is provided by the authors of the challenge and the hints are crowd-sourced from real games played by human players. Therefore, the length of a game - i.e. the number of hints - is not fixed, but determined by the individual players. Also, it should be noted that the real-world dataset, which is provided by ESSENCE Network, consists of only successfully finished games. After each guess, the describer provides not only a new hint, but also the city that the human player (wrongfully) guessed. This information may be useful, or even necessary, in order to interpret

[1] IIIA-CSIC, email: kemo.adrian@iiia.csic.es
[2] Institute for Logic, Language and Computation, University of Amsterdam, Netherlands, email: a.bilgin@uva.nl
[3] Sony Computer Science Laboratory Paris, email: vaneecke@csl.sony.fr

the next hints, as these might be relative to the guesser's previous guesses (e.g. 'north' or 'close'). Hints are usually single words, but can occasionally be multi-word expressions. According to the rules of the LTC, the hints do not include proper names. An example game, adopted from [1] is shown in Figure 1.

Target:	Venice
D:	sea
G:	Sydney
D:	festival
G:	Rio de Janeiro
D:	river
G:	Rome
D:	art
G:	**Venice**

Figure 1. Location Taboo Challenge example game, adopted from [1], where D = Describer agent, G = Guesser agent

3 Background and Previous Work

What makes the LT Challenge so interesting and difficult is that the game is not about finding a correct or objectively verifiable answer to a specific question. Instead, it is about mimicking those associations that the human players have made, for whatever possible reason. The hints provided by the describer may not be necessarily true for the target city; yet, they are the depiction of an association that a human player made with this city. Therefore, an ideal implementation of the guesser agent should not only model common sense, but also simulate human beings' associative capabilities and collaborative game-playing behavior.

There is an impressive body of previous work on modeling common sense and human behavior for game playing. Heith et al. [9] present a range of techniques for understanding and conveying concepts based on word associations. These methods utilize human word association resources such as associative thesauri on the one hand; and corpus-based approaches, in particular Latent Semantic Analysis [6], Hyperspace Analog to Language [11] and Direct Co-occurrence Counts on the other hand. The models are evaluated both in a describer and a guesser role on Wordlery, a word-guessing game that is relatively similar to Taboo. The authors find that the models based on human word association resources are superior to the ones using corpus-based approaches.

A second, more famous, relevant research project is IBM's Watson, competing in the clue-guessing game Jeopardy! [4]. Watson uses IBM's massively parallel DeepQA architecture, combining hundreds of techniques and approaches in real time [7, 8]. The main difference between LT and Jeopardy! is that LT is a collaborative game, in which the describer tries to make the clues as easy and relevant as possible, whereas in Jeopardy!, the clues are made difficult on purpose. Furthermore, the clues in Jeopardy! are crafted by a team of people having all information available and are therefore always relevant and true in some way, whereas in LT, they have to be invented on the spot by a human player.

Finally, Pincus et al. [13] present a WordNet-based describer agent that generates clues for clue-guessing games, a project complementary to the implementation of a guesser agent in the LT Challenge.

[4] Jeopardy! is an American television game show created by Merv Griffin.

4 Guesser Agent Architecture

In this section, we present the proposed architecture for our guesser agent, as well as the different techniques and experimental configurations that will be used in the results section.

4.1 Basic Architecture

The basic architecture of our guesser agent can be described as follows. For the first incoming hint, the agent calculates the semantic distance between each country in the world and the given hint, using one of the metrics discussed in Section 4.2. Then, the guesser agent selects the top *N* countries, which were closest to the provided hint, and calculates the distances between the hint and each city in these countries. The idea is to provide the city with the highest score as a guess. If the guess is correct, the game finishes successfully. If the guess is incorrect and a new hint is provided, the distance between this new hint and each country in the world is calculated and added to the score of the previous hints. Unsuccessfully guessed cities are removed from the list of cities, such that they are never guessed twice. The process continues until the guess is correct or the describer runs out of hints. The algorithm is shown in Algorithm 1.

Algorithm 1: Guesser agent main algorithm

```
input: CountryList
while Success = false and new hints exist do
    Hint ← GetNewHint();
    foreach Country in CountryList do
        Country.Hint ← CalcDistance(Hint,Country);
        Country.Score ← AggregateDist(Country);
    end
    BestCountries ← SortOnDistance(CountryList,n)
    foreach Country in BestCountries do
        Country.Cities ← GetCities(Country);
        foreach City in Country.Cities do
            City.Hint ← CalcDistance(Hint,City);
        end
    end
    BestGuess ← GetClosestCity(BestCountries)
    Success ← GuessCity(BestGuess)
end
```

We have adopted this two-level approach, first pinpointing the countries and then the cities of the highest-ranked countries, for two main reasons. The first reason is that we observed that when humans play this game, many hints are as relevant for the country as for the city itself, with some hints even being more relevant for the country than for the city (such as *tapas* being more relevant for *Spain* than for *Madrid*). The second reason is related to efficiency. Calculating the distance for each hint in combination with all countries in the world requires a much lower number of queries than calculating this for all cities in the world.

4.2 Corpora and Distance Metrics

For calculating the distance between the geographical locations and the hints, we have used two different types of resources with their associated distance metrics. The following subsections will detail the types of resources, which are WordNet and Wikipedia, together with the distance measures.

4.2.1 WordNet

The first resource is WordNet [12], a lexical database linking English nouns, verbs, adjectives and adverbs by their semantic relations, including synonymy, hyperonymy, hyponymy and meronymy. The basic idea here is to exploit these hierarchical relations for measuring the semantic distance between the geographical locations and hints. The specific metric that we use is known as the Jiang-Conrath distance [10], which was found to perform very well when applied to WordNet [3]. The Jiang-Conrath (JC) distance subtracts the sum of the conditional log probabilities (reflecting *information content*) of the two terms from the conditional log probability of their lowest super-ordinate. The lower this number is, the closer the distance between the two terms. The formula of JC distance is presented in Equation (1) where $t1$ and $t2$ represent the two terms and lso stands for their lowest super-ordinate in the database. For words to which multiple synsets are associated, all synsets are tried and the best result is taken.

$$dist_{JC}(t_1, t_2) = 2log(p(lso(t_1, t_2))) - (log(p(t_1)) + log(p(t_2)))) \quad (1)$$

4.2.2 Wikipedia

The second resource that we used consists of all pages of English Wikipedia, as consulted on June 16, 2016. Using the Wikipedia API [5], the guesser agent queries the number of hits in the Wikipedia pages for a hint, a geographical location, and the hint and the geographical location combined. Then, using these hit counts, it employs three different metrics to score the association between the hint and the geographical location.

The first metric, which we call *Normalized Wiki Distance (NWD)*, is based on the Normalized Google Distance [5], but applied to the Wikipedia corpus. The formula is presented in Equation (2). t_1 and t_2 represent the two terms, $c(t)$ stands for the page counts of term t on Wikipedia and N stands for the total number of pages in Wikipedia. A lower NWD indicates a closer association between two terms.

$$NWD(t_1, t_2) = \frac{max(log(c(t_1)), log(c(t_2))) - log(c(t_1, t_2))}{log(N) - min(log(c(t_1)), log(c(t_2)))} \quad (2)$$

The second metric, which we call *Probabilistic Distance* (PD) is based on the ratio between the documents in which both terms occur and the documents in which the most frequent term occurs. When subtracted from 1, the closer this number is to 0, the higher the association between the two terms. The formula of PD is shown in Equation (3).

$$PD(t_1, t_2) = 1 - \frac{log(c(t_1, t_2))}{log(max(c(t_1), c(t_2)))} \quad (3)$$

Finally, we also used the *Pointwise Mutual Information (PMI)* measure [4], a word association metric that is commonly used in the field of computational linguistics for collocation extraction [2]. The formula is given in Equation (4). A higher PMI indicates a higher association of the two terms.

$$PMI(t_1, t_2) = log\frac{c(t_1, t_2)}{c(t_1)c(t_2)} \quad (4)$$

[5] https://www.mediawiki.org/wiki/API%3AQuery

4.3 M Most Salient (Famous) Countries

Algorithm 1 takes a list of countries as input. Only the countries in this list will be used in the computations and therefore, only the cities in these countries may be considered as a guess. The most salient (famous) countries are extracted from a ranked list of the countries with the corresponding number of hit counts in Wikipedia. We vary the number of most salient countries throughout the different experiments using a parameter *M*. Choosing a smaller *M* bears the risk of not considering the country of the target city, which will lead to a lost game. When considering countries with too few hit counts (larger *M*) on the other hand, the distance metrics described in the previous subsections may yield unexpected results due to data sparseness.

4.4 N Top Scoring (Best) Countries

In our guesser agent algorithm (see Algorithm 1), we first calculate the distance between the hints and the different countries from the provided country list. Then, for the *N* top scoring countries (i.e. having the closest semantic distances), we calculate the distances between their cities and the hints. So, only cities of the *N* best countries are considered as guesses. This parameter *N* regulates how much weight is given to the association between countries and the hints (instead of the cities).

5 Experiments and Results

We have evaluated our guesser agent on a set of 82 real-world games provided by ESSENCE. This section presents the cross categorical experiments and their results.

5.1 Experimental Setup

We have run several experiments varying the parameters *M* and *N* as discussed in the previous section. In the experiments, *M* takes the values 0, 10, 20, 30, 40, 50 and 60. The 0 value means that the country (salience) restriction is not active and that all countries in the world are considered. The parameter *N* takes the values 1, 2, 3, 4, 5, 10, 15, 20, 25, 50, 100 and ALL. In the case of ALL, all of the cities in all *M* countries are considered. The naming of the experiments follows the same abbreviation, which can be formalized as *FMBN*. In this abbreviation, *F* refers to Famous countries as mentioned in Section 4.3 and *B* refers to Best scoring countries as mentioned in Section 4.4. The parameters *M* and *N* in the *FMBN* abbreviation take the aforementioned values and hence we have 84 experiments for each metric. It should be noted that when *M=0*, the abbreviation is represented as *BN*, rather than *F0BN*.

5.2 Results of the experiments using WordNet

5.2.1 Jiang-Conrath Distance (JCD)

In this set of experiments, we have used the Jiang-Conrath Distance on WordNet to calculate the semantic distance between the hints and the geographical locations. The results of the 84 experiments suggest that the use of the 50 most salient (famous) countries in combination with a small selection (3-5) of best scoring countries yields the best results. Table 1 displays the top 5 configurations in terms of accuracy and in terms of successful games that were solved by the guesser agent using fewer number of guesses than the human counterpart. The top configuration for this set of experiments is F50B3 with an accuracy of 6,09% and a *faster guessing performance* of 80%.

Table 1. Results of experiments using Jiang-Conrath Distance on WordNet

Experiment Type	Successful Guesses	Faster Guesses	Accuracy (%)	Relative Faster Guessing Performance (%)
F50B3	5	4	6.09	80
F50B5	5	3	6.09	60
F50B4	4	3	4.87	75
F10B15	4	2	4.87	50
F50B15	3	3	3.65	100

Table 2. Top 5 results of experiments using Normalized Wiki Distance on Wikipedia

Experiment Type	Successful Guesses	Faster Guesses	Accuracy (%)	Relative Faster Guessing Performance (%)
F30B10	16	9	19.51	56.25
F30B15	15	9	18.29	60
F20B15	15	8	18.29	53.33
F30B5	15	6	18.29	40
F60B10	15	4	18.29	26.66

Table 3. Results of experiments using Probabilistic Distance on Wikipedia

Experiment Type	Successful Guesses	Faster Guesses	Accuracy (%)	Relative Faster Guessing Performance (%)
F50B2	18	10	21.95	55.55
F50B15	18	8	21.95	44.44
B2	17	9	20.73	52.94
B3	17	9	20.73	52.94
F60B2	17	9	20.73	52.94

Table 4. Results of experiments using Pointwise Mutual Information Measure on Wikipedia

Experiment Type	Successful Guesses	Faster Guesses	Accuracy (%)	Relative Faster Guessing Performance (%)
F20B10	19	13	23.17	68.42
F30B25	17	10	20.73	58.82
F20B15	17	9	20.73	52.94
F30B30	16	11	19.51	68.75
F30B30	16	11	19.51	68.75

5.3 Results of the experiments using Wikipedia

In the following 3 sets of experiments, we have used the English Wikipedia as a corpus for calculating the semantic distance between the hints and the geographical locations.

5.3.1 Normalized Wiki Distance (NWD)

In this set of experiments, we have used the Normalized Wiki Distance as formulated in Equation (2). The results of the 84 experiments show that the use of the 30 most salient (famous) countries in combination with a medium selection (5-15) of best scoring countries yields the best results. The highest result, yielded by the F30B10 experiment, shows an accuracy of 19.51% and a relative faster guessing performance of 56.25%. Table 2 displays the results of the 5 most accurate experiments in this series.

5.3.2 Probabilistic Distance (PD)

For this series of experiments, we have used the Probabilistic Distance metric as formulated in Equation (3). Similar to the results of the experiments using WordNet, the use of the 50 most salient (famous) countries in combination with a small selection (2-15) of best scoring countries gives the best results, with F50B2 topping the list with an accuracy 21.95% and a faster guessing performance of 55,55%. Table 3 displays the 5 best-scoring configurations.

5.3.3 PMI Distance

In this set of experiments, we have used the Pointwise Mutual Information measure as formulated in Equation (4). The results are in agreement with the majority of the previously recorded results and they show that the use of 20 most salient (famous) countries in combination with a medium selection (10-30) of best scoring countries gives the best success accuracy. The best-scoring configuration here is F20B10 with an accuracy of 23.17% and a faster guessing performance of 68.42%. The 5 best-scoring configurations are shown in Table 4.

5.4 Summary of Results

Overall, we have performed 84 experiments for each resource (i.e. WordNet and Wikipedia) and the associated distance measures. In total, this makes 336 different experiments (i.e. configurations using the *M* and *N* parameters). Table 5 summarizes the success rates of both WordNet and Wikipedia and all associated distance measures. According to the results, the maximum accuracy (23.17%) was reached using the PMI distance measure on the Wikipedia corpus. On the other hand, the highest mean of the accuracy throughout the different configurations was recorded for the PD measure, on the Wikipedia corpus as well.

In this section, we have only presented the best scoring configurations, but for the sake of completeness, the results of all experiments and configurations are visualized in Figure 2. This figure clearly visualizes which configurations (M and N values) are optimal for the different metrics.

6 Discussion

The results of hundreds of experiments demonstrate that using the Wikipedia corpus yields substantially better results than using WordNet as a resource for semantic distance calculation in our guesser agent. This might be due to the very nature of the word associations that the Taboo game requires. The format of the game already rules out the best clues, i.e. the most closely associated words, from the set of hints. This means that there is always a considerable distance between the two terms. WordNet has difficulties with this, as the annotated hierarchical relations are only made between terms that are semantically very closely associated, and paths that link hints to locations might not exist, or might not be very meaningful due to their length (of the link chain). The Wikipedia approach seems to be much more robust against this. Even if the hints are not that closely related to each other, there almost always exists documents on which hint and geographical location occur together. For this task, the size of Wikipedia has the upper hand over the precision annotation of WordNet.

Throughout the different configurations in our experiments, we observed that limiting the number of countries in the country list can improve the performance. As we mentioned earlier, this has the risk that some of the games will fail because their target location falls outside the list. On the other hand, it has the advantage that countries for which the hit counts are sparser do not influence the results too much. The results show that the NWD and PMI metrics benefit from limiting the number of countries to 20 or 30, whereas PD seems to be less disturbed by the sparseness effect. Indeed, PD benefits from configurations having higher numbers such as 50, 60 and ALL.

Once the countries have been ranked based on the metric, we also limited the number of countries for which the cities were considered (the parameter *N*). This also influences the performance differently from one distance measure to another. The PMI and NWD metrics score the best with higher *N* values (10-30), whereas the PD metric scores equally well with high (15) and low (2-3) *N* values. This indicates that the PD measure performs better at ranking the countries based on the hints.

7 Future Work

The research described in this paper is only a first step towards solving the ESSENCE LT Challenge. Using well-established word association techniques and freely available corpora, we aimed to establish a baseline to which future approaches can be compared. A first, promising extension of our guesser agent would be to equip it with machinery for resolving hints that are relative to the previous answer (e.g. *close*, or *north*). Another extension, which is closely related to the diverse nature of the real-world dataset, would be to model the associative behavior of the individual describers. This is possible, as with each game in the challenge, the ID of the human describer is provided. This way, the diversity in associations and game-playing behavior of the different players could be taken into account in order to improve the number of correct guesses. Further improvements could include investigating how lemmatization of the hints influences the accuracy of the guesser agent, as well as to explore ways to fuse the different metrics that were described in this paper.

8 Conclusion

We have proposed a semantic distance based architecture for a guesser agent for the Essence Location Taboo Challenge. The proposed architecture employs a two-step approach, narrowing down

Table 5. Summary of accuracy results for each resource and distance measure

Corpora	Distance Metric	Accuracy (%)			
		Maximum	Minimum	Mean	Standard Deviation
WordNet	Jiang-Conrath	6.09	0	2.06	1.35
	NWD	19.51	8.53	15.36	2.05
Wikipedia	Probabilistic	21.95	13.41	17.16	2.18
	PMI	**23.17**	7.31	15.15	3.62

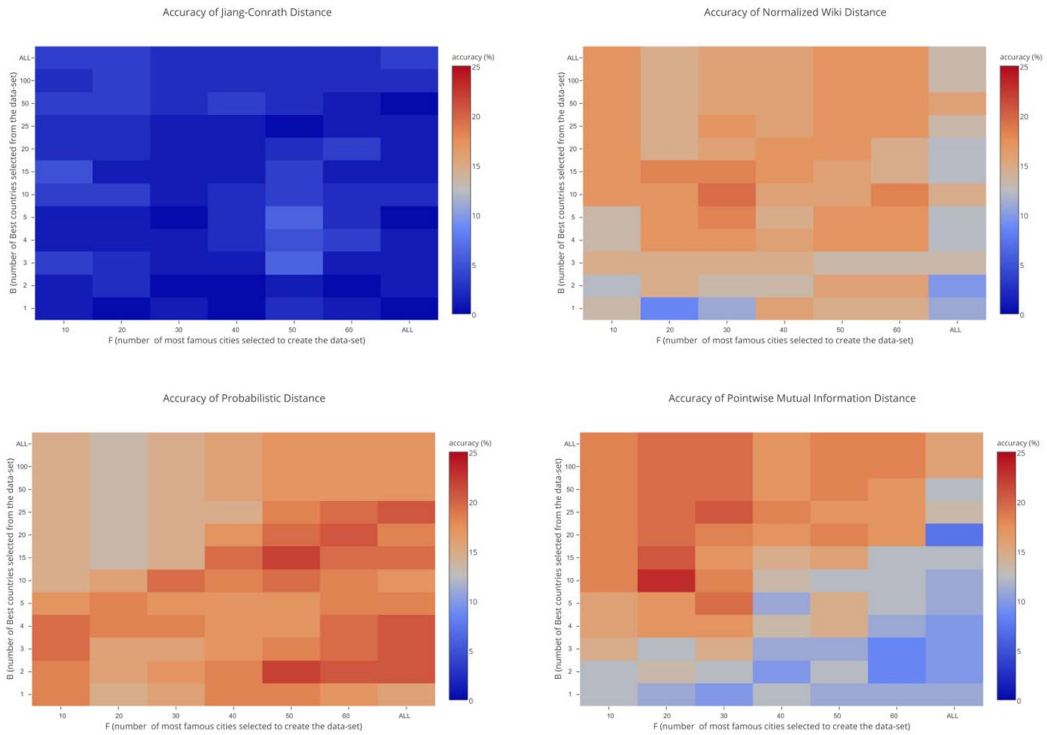


Figure 2. Results of all experiments. The X axis represents the M parameter (most salient countries) and the Y axis represents the N parameter (cities of N best countries considered). The red-blue scale indicates the accuracy of the experiment.

the geographical area of the guess first to the country and then to the city. We have explored different resources and metrics for measuring the diverse associations between the hints and the geographical locations that were made by human players with different backgrounds. The highest score with 23.17% accuracy and 68.42% of faster guessing performance was achieved with the PMI measure applied to the Wikipedia corpus. Although this research is only a first step to model the diversity in word associations that individual humans exhibit, it can serve as a strong baseline to which future attempts to solve the ESSENCE LT Challenge can be compared.

ACKNOWLEDGEMENTS

The authors are grateful to Khuyagbaatar Batsuren for his contributions to discussions on web search engines and related APIs. All authors of this paper were funded by the Marie Curie Initial Training Network (ITN) Essence, grant agreement no. 607062.

REFERENCES

[1] Kemo Adrian, Khuyagbaatar Batsuren, Nicola Bova, Thomas Brochhagen, Paula chocron, Paul Van Eecke, Mercedes Huertas-Miguelanez, Mladjan Jovanovic, Tania Marques, Julian Schloder, and Aimilios Vourliotakis, 'Taboo challenge', Technical report, Essence Marie Curie Initial Training Network, (06 2015).

[2] Gerlof Bouma, 'Normalized (pointwise) mutual information in collocation extraction', *Proceedings of GSCL*, 31–40, (2009).

[3] Alexander Budanitsky and Graeme Hirst, 'Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures', in *Workshop on WordNet and Other Lexical Resources*, volume 2, pp. 2–2, (2001).

[4] Kenneth Ward Church and Patrick Hanks, 'Word association norms, mutual information, and lexicography', *Comput. Linguist.*, **16**(1), 22–29, (March 1990).

[5] Rudi L Cilibrasi and Paul Vitanyi, 'The google similarity distance', *Knowledge and Data Engineering, IEEE Transactions on*, **19**(3), 370–383, (2007).

[6] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman, 'Indexing by latent semantic analysis', *Journal of the American society for information science*, **41**(6), 391, (1990).

[7] David Ferrucci, 'Build watson: an overview of deepqa for the jeopardy!

challenge', in *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*, pp. 1–2. ACM, (2010).
[8] David Ferrucci, 'Build watson: an overview of deepqa for the jeopardy! challenge', in *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*, pp. 1–2. ACM, (2010).
[9] Don Heath, David Norton, Eric Ringger, and Daniela Ventura, 'Semantic models as a combination of free association norms and corpus-based correlations', in *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on*, pp. 48–55. IEEE, (2013).
[10] Jay J. Jiang and David W. Conrath, 'Semantic similarity based on corpus statistics and lexical taxonomy', in *Proc of 10th International Conference on Research in Computational Linguistics, ROCLING97*, (1997).
[11] Kevin Lund and Curt Burgess, 'Producing high-dimensional semantic spaces from lexical co-occurrence', *Behavior Research Methods, Instruments, & Computers*, **28**(2), 203–208, (1996).
[12] George A Miller, 'Wordnet: a lexical database for english', *Communications of the ACM*, **38**(11), 39–41, (1995).
[13] Eli Pincus, David DeVault, and David Traum, 'Mr. clue-a virtual agent that can play wordguessing games', in *Proc. of the 3rd Workshop on Games and NLP (GAMNLP), Raleigh, North Carolina, USA*, (2014).

Interdisciplinarity as an Indicator of Diversity in a Corpus of Artificial Intelligence Research Articles

Bilge Say [1]

Abstract. The preliminary results of a corpus based study of interdisciplinarity on Artificial Intelligence (AI) on a corpus of AI research articles are presented based on an annotation scheme on nature of interdisciplinarity. I argue that the nature of interdisciplinarity of AI as seen in those articles are rather limited, where interdisciplinarity is only noted when contextualizing the problem for more than half of the articles that have an interdisciplinary interaction. Where more than one method is used, all methods used already belong to AI repertoire in most cases. Rarely a more distant method is used such as a method with a cognitive origin. These findings indicate a highly specialized approach pertains in AI journals, which is to be further verified through an extended study.

1 INTRODUCTION

The definitions and analogies used for interdisciplinarity carry clues as to how diverse a given field or specialty is, in a number of ways. Taking into account a common definition of interdisciplinarity as given by US National Academies' report [1], one can further clarify the relationship: "Interdisciplinary Research (IDR) is a mode of research by teams or individuals that integrates information, data, techniques, tools, perspectives and theories from two or more disciplines or bodies of specialized knowledge to advance fundamental understanding or to solve problems whose solutions are beyond the scope of a single discipline or research practice" (p. 27). The entities in this definition all form basis for getting diversity in a given piece of research. Although not all kinds of diversity entail interdisciplinarity; the reverse case—the existence of interdisciplinarity—hints at the existence of multiple forms of data, methods or theoretical frameworks that would entail diversity.

Other literature on interdisciplinarity seem to support this idea. Lyall et al. [2] distinguish *focused interdisciplinary research* which creates knowledge for a specific complex goal or problem possibly with societal or industry involvement and *academic oriented interdisciplinary research* where the focus is on learning and not expertise. In some cases these seemingly orthogonal dimensions can coexist, bringing about diversity to the research enterprise. Thagard [3] uses the notion of *trading zones* from anthropology where goods are exchanged—reminiscent of a diverse, vivid market place scene—by seemingly distant research communities to describe the emergence of interdisciplinarity in AI and Cognitive Science.

The aim of this ongoing study is to carry out a corpus based analysis of interdisciplinarity on a set of Artificial Intelligence (AI) research articles in established journals and, in the long run, evaluate diversity as a possible function of interdisciplinarity.

2 INTERDISCIPLINARITY WITHIN ARTIFICIAL INTELLIGENCE

Modern birth of AI along with Cognitive Science involved a lot of interdisciplinarity and diversity both in the individual interests of the founding figures and the kind of work they produced [3,4]. How can AI be characterized considering recent periods? In Web of Science subject categories, AI is seen as a subcategory of Computer Science [5]. AI's current status is seen as an "integrated speciality" in comparison with the development of Cognitive Science in [6]. Cognitive Science has had a domination of certain fields such as cognitive psychology in its developmental trajectory but has been commented to be, currently, more diverse than AI in interaction of fields and methods [6]. In a not-so-recent survey [7] in 1985, AI researchers put forward varying ideas on whether AI was a cluster of specialities engaged in intelligence and cognition, or a maturing speciality, or a set of methods under an umbrella, among definitions of AI. Van den Besselaar and Leydesdorff's [8] work on aggregated journal-journal citations have shown that between 1982 and 1992, AI has evolved to show disciplinary trends with a relatively stable set of both applied and fundamental journals, mainly with a domination of computer science with other specialites such as bioinformatics interacting. AI was found to be not closely related with Cognitive Science, not an unstructured field of philosophical questions nor a collection of specialities studying "intelligence" in comparison with some of the claims stated in [7]. The claim that AI is mainly related with computer science and does not have multiple interdisciplinary research bases have also been supported in [9].

Bibliometric and survey-based studies are surely valuable in discovering the current interdisciplinary characteristics of AI. However, complementing such methods with corpus based annotation where discourse and methods of individual articles are taken into account by expert annotators can also be valuable. This work is a first pass at such an attempt. The rest of the paper is organized as follows: in Section 3, the selection process of journals and articles as well as the annotation scheme is summarized. The results of the work carried out so far with possible implications for diversity is given in Section 4, along with future work plans in Section 5.

[1] Department of Digital Game Design, Ipek University, Ankara, Turkey, email: bsay@ipek.edu.tr

3 METHOD

3.1 Journal Selection

The current preliminary study is carried on 20 randomly[2] selected research articles from four Artificial Intelligence journals, 5 articles from each journal. Ultimate aim is to expand this corpus so that statistical significance results can be given. Issues were published between 2013-2014 so that they are recent enough but amenable to bibliometric citation analysis. Journals chosen are indexed both in Web of Knowledge and Scopus—two major reputable indices—and have been issued for continuously for the past 10 years. All journals have the term "Artificial Intelligence" in their title and are rather general in coverage. There are 14 journals in all with titles containing "Artificial Intelligence" among the 130 journals in Journal Citation Reports' "Computer Science: Artificial Intelligence" category. The median impact factor for all 130 journals is 1.403; for the 14 journals, 0.972; and for the four selected ones 1.681 (for year 2015). However, this selection of journals is rather arbitrary, and a bibliometric pre-study could better determine where AI community publish the most.

Table 1. Journals used in the AI Corpus

Journal Name	Published Since	Impact Factor	SNIP	SJR
Artificial Intelligence	1970	3.371	5.192	3.263
AI Magazine	1980	0.595	2.368	1.049
J. of Artificial Intelligence Research	1993	1.257	2.607	1.725
J. of Experimental & Theoretical Artificial Intelligence	1989	1.000	0.978	0.474

SNIP: Source Normalized Impact per Paper; SJR: SCIMago Journal Rank; All metric values are of 2014; journal names are alphabetically ordered.

Table 1 shows selected journals' names and journal metrics, where further information on journal metrics and Journal Citation Reports can be found in [10,11]. All articles sampled are research articles; reviews and editorial notes are not included.

3.2 Interdisciplinarity Annotation Scheme

The interdisciplinarity annotation scheme is based on Huutoniemi et al.'s work [12], with two extensions to be explained in the next paragraph. Particular choice of Huutoniemi and colleagues' classification is justified on the grounds that it is based on an empirical classification of a corpus of research project proposals from various disciplines. In addition, the schema is based on integrating multiple perspectives from the previous literature on interdisciplinarity research. All the feature types and values of annotation scheme—also be presented in tabular form in Section 4—will be described below.

Interaction Type forms the focus of the annotation scheme having six possible feature values for interdisciplinary interaction arranged in a spectrum. If no interaction is observed, the value is marked as *none*. *EncyclopedicMD*[3] is for when there is a topical, rather encyclopedic interdisciplinary discourse with no cognitive nor empirical interaction. This feature value is more appropriate for broad project topics where only the topic is serving as a glue for disparate works of research. *ContextualizingMD* is an interdisciplinary interaction which is limited to problem setting only but not on the actual work carried out. Using an AI method that will help present solutions to problems in another domain is such an interaction. Only the context of the interaction is explained in *ContextualizingMD* when observed as an interaction type. *CompositeMD* is a modular, borrowing interaction where the boundaries of the interaction are well defined, and the interaction is not tightly coupled. It is beyond just stating an interdisciplinary context; multiple specialities or disciplines may work separately but in interaction through the research or production process. *EmpiricalID* is where multiple kinds of empirical data are analysed, e.g. lingustic data along with neuroimaging data modeled in an AI agent in a single study. *MethodologicalID* is where multiple methods of different origins are combined in a novel and integrated way. It is to be noted that this may be different than using two AI methods so tightly coupled together that a hybrid method is in actual use. Although hybrid methods versus interdisciplinary use of multiple methods is attempted to be distinguished in this study, it may not be possible to categorically separate such uses. Methodological interdisciplinarity may in fact be precursory to the birth of a hybrid method. *TheoreticalID* aims to reach a synthesis by deriving new concepts, models or theories from the interdisciplinary interaction. Interaction types above, as listed in Table 8, are hierarchical in the sense that stronger interaction types can subsume weaker ones so only one feature value is chosen for each article in the corpus.

Scope is the conceptual and methodological distance between the interacting disciplines or specialties, *narrow* for close fields, *broad* for distant ones, if there is an interdisciplinary interaction. Since no a priori interdisciplinarity is assumed, scope can be *none* as well. Classification of Research & Development fields is loosely taken as a guide for distance judgement [13]; the author's subjective judgement is used complementarily. *GoalType* is for characterization of the aim of the research. It can be *epistomologically_oriented* in the sense that the aim is to contribute to the related knowledge, including concept, methods and theories in a novel way or *instrumentally_oriented* which aims to solve pragmatically a complex or challenging problem. Research can naturally be marked with both aims, which creates an additional *mixed* category.

Two feature types are added to Huutoniemi et al's scheme within the current research. The two *methods*—major and minor, if existent—of the research article being examined is annotated with a two level classification: whether the method falls within an AI or cognitive orientation or neither; and a specific labeling of the method for the first two cases. The method family used can be examined in next section and is collected from the current literature and textbook knowledge [14,15]. It can, of course, be argued that a specific method such as cognitive computational modeling is not specific to cognitive science but can be characterized as an AI method or can use an AI method such as neural networks. The distinction was made considering the main usage of the method family in general. In cases such as cognitive

[2] Random selection of article samples are achieved by numbering of eligible research articles (i.e. excluding reviews, notes etc.) and their sampling through http://www.random.org.

[3] All categories, where present, are spelled in the same way as in [12] with the exception of removal of asterisks from the entries.

modeling with neural networks, a second (minor) method annotation is done to denote the co-use. If a second method is annotated, another feature, *method use*, characterizes whether the method's use is an *interdisciplinary* use. If both methods come for an AI origin and are highly integrated, a *hybrid* use is marked. The method use is marked as *discussion only*, where there is no actual application of the method but a comparative discussion. The last feature type called *DomainType*, is marked again for two data types—major and minor—used in the research (if more than one data type is used) as to whether it is *human* data, (collected behaviourally or otherwise, e.g. linguistic data from the web); *machine* generated, *animal* data, other type of data from the *real world* or *logical and mathematical* statements only. The distribution of the features values will be given in the next section.

3.3 Coding

Coding of the articles according to the scheme above is done using UAM Corpustool [16]. Currently only the author, with previous experience in AI and Cognitive Science, did annotation judgements, which is a limitation to be overcome in further phases of the study[4]. Each article is read, particular interdisciplinary-flavored sentences are marked as segments, and the document level judgement is based on the general reading of the articles and the discourse of the segments, which takes about an hour per article.

4 RESULTS AND DISCUSSION

In the ongoing work, the results are too few for statistical analysis; hence frequencies will be presented alongside discussion points for each feature type.

4.1 Methods

As can be seen in Table 2, most of the articles used a dominating (major) method belonging to the AI family of methods. It was a rarer case when the major method was from a cognitive orientation or the researchers borrowed a method from (usually) a neighboring discipline such as Operations Research or Electrical Engineering.

Table 2. Method Types Used (Major)

Method (Major)	Frequency (N=20)	
AI	16	80.00%
cognitive	2	10.00%
other	2	10.00%

Nearly half the articles in the corpus did not involve the use of a second method as can be seen in Table 3. In the case they did, such methods again were mostly characterized as AI methods. Other methods were used in lesser frequency.

[4] A graduate student in Cognitive Science will annotate the same corpus independently and interrater agreements will be presented.

Table 3. Method Types Used (Minor)

Method (Minor)	Frequency (N=20)	
none	8	40.00%
AI	8	40.00%
cognitive	1	5.00%
other	3	15.00%

In the case that a secondary method was used, most of such uses were of hybrid nature, jointly using two AI methods for improvement of a certain set of metrics on a given problem, as shown in Table 4. Some uses involved a comparative discussion of previous results from a second method and not actual use, and only %15 of the total article set involved the use of a second method that contributed to the diversity of the article by explicitly creating an interdisciplinary interaction such as behavioural experimentation or cognitive modeling.

Table 4. Minor Method Use

Method Use (Minor)	Frequency (N=20)	
none	8	40.00%
hybrid	6	30.00%
interdisciplinary	3	15.00%
discussion only	3	15.00%

AI methods used can be said to be quite dispersed (Table 5) both for major methods and minor methods. Statistical methods dominate, which is not surprising, given recent trends in AI [14,15]. A more topical distribution can also be made in future with marked subspecialties of AI (planning, natural language processing, etc.) to get a picture of diversity with respect to subspecialties.

Table 5. AI Methods Used

AI Method	Major/Frequency (N=16)		Minor/Frequency (N=8)	
a-life-other	0	0.00%	1	12.50%
case-based-reasoning	0	0.00%	0	0.00%
constraint-programming	1	6.25%	1	12.50%
evolutionary-algorithms	1	6.25%	1	12.50%
fuzzy-logic	0	0.00%	0	0.00%
knowledge-based	4	25.00%	1	12.50%
machine-learning-other	2	12.50%	1	12.50%
neural-networks	0	0.00%	0	0.00%
probabilistic-statistical	6	37.50%	2	25.00%
reinforcement-learning	0	0.00%	0	0.00%
search-heuristic-methods	2	12.50%	1	12.50%

Cognitive oriented methods used were few both in major and minor methods (Table 6). This might be a limitation of the particular set of articles as well as the particular set of journals in this preliminary study.

Table 6. Cognitive Methods Used

Cognitive Method	Major/Frequency (N=2)		Minor/Frequency (N=1)	
behavioral	0	0.00%	1	100.00%
computational-modeling	1	50.00%	0	0.00%
experimental-other	0	0.00%	0	0.00%
neuroimaging	0	0.00%	0	0.00%
linguistic-analysis	0	0.00%	0	0.00%
argumentative	1	50.00%	0	0.00%
survey	0	0.00%	0	0.00%

4.2 Domain

The data domains used can be said to be quite diverse (Table 7). Human data was mostly natural language utterances from web or data sets; other real world data came from the problem context. Papers that worked on mathematical or logical representations only, were not negligible either.

Table 7. Data Domains Used

Data Domain	Major/Frequency (N=20)		Minor/Frequency (N=9)	
human	6	30.00%	2	10.00%
machine	4	20.00%	1	5.00%
animal	0	0.00%	0	0.00%
real-world-other	5	25.00%	2	10.00%
logic-maths	5	25.00%	4	20.00%

4.3 Interdisciplinary type and scope

The main focus of the article, the diversity resulting from interdisciplinarity, is to be seen in Table 8. More than half of the interactions are contextualizing the problem domain, where a given method or its improvement help the solution of a problem in another domain that is usually close (Table 9). In fewer but non-negligible number of cases there was more interdisciplinarity, where either a novel, integrated method was developed or contextual interaction was in constant dialogue throughout the research process e.g. not only as a justification statement in "introduction" part of an article, but via active evaluation in the problem domain.

Table 8. Interdisciplinary Interaction Type

Interdisciplinarity Type	Frequency (N=20)	
no-interaction	3	15.00%
encyclopedicMD	0	0.00%
contextualizingMD	11	55.00%
compositeMD	4	20.00%
empiricalID	0	0.00%
methodologicalID	1	5.00%
theoreticalID	1	5.00%

However, Table 9 indicates most diversity is in narrow fields where conceptual assumptions and methods seem to be already similar. Only in two cases interdisciplinarity interaction was observed between distant fields; both articles were in AI Magazine.

Table 9. Scope of Interdisciplinary Interaction

Scope	Frequency (N=20)	
narrow	15	75.00%
broad	2	10.00%
no-interaction	3	15.00%

4.4 Goal

Finally, as can be seen in Table 10, more than half of the articles in the corpus have explicit problem-oriented goals; fewer, theoretical papers make contributions to the mathematical/logical knowledge without necessarily undertaking the applied implications. Fewer still have something to achieve in both arenas. Goal types can be said to be diverse in this respect.

Table 10. Goal of the Research Article

Goal	Frequency (N=20)	
epistemologically-oriented	6	30.00%
instrumentally-oriented	11	55.00%
mixed	3	15.00%

5 CONCLUDING REMARKS AND FUTURE WORK

The main remark that might be drawn from the limited corpus is that the majority of interdisciplinary interaction is limited; often the context of the interaction is stated but rest of the article is rather

specialized. Generally, a mention of the context, where the proposed improvement on an AI method or concept can be useful, is given. In some cases, evaluation of the study is carried out on public data sets, but rarely evaluation is carried out using diverse methods from multiple fields. In more theoretical papers, after an initial mention of context, rest of the mathematical discourse with proofs entirely remain within a mathematical/logical framework. Only in one case a philosophical/cognitive subject, was the main subject and formed the argumentative discourse of the article. This observation is inline with previous work [6,8]; indicating disciplinary and integrated speciality trends. With an extended corpus of different AI journals and articles, it may be also be possible to distinguish between article level interdisciplinarity versus journal level interdisciplinarity. With the tentative data at hand, it may be possible to imply AI Magazine, the journal for Association for the Advancement of Artificial Intelligence (AAAI), can be said to have more interdisciplinary interaction in methods and data, and broader scope in those interactions, therefore more diversity because of the applied academic role it carries for general public and general AI community. As such the diversity of the methods, domains and interactions styles of the articles forming the corpus seem somewhat limited. The link between interdisciplinarity and taking full advantage of actual potential diversity in intelligent systems is yet to be clarified further. It would be interesting to observe whether the nature of interdisciplinarity and diversity is different, if other types applied outputs of AI projects are examined, e.g. project proposals in AI, major conference proceedings and intellectual property applications.

In near future, the corpus will be enlarged with enough articles from the same journals to be able to carry out statistical analysis and a significant portion of the corpus will be annotated by a second annotator to be able to evaluate interrater agreement. Further work is planned for two parallel avenues: One is to bibliometrically calculate the degree of interdisciplinarity of the references of all the articles published in the journal set between 2013 and 2014. This can be done through bibliometric data mining using Rafols-Porter integration score [17], a variant of Rao-Stirling Diversity Index [18] with Web of Knowledge Subject Categories used as subject areas. This measure calculates the number and the diversity of different subject areas in a corpus of articles' references (termed variety and disparity respectively), as well as how balanced the distribution of references are in the subject areas referred. Such a bibliometric study can reveal better whether interdisciplinarity in citations of Artificial Intelligence articles is narrow and local (between close disciplines, with a disciplinary focus, characterized by relatively higher variety with lower balance and disparity) or distal and broad (between far-away disciplines, characterized by on relatively high balance and high disparity) [19]. Such a bibliometric study can be repeated on a larger pool of AI journals, e.g. the 130 journals characterized under Artificial Intelligence in Journal Citation Reports [11].

Another avenue is to repeat the same study both bibliometrically and by way of corpus annotation on a corpus of research articles for major Cognitive Science journals and carry out a comparative evaluation. It has been previously claimed that Cognitive Science and Artificial Intelligence have completely separate citation networks, that is, they do not cite references from each other's reference networks [6]. As well as validating whether such a claim holds on a different corpus of articles, the corpus analysis part of such a study will let us know whether two fields' diversity as characterized by their type of interdisciplinarity interactions and use of multiple methods differ. Moreover, the interaction of Cognitive Science and AI is worth examining as to how much two fields bring diversity to each other.

REFERENCES

[1] Committee on Facilitating Interdisciplinary Research, Facilitating Interdisciplinary Research. Washington, US: National Academies Press, 2004.

[2] C. Lyall, A. Bruce, J. Tait, and L. Meagher, Interdisciplinary Research Journeys: Practical Strategies for Capturing Creativity. Bloomsbury Publishing PLC, 2011.

[3] P. Thagard, 'Trading Zones in Cognitive Science,' in Interdisciplinary Collaboration: An Emerging Cognitive Science, Sharon J. Derry, Morton Ann Gernsbacher, and Christian Schunn, Eds. Informa, 2005, pp. 316–339.

[4] AI Topics, Brief History. Retrieved June 10, 2016 from http://aitopics.org/misc/brief-history, n.d.

[5] Web of Science Subject Categories. Retrieved June 10, 2016 from http://incites.isiknowledge.com/common/help/h_field_category_wos.html , n.d.

[6] L. Leydesdorff & R.L. Goldstone. 'Interdisciplinarity at the journal and specialty level: The changing knowledge bases of the journal cognitive science', *Journal of the Association for Information Science & Technology* **65,** 164–177 (2014).

[7] D. G. Bobrow & P. J. Hayes. 'Artificial intelligence — Where are we?', *Artificial Intelligence* **25,** 375–415 (1985).

[8] P. van den Besselaar. & L. Leydesdorff. 'Mapping change in scientific specialties: A scientometric reconstruction of the development of artificial intelligence', *J. Am. Soc. Inf. Sci.* **47,** 415–436 (1996).

[9] Y. J. Khawam. 'The AI interdisciplinary context : single or multiple research bases?', *Library & information science research* **14,** 57–74 (1992).

[10] S. Haustein. Knowledge and Information : Multidimensional Journal Evaluation : Analyzing Scientific Periodicals beyond the Impact Factor. (De Gruyter Saur, 2012).

[11] 2015 Journal Citation Reports® (Thomson Reuters, 2016)

[12] K. Huutoniemi, J. T. Klein, H. Bruun & J. Hukkinen. 'Analyzing interdisciplinarity: Typology and indicators',*Research Policy* **39,** 79–88 (2010).

[13] OECD. *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*, OECD Publishing, Paris.
DOI: http://dx.doi.org/10.1787/9789264239012-en (Organisation for Economic Co-operation and Development, 2015).

[14] S. Russell & P. Norvig. *Artificial Intelligence: A Modern Approach*. (Pearson, 2009).

[15] Z. Shi. *Series on Intelligence Science: Advanced Artificial Intelligence.* (World Scientific, 2011).

[16] M. O'Donnell. 'The UAM CorpusTool: Software for corpus annotation and exploration', in *Proceedings of the XXVI Congreso de AESLA*, (2008).

[17] A. L. Porter, D. J. Roessner & A. E. Heberger. 'How interdisciplinary is a given body of research?', *Research Evaluation* **17,** 273–282 (2008).

[18] A. Stirling. 'A general framework for analysing diversity in science, technology and society', *Journal of The Royal Society Interface* **4,** 707–719 (2007).

[19] D. Chavarro, P Tang. & I. Rafols. 'Interdisciplinarity and research on local issues: evidence from a developing country', *Research Evaluation* **23,** 195–209 (2014).

Managing human diversity in diverse multi-agent collaborative intelligence systems

Mark Hartswood and Kevin Page and Avi Segal and Ya'akov (Kobi) Gal and Marina Jirotka and Ronald Chenu Abente Acosta[1]

Abstract. This position paper is concerned with forms of diversity in collaborative intelligence systems involving collectives of human and machine agents working together to achieve both individual and global goals The paper draws on a case study in citizen science to examine diversity from sociological and economic perspectives and explore the dynamics that arise between diverse groups of participating human and computer agents. It argues that within *collaborative intelligence* systems that complex social activities of diverse groups of humans play a significant role in producing the overall 'intelligence' of the system. Finally we propose guidelines for encouraging and managing diversity within collaborative intelligence systems.

1 INTRODUCTION

Understanding how to motivate participation is an important aspect of social computations [12] and collaborative forms of intelligence [4] and these aspects are often considered from an economic or game theoretic perspective [8] where the incentives are taken to be important levers to drive and shape participation (e.g. [1][6]). An alternative perspective, emerging from the field of Human Computer Interaction (HCI), explores more thoroughly the social dynamics and circumstances of participation, offering a broader range of strategies beyond incentives for helping collaborative intelligence platforms to flourish (e.g. [3][5][10]). Our contribution is to unpick some of the social dynamics visible in collaborative intelligence platforms that arise due to the diversity of participating groups. In particular, we draw on two case studies to explore how system design and active management by platform operators permits beneficial interactions between groups who contribute in different (but crucial) ways and take on different (but significant) roles within the platform.

Our paper may seem a little incongruous at an Artificial Intelligence (AI) workshop as it does not address the design of AI concepts or tools directly. We justify this by noting that in many contexts AI operate within as wider system as part of an overall 'collaborative intelligence' that depends on human collectives working in synergy with intelligent machine agents [4]. Our contention is that within these systems then design of the *overall intelligence* has to account for the wider social environment in which the AI is but one element. Despite being an essential substrate of collaborative intelligence platforms, human diversity is also challenging to foster and maintain (e.g. [9]). The purpose of this paper is to provide some initial conceptual tools and guidance for nurturing and sustaining human diversity in these contexts. Our paper addresses two main issues in relation to diversity: (1) How do we come to understand interactions between the diverse groups that are naturally attracted to collaborative intelligence platforms; and (2) How can that diversity be managed so that participating groups remain harmonious and do not fragment too readily due to conflict?

2 CASE STUDIES

To date we have studied sharing economy platforms such as Uber, viewing these as distributed forms of collaborative intelligence. In the case of Uber, human agents (drivers) and computer agents (algorithms that measure and predict demand) 'collaborate' to solve a global resource allocation problem. There are powerful economic drivers in this type of system that motivate the participation of drivers and passengers, but also drive actions of Uber (the platform owner), and in particular, how Uber configures the interplay of the intelligences within the system. For example, when requesting a lift, passengers are shown the locations of nearby drivers on a map, but this information is withheld from drivers themselves, as it is not reproduced on the separate app used by drivers. This prevents the drivers using their own intelligence to collectively manage how supply within the system is organized, which they could do by adjusting their own local position on the basis of information about how supply is distributed globally. Thus the potential of Uber to function as a collaborative intelligence system will depend upon subtle configuration decisions as well as any explicit supportive mechanisms[2]. Allowing a greater contribution of driver intelligence also admits a greater influence of driver agency over the system, which in turn implies a greater expression of driver interests and even a different apportioning of economic benefits. We see these aspects of intelligence, agency, interests and economics to be intimately connected, and so any analysis of a collaborative intelligence system needs to consider these elements together. If we consider any contribution to collaborative intelligence as having consequences in relation to the diverse interests of the other participants, we can see how such

[1] Mark Hartswood and Marina Jirotka. Computer Science, University of Oxford, U.K. mark.hartswood@cs.ox.ac.uk, marina.jirotka@cs.ox.ac.uk. Kevin R. Page. Oxford eResearch Centre, University of Oxford, U.K kevin.page@oerc.ox.ac.uk. Avi Segal and Ya'akov (Kobi) Gal. Ben-Gurion University, Israel. avise@post.bgu.ac.il, kobig@bgu.ac.il. Ronald Chenu Abente Acosta, chenu@disi.unitn.it

[2] As a workaround, drivers may log into the passenger app on a second separate phone to access this information.

systems entail political as well as economic dimensions that are crucial for how diversity is managed.

While discussion of Uber helps lay out some of the issues discussed in this paper, our main case study relates to the work we have undertaken collaboratively with the citizen science platform 'The Zooniverse'. Citizen science (particularly within the Zooniverse) can be seen as a form of collective intelligence (whether it is also a form of *collaborative intelligence* is an issue we discuss later in this paper). The Zooniverse attracts interested volunteers to provide interpretations of scientific images, usually in the form of discrete annotations. Working as a crowd, volunteers are able to contribute many more annotations than scientists can accomplish unaided. Volunteers perform tasks that are still resistant to automation and achieve a high degree of accuracy when several annotations are combined. An ongoing issue for the Zooniverse is to understand its community – where they come from, what motivates them to participate; and particularly, how to understand the evident variations in types and degrees of the contributions made by its volunteers [3].

3 DIVERSITY IN MOTIVATION AND COMMITMENT

Participation in collaborative intelligence systems needs to be sufficiently diverse to provide the right mix of capabilities to sustain the operation of collaborative intelligence platforms. For example, Uber requires a diverse population of Uber drivers who are willing and able to drive at different times of the day to enable them to offer a comprehensive service. Similarly, the Zooniverse needs volunteers with a diversity of interests (to be attracted to the different projects) and has to be encouraging of any level of contribution – particularly since a bulk of annotations are made by the majority of visitors who contribute only a handful of annotations each.

Our first case study with the Zooniverse was directed towards understanding the reasons that volunteers engage with the Zooniverse, what leads them to disengage, and what might prevent them from re-engaging again in the future. To do this, we used a mixed method approach where we combined a qualitative survey of Zooniverse volunteers with statistical descriptions of participation to tease out an identifiable sub-population who may be susceptible to an intervention. Our survey results [11] echoed those of Eveleigh [3], who had identified a sub-group of 'Dabblers', who are characterized by a loose commitment to the platform, which they may sporadically renew. In our survey, respondents reported enjoying participating on the Zooniverse, but that they were prone to 'forget' about it as they become busy with other life and work related events. We hypothesized that for this group a reminder message would be effective at bringing them back to the site, and indeed we were able to demonstrate an effect from this type of intervention [11].

4 PARTICIPATION CURVES

We have further developed our approach to suggest the concept of 'participation curves' as a way to understand and characterize participation within populations comprised of diverse groups, and as a way to reason about interventions and the extent of their impact. In practice, a participation curve is simply a statistical view on some aspect of participation, such as those shown in figures 1 and 2. These curves demonstrate a 'power law' in the relationship between the time Zooniverse users are engaged in the system and the proportion of total users, showing how the majority of users are only active within the system for a very short time.

Different portions of the curve hint at different styles of interaction and different degrees of commitment. The shape of the participation curve can be viewed as a compositional representation of the activities of diverse sub-groups behaving in different ways.

As well as providing this integrated view of participation, participation curves also provide a scaffold for decomposing participation into more individualized categories of differently motivated groups of participants. Thus, we can read left-most side of the graph in figure 1 as comprising those large numbers of participants who stay in the system for a short while, while on the right hand side we see the smaller numbers of persistent volunteers.

At this point it becomes clearer how our qualitative understandings come to play as we forge connections between the kinds of participation implied by segments of the curve and the experiences of volunteers reported in our survey and reported in the literature. What emerges from this process is something very much like a series of 'personas' – sketches of types of user that are associated with particular stories or narratives explaining how a given group experiences the Zooniverse - which are each linked to different parts of the curve. In our concrete case we connected transient contributors with elaborate narratives around boredom, distraction and anxiety.

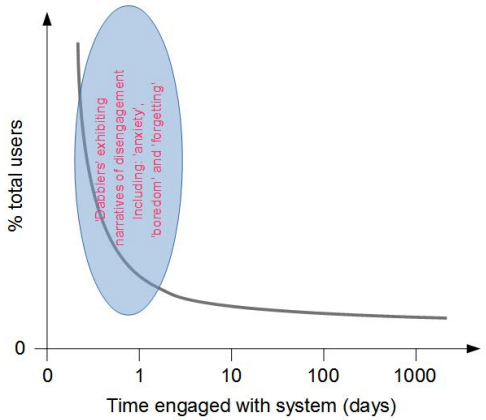


Figure 1. Conception/Diagrammatic view of a generalized participation curve

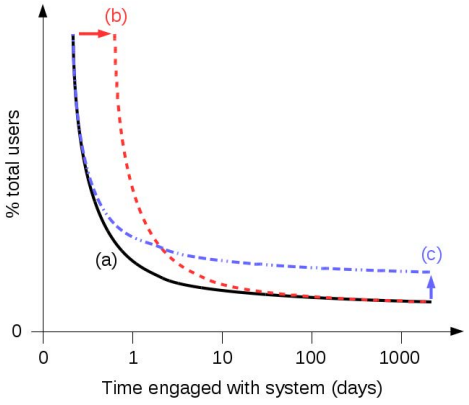


Figure 2. Local effects of interventions on participation curves

Another feature of participation curves is that they also serve to frame interventions and their likely effects. For example, our e-mail intervention was designed to encourage transient contributors to return. It acted modestly to transform the shape of participation by tugging a segment of the existing curve (Fig 2.a) into a new position (2.b). Since our intervention singles out a specific sub-group we might expect that its effects will be local. If our focus had been different, perhaps directed towards persistent volunteers, the effects of *that* alternative intervention would occur at a *different* segment of the participation curve, for instance, as indicated by the transition to fig 2.c.

Thus we regard participation curves as a gloss of much deeper and intricate group structures, such that the effort required in altering a curve is proportional to the resistance of the many real world obstacles experienced by specific groups 'responsible' for that portion of the curve. It is precisely the qualitative approaches, such as surveys (but also ethnographies and interviews) that allows us to dig into these circumstances and to help formulate relevant (but locally effective) remedies.

5 DIVERSITY IN TYPE OF CONTRIBUTION

Our second case study explores how, within the Zooniverse, scientists and diverse groups of volunteers participate together in complex and mutually sustaining networks of value coproduction. We use the concept of 'value coproduction' as a lens to understand community activity on the Zooniverse. The concept was developed in management science to account for the changing relationships between manufacturers and customers, largely brought about by new forms of interaction that the Internet, and social media in particular, have made possible. It recognizes the increasing role played by the customer in creating the value in goods and services. Examples include situations where customers assist with the configuration of the goods, perhaps by solving design problems as a crowd; or assist with marketing them by reviewing goods online; or play a role in helping others adopt those goods by sharing advice in online fora [13].

The 'economics' of coproduction differ in the Zooniverse as compared to monetized services such as Uber. The Zooniverse depends upon voluntary labor forces that are happy to contribute to a socially beneficial collective endeavor. The principal mode of exchange is around the annotation task, where volunteers benefit from an engaging experience and fulfillment through making a contribution to science. The scientists benefit from accumulating annotations which they can transform into scientific knowledge. Attached to this exchange are various obligations and commitments, not least those whereby the scientists are obliged to share credit for scientific discovery, and to 'account' for progress within the projects. These elements to the bargain can be seen in the various blog posts that scientists use to communicate with their volunteers.

As we have seen, volunteers exhibit diversity in the volume of classifications they contribute, the roles they adopt, their interests and motivations and their personal circumstances. To capitalize on this diversity, the Zooniverse has developed several arenas in which participants may create value in different ways for themselves and for each other. These arenas may be thought of as niches that diverse participating groups may colonize. The interfaces for classifying scientific objects (fulfilling the main objective of the site) are carefully designed to support the varied needs of participating groups. Different tasks from a range of scientific domains with varying levels of difficulty are offered to entice participants with varied interests and inclinations to take part. Classification interfaces are carefully designed to ensure low entry costs, enabling volunteers to begin classifying scientific objects as soon as they access the platform. This encourages the many transient visitors who contribute only a handful of classifications each. For groups who develop a deeper interest and stronger commitment, there are links to further arenas that invite alternative forms of participation. These include fora where findings may be discussed, tagged and shared, links to star catalogues where objects may be further researched, and links to social media where interesting images may be shared. In these ways, the Zooniverse caters for and encourages a diverse range of participating sub-groups.

These diverse sub-groups have different (but interconnected) roles in sustaining the Zooniverse. The large numbers of transient volunteers are a significant workhorse for amassing classifications. The fewer more dedicated volunteers contribute to fora and site management. This keeps the site vibrant and attractive, as well as providing bridges between the volunteer community, the science teams and platform maintainers. Various synergies exist and are encouraged between these diverse groups. Thus, while most participants will never post, many will benefit from those who do by passively engaging with the fora as a resource to solve problems or to seek inspiration. In this way, and in others, value is created and exchanged, or coproduced, between diverse participating sub-groups (as it is exchanged between the scientists and volunteers), in ways that form bonds and add to the cohesion of the overall community.

Managing this diverse ecosystem of participation is often tricky. Scientists have to carefully compose community messages so as to acknowledge the varied contributions of different groups so as not to create a hierarchy where one group may feel favored and another alienated.

Gamification strategies can have ambiguous outcomes in these complex participatory spaces, with competitive motivational devices, such as leader boards, being potentially divisive. One reason for this is that they privilege contributions based upon a single type of value (e.g. volumes of classifications), which distorts the delicate ecology of exchange in where many forms of value are traded in a balanced way across the platform.

A further issue is how *values* are consistent across the different uses of the value created by participation. For example, participation in the Zooniverse is driven by values attached to voluntarism, and contributions are expected to advance science and not to deliver (say) monetary gain.

Activities that seem peripheral to what might be perceived the 'central task' of producing annotations are actually highly important to sustaining the Zooniverse as an active and ongoing endeavor. Discussions over specific cases on Zooniverse project fora, for example, can be viewed as 'hidden work' that contributes directly to the overall quality of annotations and add to the 'overall quality' of the Zooniverse workforce.

6 DESIGN FOR DIVERSITY IN COLLABOATIVE INTELLEGENCE

The phrase Collaborative Intelligence has been coined to express the idea of shifting from the aggregation of anonymous, passive contributions that characterizes collective intelligence, towards instead an approach that supports and benefits from the diverse capabilities of identifiable human agents, each actively contributing towards the collective formation of a solution to a complex problem [4]. Thus within *collaborative* intelligence, the diversity of individual interests and perspectives of human agency is actively recruited and put to work, as opposed to the centralized control, and passive contributions as is perceived to be the case for *collective* intelligence systems (ibid). We argue that the diverse activities of Zooniverse participants around the core annotation task create a bridge from the Zooniverse as a collective intelligence platform, towards a Zooniverse that exhibits *collaborative intelligence*, as volunteers increasingly impart a stronger influence over the trajectory of the research and take on analytical roles that are closer to the domains of professional scientists.

Thus attracting and maintaining the contributions of diverse participating groups is a fundamental operational consideration for collaborative intelligence platforms. Considerations of how to do this need to go beyond simple notions of individual motivations, but instead they need to consider the dynamics of the interactions between diverse sub-communities and how these are managed by platform operators. We have distilled the following guidelines and strategies from our case studies in encouraging and sustaining diverse forms of participation in these arenas:

- **Algorithm design:** Understand how the contributing population is composed, for example, in terms of diversity of experience as revealed by participation curves. In the Zooniverse, understanding diversity is key to selecting appropriate statistics for aggregating classifications [2].
- **Community management:** Promote an ethos where different degrees and styles of contribution are each valued throughout the community. This is reflected in the style of 'community management' adopted within the Zooniverse that avoids creating a hierarchy where one style of contribution is perceived to be valued above another.
- **Experience design:** Sustain the participation of groups with diverse motivations and interests by creating a platform that has diverse opportunities for contributing, and supports diverse modes of engagement (e.g. casual to committed; contributor to community leader, etc).
- **Ecosystem design:** Attend to how diverse elements fit together to create a cohesive whole. Configure the platform to encouraged participants to work in mutually sustaining ways, e.g. by supporting them to create value for themselves and for each other, and through promotion of shared values.
- **Evolution:** As the population of participants grows, explore how alterations to platform elements or management approaches become necessary to support a growing and diversifying population. The Zooniverse platform did not emerge fully formed in the configuration we see it today. Instead, and over time, it has responded to the expanding expertise of sections of its community by incorporating features that cater for more experienced and self-directed forms of participation.

  Whilst we have not yet studied the decline of a platform, we suspect it would be wise to be alert to trends towards a shrinking or homogenized population and consider the effects this will have on the collaborative intelligence's ability to function.
- **Consider the limits to diversity:** Set limits to diversity by considering what sorts of and extents of diversity are desirable, and what forms of diversity would be hard to sustain. E.g. the Zooniverse is about undertaking scientific tasks, and adding gaming elements (to also attract those motivated by gaming) has proved to be problematic in the past. However, other approaches to Citizen Science wholly dedicated to gaming (e.g. FoldIt) seem to work well on their own terms.
- **Holistically:** Work towards an ecosystem that favors diversity via the interplay between all technical elements of the system – including the algorithms, incentives, reputation mechanisms, fora, and gamificaiton elements, by considering how these interact with the ethos, value creation and social conventions established within the collaborative intelligence community.

7 CONCLUSIONS

Through our examples we have aimed to show some of the complexity of managing a collaborative intelligence platform as a social and economic space where interactions between diverse groups of participants, often in arenas away from the 'core task', play crucial roles in sustaining and developing the 'intelligence' of the platform as a whole. We have shown how participation curves and the concept of value coproduction can be useful tools for exploring these social dynamics in ways that help frame possible interventions, while at the same time revealing the trade-offs and limitations of any given approach. Drawing on our experience of applying these tools within our Zooniverse case study, we have formed some initial high-level guidelines for managing ecosystems of diverse participating groups in collaborative intelligence platforms. These guidelines are not specific instructions, but are intended instead to provide an orientation or framing for the many design and operational decisions continually being made as a collaborative intelligence system evolves.

ACKNOWLEDGEMENTS

This work was supported in part by EU FP7 FET SmartSociety project (http://www.smart-society-project.eu/) under the Grant agreement n.600854. It was also supported in part by SOCIAM: The Theory and Practice of Social Machines, funded by the UK. Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J017728/1 and comprises the Universities of Southampton, Oxford and Edinburgh.

REFERENCES

[1] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. "Steering user behavior with badges." In Proceedings of the 22nd international conference on World Wide Web(WWW '13). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 95-106, 2013

[2] Darch, P. Managing the Public to Manage Data:Citizen Science and Astronomy. International Journal of Digital Curation, 9(1), pp. 25–40, 2014.

[3] Eveleigh, A., Jennett, C., Blandford, A., Brohan, P., and Cox, A., L.. Designing for dabblers and deterring drop-outs in citizen science. In

Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14). ACM, pp. 2985-2994, 2014.

[4] Gill, Z., User-driven collaborative intelligence: social networks as crowdsourcing ecosystems. In CHI'12 Extended Abstracts on Human Factors in Computing Systems. ACM. pp. 161-170, 2012.

[5] Haythornthwaite, Caroline. "Crowds and communities: Light and heavyweight models of peer production." In System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on System Sciences, pp. 1-10. IEEE, 2009.

[6] Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M. and Horton, J., February. "The future of crowd work." In Proceedings of the 2013 conference on Computer supported cooperative work ACM, pp.1301-1318, 2013.

[7] Luczak-Roesch, Markus, Ramine Tinati, Elena Simperl, Max Van Kleek, Nigel Shadbolt, and Robert Simpson. "Why won't aliens talk to us? Content and community dynamics in online citizen science.", ICWSM, 2014.

[8] Miorandi, D. and Maggi, L. "Programming" social collective intelligence. Technology & Society Magazine 33(3), pp. 55–61, 2014.

[9] Preece, J., and Shneiderman, B, The Reader-to-Leader Framework: Motivating Technology Mediated Social Participation. AIS Transactions on Human-Computer Interaction, (1)1, pp.13-32, 2009.

[10] Rotman, D., Preece, J., Hammock, J., Procita, K., Hansen, D., Parr, C., Lewis, D. and Jacobs, D. Dynamic changes in motivation in collaborative citizen-science projects. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, pp. 217-226, 2012

[11] Segal, Avi, Ya'akov Kobi Gal, Robert J. Simpson, Victoria Victoria Homsy, Mark Hartswood, Kevin R. Page, and Marina Jirotka. "Improving productivity in citizen science through controlled intervention." In *Proceedings of the 24th International Conference on World Wide Web Companion*, International World Wide Web Conferences Steering Committee, pp. 331-337, 2015.

[12] Smart, Paul, Elena Simperl, and Nigel Shadbolt. "A taxonomic framework for social machines." *Social Collective Intelligence*. Springer International Publishing, pp.51-85, 2014.

[13] Vargo, Stephen L., and Robert F. Lusch. "Service-dominant logic: continuing the evolution." *Journal of the Academy of marketing Science* (36)1, pp. 1-10, 2008.

[14] Von Krogh, Georg, Stefan Haefliger, Sebastian Spaeth, and Martin W. Wallin. "Carrots and rainbows: Motivation and social practice in open source software development." Mis Quarterly 36(2) pp. 649-676, 2012.

Analysing communicative diversity via the Stag Hunt

Robert van Rooij[1] and **Katrin Schulz**[2]

Abstract. What is the influence of the diversity of the targeted audience on how and what is communicated? Although Gricean pragmatics studies the effect of context on *what* is being communicated, the question *how* things are communicated is mostly ignored. Moreover, the impact of the size (and thus expected diversity) of the targeted audience is typically not addressed at all. In this paper we will study these questions making use of game theory, the theory of rational interaction. In particular we will argue he above questions can be addressed making use of insights gathered on the equilibria solutions of the Stag Hunt game.

1 Introduction

The original stag hunt game traces back to 1773, when Rousseau proposed the story of a stag hunt to represent a choice in which the benefits of cooperation conflict with the security of acting alone. In the story, two individuals must each choose to hunt a stag or to hunt a hare. Hunting stags can only be successful with cooperation, while hunting a hare does not require the other players help. The idea is that the stag offers both hunters a lot more meat than the hare. Thus, the stag hunt obliges a choice between productivity and security. Studies show that cooperatively hunting stag will only come out in case the trust between the partners is high.

Communication behaviour can be characterised as either safely (explicit and polite language use), or more efficient but also more risky (implicit, using e.g. irony) as well (Sally, 2003; van Rooij & Sevenster, 2006). Therefore, modelling these strategic communication choices in terms of the Stag Hunt game seems very natural. And also here it holds that the form of communication depends on the level of trust between the participants involved.

Language Typologists note that there is a difference between languages used by a lot of people (e.g., languages used as lingua franca) and languages used by smaller groups. The latter languages tend to be much more specific and complex both syntactically and semantically. In this paper we will also sketch how the diversity between the different languages (lingua franca versus not) can be explained making use of insights gathered on the Stag Hunt game.

2 The Stag hunt

Rousseau's Stag Hunt is described by Lewis (1969) as a simple two-player symmetric game with two strict equilibria (if both ϵ and ϵ' are higher than 0): both playing Risky (hunting Stag) or both playing it Safe (hunting Hare). It is obvious that equilibrium $\langle$Risky, Risky$\rangle$ is payoff-dominant.

Stag hunt:

	Risky	Safe
Risky	$1+\epsilon, 1+\epsilon$	$-\epsilon', 0$
Safe	$0, -\epsilon'$	$1, 1$

Following Harsannyi and Selten (1988), we will say that Nash equilibrium $\langle a^*, b^* \rangle$ is risk-dominant iff for all Nash equilibria $\langle a, b \rangle$ of the game,

$$(U_i(a^*, b^*) - U_i(a, b^*)) \times (U_i(a^*, b^*) - U_i((a^*, b)) \geq (U_i(a, b) - U_i(a^*, b)) \times (U_i(a, b) - U_i((a, b^*))$$

In the above example this is exactly the case for $\langle$Safe, Safe$\rangle$ if $\epsilon' \geq \epsilon$. For this reason, we will call a player *risk-loving* iff $\epsilon > \epsilon'$, he is *risk-neutral* iff $\epsilon = \epsilon'$, and he is *risk-averse* iff $\epsilon < \epsilon'$.

For the analysis in this paper it is useful to also consider the following non-symmetric variant of the Stag hunt game:

Stag hunt* :

	S_1	S_2
Risky	$1+\epsilon, 0$	$1-\epsilon', 0$
Safe	1,1	1,1

Also this game has two equilibria, $\langle$Risky, $S_1\rangle$ and $\langle$Safe, $S_2\rangle$, but in contrast to the previous game one equilibrium $\langle$Safe, $C_2\rangle$ is not a strict one: it doesn't matter what Column plays if Row plays Safe. If Row takes both Collumn strategies to be equally likely, the expected utility of playing Risky is higher/equal/lower than the expected utility of playing Safe if and only if $\epsilon > / = / < \epsilon'$, just as in the original Stag hunt.

3 Irony and metaphor

In western cultures we can think of the use of irony and metaphor as risky speech. In Easern cultures, instead, risky speech is being implicit: speaking with hints (where the likely costs are miscoordination, rather than irony, where the likely costs are insults). We can model the usefulness of irony in western cultures as follows, using two different types of audiences: intimates and strangers

Irony:

	Intimate	*Stranger*
Irony	$1+\epsilon, 1+\epsilon$	$-\epsilon', 0$
literal	$0, -\epsilon'$	$1, 1$

Te above payoffs can be at least partly explained by the following quote of Fowler:

> Irony is a form of utterance that postulates a double audience, consisting of one party that hearing shall hear and shall not understand, and another party that, when more is meant than meets the ear, is aware both of that more and of the outsiders'

[1] Institute for Logic, Language and Computation (ILLC), email: r.a.m.vanrooij@uva.nl
[2] Institute for Logic, Language and Computation (ILLC), email: k.schulz@uva.nl

incomprehension. [It] may be defined as the use of words intended to convey one meaning to the uninitiated part of the audience and another to the initiated, the delight of it lying in the secret intimacy set up between the latter and the speaker. Fowler (1965, pp. 305-306)

The conclusion is indeed that the use of irony is risky, but can be advantaguous.

4 Risk of implicit communication

Suppose that two meanings, t_1 and t_2, can be expressed literally by m_1 and m_2, respectively.. In addition, however, we have a lighter expression m_u whose use yields a bonus of $\epsilon > 0$ above the others. If the relative probabilities of t_1 and t_2 are not shared between speaker and hearer, the benefit of communicating with a light expression must be very high in order to overcome the *risk* of miscommunication. We are going to discuss a case like that of game below.

implicit :

	S_1	S_2
Risky	$1+\epsilon, 1$	$1-\epsilon', 0$
Safe	$1, 1$	$1, 1$

The Safe strategy is to send the correct explicit message in the relevant state, while the Risky strategy is to use the light message with the underspecified meaning. S_1 and S_2 are the strategies that interpret the explicit messages in the expected way, and they interpret m_u as t_1 and as t_2, respectively. Unsuccessful communication has a payoff of 0, i.e., we assume that ϵ' = 1; and the benefit of successful communication with the light underspecified expression m_u instead of the conventional explicit expression m_1 is ϵ, which is higher than 0. Notice that the speaker prefers to play Risky if she takes strategies S_1 and S_2 to be equally likely iff $\epsilon > 1$, (given that $\epsilon' = 1$).

The hearer interprets m_u as t_1 if he takes t_1 to be more likely, and he interprets m_u as t_2 if he takes t_2 to be more likely. The speaker's payoffs of these two strategies in the different situations are given by following tables:

t_1	S_1	S_2
Implicit	$1+\epsilon$	0
Explicit	1	1

t_2	S_1	S_2
Implicit	0	$1+\epsilon$
Explicit	1	1

The speaker doesn't know how the hearer will interpret the underspecified message m_u because she does not know whether the hearer will take t_1 or t_2 to be more likely. We have seen above already that if the speaker takes S_1 and S_2 to be equally likely, the benefit of using the underspecified message has to be at least 1, $\epsilon \geq 1$. But what if the speaker doesn't think strategies S_1 and S_2 are equally probable? Let us assume that the speaker believes with probability n that $P(S_1) > P(S_2)$ (and thus with probability $1-n$ that $P(S_1) \leq P(S_2)$). It follows that the speaker takes implicit communication to be worthwhile in situation t_1 if and only if $n \times (1+\epsilon) > 1$. That is, for the expected utility of being implicit to be higher than the expected utility of being explicit it has to be the case that $\epsilon > \frac{1-n}{n}$.

Obviously, if n is very close to 0 the use of m_u will be a bad choice, but also for other choices of n, it probably won't pay to be implicit: if n is $\frac{1}{3}$ or $\frac{1}{4}$, for instance, the value of ϵ has to be 2, or 3, respectively, which seems to be much too high.

Being explicit is a *safe* strategy. It is optimal under the maximin strategy and the minimax strategy. Things are more complicated when expected utility is at issue, for now it also depends on the relative weight of n and ϵ. But the obvious, conclusion is always that it is safer to be explicit if—because of diversity— you don't know (for sure) what your conversational participants takes to be the most salient situation of T, and that it is risky to be implicit.

5 Risky lying or not?

As another example of risky communication, we will consider under which circumstances it is advantageous to lie. In this example, the preferences are diametrically opposed for one choice of action of column player. The two players of the game are speaker (the row-player) and hearer (column-player). Suppose the speaker wants the hearer to perform a certain action, say a_1, but that it is commonly known between speaker and hearer that the latter will only perform a_1 if he thinks the speaker is of a high quality. Otherwise, hearer will perform action a_2 which the speaker disprefers to a_1. In fact, the speaker is not of a high quality. The hearer, however, doesn't know this, which gives the speaker the possibility to mislead her conversational partner by lying about her quality. Thus, the speaker has two strategies: she either is honest, or lies about her quality. We assume that the hearer will always perform action a_2 in case the speaker is honest about his low quality. However, in case the speaker says that she is of a high quality, and thus is lying, the speaker might be able to verify (or better, falsify) what the speaker says. Thus the hearer has now two strategies: he either checks whether what the speaker said is true or he trust the speaker on her words. In the actual situation where the speaker has a low quality, this means that if the hearer does not check the truth of what the speaker said, he will play a_1, otherwise he will play a_2. Moreover, we will assume that the hearer will punish the speaker in case he finds out that the latter was lying. In that case the hearer gets, let us say, a payoff of $-\epsilon$. This situation might be described by the following kind of utilities for the speaker (where v stands for the action of verifying):

	'I am low' $\rightarrow a_2$ 'I am high' $\rightarrow \neg v + a_1$	'I am low' $\rightarrow a_2$ 'I am high' $\rightarrow v + a_2$
Lying	$1, 0$	$-\epsilon, 1$
Honest	$0, 1$	$0, 1$

We want to know under which circumstances it is still favorable for the speaker to lie. Suppose n is the probability that the hearer will not verify whether the speaker is lying. We want to know what the value of ϵ should be in order for it to be beneficial for the speaker to lie. This is so if the expected utility of lying is higher than the expected utility of being honest, $EU(lying) > EU(honest)$. This is the case when $n > (1-n)\epsilon$. This inequality this gives rise to the function $\epsilon = \frac{n}{(1-n)}$, which can be plotted in the following graph.

Table 1 :

	$n + n\epsilon < \epsilon$ ('I am t_L' is best) iff
$n = 0$	always
$n = 0.25$	$\epsilon > \frac{1}{3}$
$n = 0.5$	$\epsilon > 1$
$n = 0.7$	$\epsilon > \frac{3}{2}$
$n = 0.75$	$\epsilon > 3$
$n = 0.8$	$\epsilon > 4$
$n = 0.9$	$\epsilon > 9$
$n = 1$	impossible

We can certainly assume that the cost of lying in case you are verified is higher than its potential benefit. Thus $\epsilon > 1$. The table above shows that if the chances that the hearer will verify the speaker's message increase, the benefit of lying, $1+\epsilon$, has to increase rapidly in order for it to be expected.

6 Linguistic Complexity

In linguistics it is generally agreed that there are no differences in languages in terms of over-all complexity: all languages are supposed to be equally complex. Indeed, it is very unclear how to compare languages in terms of *theoretical complexity*. Should one compare languages in terms of the number of (syntactic) rules it can be described theoretically? But how then to compare a language with more but less complex rules against another language with less rules which are more complex? Moreover, given that the rules used to describe the language are theory-dependent, which theory should be used to describe the languages? On the other hand, Jacobson (1929) argued that the more diffuse the geographical range of languages, the simpler the system, because of ease of learning and discrimination, or understanding. In dialectology Trudged (2001) argued that languages differ in complexity, both in their phonology, morphology and syntax, and that these differences can be related to characteristics of speech communities. Others (e.g. Thurston, 1992; McWhorter, 2001) have argued that elaborateness and esotericity are more likely to be found in small closed speech communities, where the language has more a symbolic than a communicative function; languages used by larger populations tend to be less complex.

In this paper we will follow Kusters (2003), who studied complexity from an *empirical* point of view. He defines complexity from the point of view of an *outsider*. An outsider, in turn, is defined as a non-native speaker who learns the language at a later stage, who does not have much shared background knowledge with other members of the speech community, and is more interested in clear transmission of information than in expressing personal and group identity and aesthetic feelings.

He finds that languages typically used as 'lingua franca' (such as Arabic, Quechua, Swahili) are more adapted to outsiders, ie. less complex. In communities where a more complex language is spoken, emphasis is laid on native language learning, production and symbolic use.

Let us represent the situation as a game. We assume that the speaker either uses a complex language or a simple one. The hearer is either an insider (an 'Inner') or an outsider (an 'Outer').

Language game :

	Inner	Outer
Complex	$1+\epsilon$,$1+\epsilon$	$0, 0$
Simple	$1, 1$	$1, 1$

If n is the chance that the speaker meets an insider, the expected utility of using a complex language, $EU(C)$, is $n \times (1+\epsilon)$, while the expected utility of using a simple language, $EU(S)$, is 1. It depends on both ϵ and n when it pays off to use a complex language. For instance, if $\epsilon > 1$, it pays of to use a complex language if you think it is at least as likely that the hearer is an insider than that (s)he is an outsider. In general, $EU(C) > EU(S)$ iff $n > \frac{1}{1+\epsilon}$.

This shows that using a complex language only pays off if you have a good chance to meet an insider. In case you want to communicate a lot with outsiders, it is better to use a simple language. It follows that languages that are used a lot as 'lingua franca', such as Arabic, are by this analysis predicted to be simpler (in the above mentioned sense) than those not used as a lingua franca. As such, our analysis of risky speech seems a natural model to account for the data found by Kusters (2003).

The use of language (at least) fulfils two functions. On the one hand, they are used to reliably communicate information. On the other hand, they can be used to express one's own identity, or that of the group to which one belongs. The above considerations suggest the natural conclusion that for languages used as lingua franca, the second function is less important than for more 'local' languages.

7 Conclusion

Sally (2003) discusses how the notion of 'risk' might be important in conversational situations between speakers and hearers. In van Rooij & Sevenster (2006) it is shown how Sally's work can be embedded within Lewisian signaling games, and how some additional ways of speaking can be considered to be risky. In this paper we extended this work again, by making a link between the extend of diversity between speaker and hearer and the risk that is (rationaly) taken by the speaker, and by suggesting the complexity of languages can be explained in terms of risky linguistic behavior as well.

In a sense, these insights were already present in Hume's Treatise:

> Two men who pull the oars of a boat, do it by an agreement or convention, tho' they have never given promises to each other. Nor is the rule concerning the stability of possession the less derived from human conventions, that it arises gradually, and acquires force by a slow progression In like manner are languages establish'd by human conventions without any promise.
>
> Two neighbors may agree to drain a meadow, which they posses in common; because 'tis easy for them to know each others mind, and each may perceive that the immediate consequence of failing in his part is the abandoning of the whole project. But 'tis difficult, and indeed impossible, that a thousand persons shou'd agree in any such action.

ACKNOWLEDGEMENTS

We would like to thank Michael Rovatsos for his encouragement to us to write this paper.

REFERENCES

[1] Fowler, H.W. (1965), *A Dictionary of Modern English Usage*, 2nd edition, revised by E. Gowers, Oxford: Clarendon Press.

[2] Grice, H.P. (1967), 'Logic and conversation', *William James Lectures*, Harvard University, reprinted in *Studies in the Way of Words*, 1989, Harvard University Press, Cambridge, Massachusetts.

[3] Harsanyi, J. C. and R. Selten (1988), *A General Theory of Equilibrium Selection in Games*, MIT Press, Cambridge, Massachusetts.

[4] Hume, D. (1739/2000), *A Treatise of Human Nature*, Oxford University Press.

[5] Jakobson R. (1929). *Remarques sur l'evolution phonologique du russe comparée a celle des autres langues slaves*, Prague.

[6] Kusters (2003), *Linguistic Complexity. The Influence of Social Change on Verbal Inflection*. PhD dissertation, University of Leiden.

[7] McWhorter, J. (2001), 'The world's simplest grammars are creole grammars', *Linguistic Typology*, **5**: 125-166.

[8] Lewis, D. (1969), *Convention: A Philosophical Study*, Harvard University Press, Cambridge, Massachusetts.

[9] Rooij, R. van & M. Sevenster (2006), 'Different faces of risky speech', in; A. Benz, G. Jäger and R. van Rooij (eds.), *Game Theory and Pragmatics*, Palgrave MacMIllan, pp. 155-178.

[10] Rousseau, J.J. (1755), *Discours sur l'origine et les fondement de l'inégalité les hommes*, chez Marc Michel Rey, Amsterdam.

[11] Sally, D. (2003), 'Risky speech: behavioral game theory and pragmatics', *Journal of Pragmatics*, **35**: 1223-1245.

[12] Skyrms, B. (2004), *The Stag Hunt and the Evolution of Social Structure*, Cambridge University Press, Cambridge, Massachusetts.

[13] Thurston, W.R. (1987), *Processes of change in the languages of northwestern New Britain*, Comberra: Pacific Linguistics B-99.

[14] Trudgil, P. (2001), 'Contact and simplification: Historical baggage and directionality in linguistic change', *Linguistic Typology*, **5**: 371-374.

Domain-Based Sense Disambiguation in Multilingual Structured Data

Gábor Bella and **Alessio Zamboni** and **Fausto Giunchiglia**[1]

Abstract. Natural language text is pervasive in structured data sets—relational database tables, spreadsheets, XML documents, RDF graphs, etc.—requiring data processing operations to possess some level of natural language understanding capability. This, in turn, involves dealing with aspects of diversity present in structured data such as multilingualism or the coexistence of data from multiple domains. Word sense disambiguation is an essential component of natural language understanding processes. State-of-the-art WSD techniques, however, were developed to operate on single languages and on corpora that are considerably different from structured data sets, such as articles, newswire, web pages, forum posts, or tweets. In this paper we present a WSD method that is designed for short text typically present in structured data, applicable to multiple languages and domains. Our proof-of-concept implementation reaches an all-words F-score between 60% and 80% on both English and Italian data. We consider these as very promising first results given the known difficulty of WSD and the particularity of the corpora targeted with respect to more conventional text.

1 INTRODUCTION

While current formal or semi-formal data models—spreadsheets, XML trees, RDF graphs, etc.—were designed to ease the processing of data by machines, structured data sets still tend to contain a large amount of informal text expressed in natural language within schema elements, data values, and metadata.

Ever more often, applications need to exploit data sets—link, integrate, query, and search them—facing various aspects of diversity in textual data in the process, e.g., the diversity of the languages used, of terminology, or of the domains covered. Picture multilingual Switzerland where a French application on tourism may need to use travel information available in German as well as geographical open data in English, needing to connect data in multiple languages and from multiple domains. Another example is medical data of patients being exchanged across countries for research purposes, again expressed in different languages and using different national standards.Fig. 1 shows examples of natural language text content extracted from real-world data sets that we will use as running examples:

- open government data in Italian and English from the tourism domain containing points of interest in Trento;
- healthcare data in English containing dosages of drugs;
- university data in English and French on papers published by staff containing abstracts, keywords, titles, etc.

[1] University of Trento, via Sommarive 5, 38123 Trento, Italy. {gabor.bella, alessio.zamboni, fausto.giunchiglia}@unitn.it.

Techniques such as cross-lingual semantic matching [2], semantic search [7], or semantic service integration [13] were designed to tackle diversity in data and therefore invariably have some kind of built-in meaning extraction capabilities. In semantic search, natural language queries should be interpreted and matched to data in a robust way so that search is based on meaning and not on surface forms of words (a tourist's query on *'bars'* should also return establishments indicated as *'winebar'*, cf. fig. 1 a, but preferably no results on the physical unit of pressure). In classification tasks, natural-language labels indicating classes need to be formalised and compared to each other (establishments categorised as *'malga'*, i.e., Alpine huts specific to the region of Trento, should be classified as lodging facilities, cf. fig. 1 a). In service integration, on the schema level, attribute names need to be mapped using schema matching techniques (the English *'address'* mapped to the Italian *'indirizzo'*); while on the data level, heterogeneous terminology used across data sets needs to be mapped to common meanings in order to allow interoperability (*'PhD thesis'* equivalent to *'doctoral thesis'*).

We argue that sense disambiguation that relies on conventional natural language processing methods and toolkits, while still applicable, is suboptimal for dealing with diversity in structured data. First, NLP tools and resources tend to be developed for single languages and the representations they use for word senses do not always allow cross-lingual interoperability. Secondly, the type of text appearing in structured data is considerably different from those targeted by state-of-the-art NLP tools. Most existing efforts on NLP concentrate either on 'conventional' text with full, grammatically correct sentences and standard orthography (e.g., newswire, encyclopedia entries, literature, general web content) or on short and noisy text (e.g., tweets, forum comments).

Compared to these cases, text in structured data tends to be shorter and follows different conventions of orthography and syntax. We believe the best-fitting linguistic category to be that of *block language*, defined in [4] as *'abbreviated structures in restricted communicative contexts, especial use being made of the word or phrase, rather than the clause or sentence.'* In such text *'communicative needs strip language of all but the most information-bearing forms'* [3].

The result is that techniques and resources typically used in NLP, such as machine learning models trained on 'conventional' corpora, can only be applied to structured data with a loss in accuracy. Retraining is not in itself a satisfying answer as *sequence labelling* on words—the usual mode of operation of machine learning tools in NLP—relies on an adequate amount of *co-text*, i.e., preceding and following words, the lack of which in structured data again leads to lower accuracy.

Nome	Categoria	Desc_EN	Indirizzo
AI VICOLI	ristorante	Restaurant and Winebar	Piazza Santa Teresa Verzieri, Trento
ORSO GRIGIO	ristorante	typical restaurant	Via degli Orti, 19, Trento
MALGA CANDRIAI	malga	Alpine hut with typical restaurant	Strada di Candriai, 2, Monte Bondone

(a) Open data from Trentino, Italy, on points of interest.

active_substance	name	note	dosage
apixaban	Eliquis	5 mg of 60 film coated tablets	10 milligrams oral
warfarin	Coumadin	30 tablets 5 mg	7.5 milligrams oral, 7.5 milligrams injected

(b) A data set from a healthcare agency on available drugs and their dosages.

Title	Abstract	Date	Type	Keywords
Concept Search: Semantics-Enabled Information Retrieval	The goal of information retrieval is to map a natural language query…	2010	PhD thesis	semantic search - information retrieval - classification
L'oublie la présence le jeu		2009	article	théâtre - comédie - représentation

(c) A data set of university publications.

Figure 1. Simplified examples extracted from real-world data sets, showing various types of natural language text commonly found in structured data.

This paper provides an approach to word sense disambiguation that is adapted to text in structured data and is based on the following principles.

A language-independent representation of meaning. The disambiguation method is designed to be applicable to multiple languages. The hypothesis underlying this design choice is that meanings of words can efficiently be represented as language-independent concepts. We tackle multilingual diversity at design time through the use of multilingual NLP preprocessors, followed by a language-agnostic WSD method that operates on the concept level and can thus be applied to any language.

Domain-Based WSD suits structured data. The backbone of our method is a domain-based WSD algorithm, for two reasons. First, we observe that contents of structured data tend to be domain-specific. Secondly, we argue that formalising the notion of domain is the first step towards tackling this aspect of diversity in structured data. We tackle the diversity of domains at runtime through automated domain extraction and domain-based WSD.

Weaker reliance on co-text. Due to the shortness of text, the output of machine learning methods trained on long text using sequence labelling—such as state-of-the-art part-of-speech taggers—has to be considered as less reliable and 'taken with a pinch of salt.' For this reason we only very minimally rely on co-text in our WSD approach.

Stronger reliance on structural context. Instead of relying on surrounding words, the context encoded in the surrounding data structures (data set, records, attributes) is exploited for additional clues in order to help disambiguation.

The rest of the paper is organised as follows. Section 2 provides a succinct description of linguistic and structural features of text commonly appearing in structured data. Section 3 develops the general architecture and a theoretical description of the solution, as well as implementation details. Section 4 provides evaluations. Section 5 discusses results and problems yet unsolved. Finally, section 6 presents related work.

2 TEXT IN STRUCTURED DATA

2.1 Linguistic Features

In this section we briefly examine the linguistic characteristics of text in structured data.

Languages. It is not uncommon for data sets to mix languages if they were aggregated from heterogeneous sources or if they were produced in geographical areas or usage domains where multilingualism is common practice. The language may change across records (fig. 1 c) or across attributes (fig. 1 a).

Text length. The typical length of textual attribute values is that of a single phrase with less than 10 tokens (words and punctuation). In attribute names 1–3 tokens are typical.

Orthography. The divergence from standard orthography is considerable:

- capitalisation is used arbitrarily: *ALL CAPITALS* or *no capitals* are frequent, as is *Capitalisation Of Each Word* (all of which can be found in fig. 1 a); capitals are therefore not reliable linguistic indicators and, worse, they can confuse machine learning components trained on conventional text,
- punctuation is often omitted or inconsistently used (e.g., dashes instead of commas are used to separate enumerated items, fig. 1 c),
- abbreviations are frequent, especially in attribute names (fig. 1 a),
- in attribute names dash, underscore, or *CamelCasing* are often used for word separation (figs. 1 a and b);

Parts of speech. Nouns are the most frequent, followed by adjectives, prepositions, verbs, and adverbs. Verbs are rare and are mostly limited to present or past participle form (*'coated'* in fig. 1 b). Consequently, the ability to perform lemmatisation (and more generally, morphological analysis) on verbs is not as crucial as on nouns.

Syntax. In rare cases, attribute values may contain text consisting of full grammatical sentences including noun and verb phrases (such as abstracts in fig. 1 c). Most pieces of text, however, are non-sentential and can better be described through the linguistic notion of *minor sentence*, with the following specific characteristics:

- they consist either of a single noun phrase (*'ristorante'*) or noun phrases connected by conjunctions (*'Restaurant & Winebar'*, *'théâtre, comédie, représentation'*);
- the noun phrase can be simple or contain embedded prepositional and noun phrases (*'5 mg of 60 film coated tablets'*);
- ellipsis and other forms of compression are frequently used ('*30 tablets* [of] *5 mg'*).

Semantics. Attribute names frequently express atomic concepts (*'name'*) but sometimes also complex concepts (*'English description'*), mostly using common nouns and adjectives. Proper nouns denoting named entities rarely also appear in attribute names (*'resident of Italy'*). In attribute values, we distinguish between those that encode sets of concepts (typically *class*, *type*, *category* attributes such as *'Categoria'* in fig. 1 a), those that encode sets of named entities (*'name'* in fig. 1 a and b), and the more complex case of descriptive text that may contain both (*'Abstract'* in fig. 1 c).

2.2 Structural Features

While individual pieces of text tend to be short and thus offer a limited opportunity for context-based analysis of meaning, the data structure itself proves to be a alternative source of contextual information. One may draw an analogy between discourse or pragmatics across sentences in conventional long text and high-level meaning that can be extracted across pieces of text within the data structure.

Our analysis on using data structures as context is intended to be as general as possible, applicable invariably to tabular, tree-based, and graph-based structures. However, it is possible to develop more fine-grained context extraction techniques adapted to specific data structures such as trees or graphs. We leave this problem as future work, referring the reader to [12] that provides a deep analysis of context extraction specifically for data schemas.

2.2.1 Structural Context of Data Values

In this section we examine the structural elements of data sets that may serve as context for texts appearing as attribute values.

Other values of the same attribute. Textual values from other records for the same attribute[2] can be used to derive contextual information:

- the larger lexical context of values considered together may help disambiguation (in fig. 1 c the word *'article'* may in itself be ambiguous but is less so in the context of the preceding attribute value *'PhD thesis'*;
- values of an attribute tend to fall under the same domain (in fig. 1 b the domain that can be associated to the attribute *'note'* is *'medicine'*, which makes the meaning of *'film'* less ambiguous);
- repetitions of words and phrases are very frequent in structured data (the word *'milligrams'* in fig. 1 b), a phenomenon that may introduce severe bias in WSD algorithms if not properly accounted for.

Attribute names. WSD does not need to be applied to all types of text: for example, while names or addresses may need to be disambiguated as named entities, the meanings of words composing them are irrelevant in a lot of use cases. While *named entity recognition* techniques can be used to detect such cases, attribute names such as *'name'* or *'address'* may also be indicative of values holding named entities that do not need WSD.

Other values of the same record. Textual values of other attributes in the same record may provide useful context for disambiguation. In tables b and c of fig. 1 the record-level context provides further domain-specific vocabulary.

2.2.2 Structural Context of Schema Elements

Schema elements are not named nor formalised the same way across data formats: OWL has *properties* and *classes*, XML has *attributes* and *elements*, spreadsheets have *column headers*, etc. Extraction methods of structural context vary depending on the goals and on the type of structure (relation, tree, graph). [12] presents a unified approach to modelling various data formats and to context extraction, while here we only provide a high-level summary.

We consider the context of a schema element to consist of other schema elements and of metadata describing the element.

Metadata, in the form of natural language descriptions of the schema element, are frequent in ontologies (e.g., annotation properties) and in open data (e.g., DCAT metadata).

In order to extract context from other schema elements, it is common practice to consider those elements that are directly or transitively related to it, possibly within a given distance.

Trees. In a tree-shaped data structure (a classification, an XML or JSON file) the parent-child relation is used to extract the context of a given node. The context is selected from the set of ancestor and descendant nodes, including the root.

Tables. Tabular data structures can be considered as shallow trees, with the root being the name of the table or relation. The context of an attribute (of a column header) consists of the parent, i.e., the root.

Graphs. In graph-shaped ontological schemas (such as OWL ontologies) relations are named and are often qualified with metaproperties (reflexivity, symmetry, transitivity). The degree of freedom to select the relations that are included in the context of a node is thus much higher than in the previous cases. In this paper we do not consider this use case and direct the reader to [12] for more details.

3 WSD ON STRUCTURED DATA

3.1 General Architecture

Based on our analysis in the previous section, and focusing in particular on the use cases evoked in section 1 (semantic search, classification, query answering, data integration), we identify the following types of meaning extraction tasks relevant for structured data (not pretending to be exhaustive):

concept extraction: this semantic-level operation is commonly solved as the NLP task of *word sense disambiguation*;

[2] The term *record* may unintentionally imply a relational data structure. In case of ontological or object-oriented resources the term *instance* might be more appopriate. We do not intend to restrict the scope of our work to a specific type of data model so we use the term *record* in the most general sense possible.

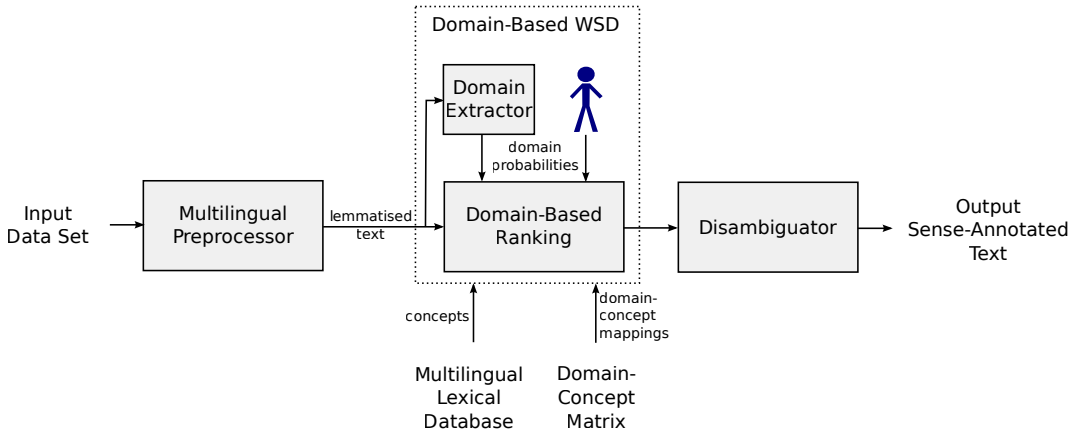


Figure 2. WSD architecture for text in structured data.

named entity extraction: this semantic-level operation is commonly solved as a *named entity recognition and disambiguation* problem;

domain extraction: this pragmatic-level operation is commonly solved as a *document classification* problem and, as we will show, can serve as input for the WSD task.

In this paper we do not consider the extraction of named entities; rather, we set out to provide a WSD method that is able to extract concepts from text appearing in structured data. The high-level architecture of our method is shown in fig. 2:

1. a *Multilingual Preprocessor* provides lemmas and parts of speech for texts in multiple languages extracted from the input data set;
2. possible meanings for lemmas are retrieved from a *Multilingual Lexical Database*;
3. domains relevant to the processed text are estimated by providing *domain probabilities* for lemmatised texts, either by a human user or by a *Domain Extractor* algorithm using a *Domain–Concept Matrix*;
4. based on the domains estimated the *Domain-Based Ranking* of concepts provides preliminary scores to meanings of polysemous words;
5. based on rankings and on hints computed during preprocessing, a *Disambiguator* component produces a final ranking where the top-ranked concepts are the disambiguated ones.

3.2 The Multilingual Lexical Database

WSD annotates words by labels representing formally defined meanings that are usually taken from some form of knowledge resource. In particular, we call *lexical-semantic concepts* the basic meanings defined by two well-known types of linguistically-oriented knowledge resources: *wordnets* and *term bases*. In the case of wordnets the lexical-semantic concept corresponds to the *synset* (i.e., set of synonyms, see [14]) while in term bases it is the *terminological entry*.

The main difference between wordnets and term bases is that the former are single-language multi-domain resources while the latter are (usually) designed to be multi-language and single-domain. Thus, a wordnet maps a lexical entry to one or more possible language-dependent lexical-semantic concepts. These concepts tend to be characteristic of various domains that, however, are usually not explicitly indicated by the wordnet. A term base, on the other hand, maps terms in multiple languages to a single language-independent terminological meaning from a specific domain.

For our purposes of WSD we use a hybrid knowledge resource that we call a *multilingual lexical database* (MLDB). It is defined as a multi-language multi-domain resource where by *multi-language* we understand that it maps lexical entries from multiple languages to language-independent lexical-semantic concepts, and by *multi-domain* we understand that lexical-semantic concepts may belong to different domains. However, we do not require the domains of lexical-semantic concepts to be explicitly indicated within the MLDB.

Let C be an ontology of language-independent lexical-semantic concepts (in short: concepts) c_i. Let l be a lexical entry defined as $l = (\mathfrak{l}, L)$ where $\mathfrak{l}$ is a lemma (word in dictionary form) and L is the language of the lemma. Then the MLDB is defined as

$$\text{MLDB} = \left\{ (l, \{c_i^l\}) \right\}$$

where c_i^l are the language-independent meanings of the lexical entry l.

Existing MLDBs include *EuroWordNet* [17], *MultiWordNet* [15], the *Universal Knowledge Core* (UKC) [5] implemented at the University of Trento and, more recently, *BabelNet* [6]. We used the UKC for our research, also integrating in it some content from MultiWordNet.

In reality, wordnets and MLDBs are more complex than what our definitions above may suggest—in particular, they also encode relations among concepts—but these aspects are not relevant for our paper.

3.3 Multilingual Preprocessing

The MLDB serves the purpose of providing meanings associated to lemmas in multiple languages. Consequently, the text to be sense-disambiguated first needs to be lemmatised. We achieve this using multilingual NLP pipelines specially optimised for the parsing of short text. The languages currently supported are English, Italian, Spanish, and Mongolian. Pipelines consist of the following components, in this order:

Language detector. This component allows the correct language-specific pipeline to be called without an explicit indication of language by the user, which is practical for data sets that contain attribute names or values in multiple languages.

Pipeline selector. Using a heuristic based on text size, one of two pipelines is instantiated:

- a conventional NLP pipeline for longer texts composed of full sentences (we do not discuss this pipeline in the paper), such as abstracts in fig. 1 c;
- a pipeline optimised for short text.

Tokeniser. Tokenisation is optimised to the characteristics of short text as described in section 2.1. We currently use regular expressions but the training of a learning-based tokeniser on short text is also a possibility.

Part-of-speech tagger. As conventional learning-based POS taggers (such as OpenNLP) are suboptimal on short text, we use their output cautiously. First, we distinguish between closed-class and open-class words (nouns, verbs, adjectives, adverbs). For the latter, any further detail provided by the POS tagger is used merely as a hint. Based on prior frequencies of parts of speech we observed in structured data, in cases of polysemous open-class words we use a heuristic scoring system to favour certain parts of speech: nouns $>$ adjectives $>$ verbs. For example, the noun meanings of the word *'search'* in fig. 1 c will be preferred over its verb meanings. Scores are currently hard-coded but in the future we are planning to use syntactic analysis to improve guesses. These scores are used in the final disambiguation phase in combination with domain-based ranking of meanings.

Lemmatiser. Lemmatisation retrieves for every word form all possible lemmas (e.g., *'tablet'* for *'tablets'*. As due to text shortness parts of speech cannot be guessed with a high enough certainty at this point, no POS-based filtering of lemmas is applied.

Multiword detector. Dictionary-based multiword detection is applied to find lemmas composed of multiple words.

Due to the non-conventional syntax of text in structured data (cf. section 2.1), we currently do not apply syntactic parsing. We however plan to research the syntactic properties of such text as future work in order further to improve disambiguation accuracy.

3.4 Domain-Based WSD

The adoption of a domain-based approach as the backbone of our WSD method is motivated by the observation that the contents of structured data sets, and even more of individual attributes within data sets, tend to belong to specific domains. This can be considered as an adaptation of the *one-domain-per-discourse hypothesis* that claims that *'multiple uses of a word in a coherent portion of text tend to share the same domain'* [9, p. 28].

3.4.1 A Formal Notion of Domain

As a simple definition, we represent a *domain label* d_j as a concept taken from the MLDB (such as *'travel'*, *'medicine'*, *'sport'*, *'education'*). We suppose that the set $D = \{d_j\}$ of domains is closed and is relatively small (no more than a couple hundreds), although these are more practical than theoretical requirements.

In conformance to real-world resources, we defined MLDBs not to possess an explicit notion of domain. We therefore provide explicit domain information at this point as an extension to the MLDB. Inspired by [11] and [10] we add domain information to a concept c_i through a mapping to a domain label:

$$m_j = \Big(d_j, \bigcup_{i=1}^{|C|} \{(c_i, w_{ij})\}\Big)$$

meaning that for each concept c_i we provide a weight w_{ij} linking that concept to the domain d_j. The weight w_{ij} is a rational number between 0 and 1. For example, the concept of *'film* [as a movie]*'* will be mapped to the domain label *'media'* with a strong weight while the concept of *'film* [as coating]*'* will be mapped to it with a much lower weight.

The *domain* j is then formally defined as the domain label d_j together with the union of its mappings: $(d_j, \cup_{j=1}^{|D|}\{m_j\})$.

3.4.2 The Domain–Concept Matrix

All mapping of concepts to domains are collected in a resource called the *domain–concept matrix*, $\mathbf{W}$, defined as

$$\mathbf{W} = (w_{ij}) \in \mathbb{Q}^{|C|\times|D|}$$

where $\mathbf{W}$ has as many rows as there are concepts and as many columns as there are domains.

Note that by mapping domain labels to language-independent concepts we obtain a resource $\mathbf{W}$ that can be used across languages.

We are aware of three existing resources mapping meanings from lexical databases to domains:

- *WordNet Domains* (WND) by Magnini et al. [11];
- *Extended WordNet Domains* (XWND) by González-Agirre et al. [10];
- *WordNet Topics* (WNT) included in Princeton WordNet itself starting from version 3.0.

All three use Princeton WordNet as lexical database, thus they can be considered as monolingual English-only resources. WND is an earlier work defining about 170 domains and using binary weights (0 or 1) to model English synsets either belonging or not belonging to domains. XWND was developed as an improved and extended version of WND: it maps *all* concepts to *all* domains using rational numbers for weights, each concept having a positive nonzero weight with respect to each domain. Finally, WNT defines about 440 topics but only annotates a subset of its synsets by topic.

For our work we reused XWND because of its full coverage of WordNet synsets and because of its use of weighted mappings between domains and meanings, lending itself better to statistical methods. For our purposes we modified the XWND resource as follows: first, we converted mappings so that they map domains to language-independent concepts of the UKC instead of English synsets. This way the resource became reusable across languages. Secondly, we made sure that weights of concepts always add up to 1 for any given domain, so that we can consider the set of mapping weights for each domain as a distribution of conditional probabilities $\mathrm{P}(c_i|d_j)$: given a (meaningful) word in a text that we know belongs to domain d_j, $\mathrm{P}(c_i|d_j)$ is the probability of c_i being its meaning. This interpretation allows us to formalise disambiguation using basic probability theory.

3.4.3 Domain-Based Ranking

The domain-based meaning ranking algorithm takes the following inputs:

- the input text as a series of lemmatised tokens;
- the MLDB providing possible concepts for each lemma;
- $\mathbf{W}$ providing domain–concept mappings of the form $\mathrm{P}(c_i|d_j)$;
- a *domain vector* $\bar{d}_t$ that for each domain d_j provides the probability $\mathrm{P}(d_j|t)$ of the input text t belonging to that domain.

Note that having $\bar{d}_t$ as input supposes that the algorithm has prior knowledge of the input domains. This makes sense as data sets are often categorised into domains, e.g., in CKAN open data catalogues, or else they can easily be categorised by the user. Still, in case such information is not available we provide in the next section an automated domain extraction method that computes $\bar{d}_t$ from t.

Based on these inputs, domain-based ranking is provided by the simple formula below that outputs a probability for each concept of each lemma given the input text:

$$\mathrm{P}(c_i^l|t) = \sum_{j=1}^{|D|} \mathrm{P}(c_i^l|d_j)\mathrm{P}(d_j|t)$$

where l is the lemma to be disambiguated and c_i^l are the possible concepts of the lemma provided by the MLDB.

Note that the disambiguation method is context-independent in the sense that it does not take surrounding words into account. This is a deliberate feature that allows the method to work on very short text without affecting performance. Contextual information is present in an implicit manner in the input domain vector $\bar{d}_t$, i.e., in the values $\mathrm{P}(d_j|t)$ that characterise the text as a whole. Note that this supposes that a whole piece of text can entirely be characterised by a single domain vector, in other words, that domains do not change along the text. This is a reasonable hypothesis for short text typically present in structured data.

3.4.4 Domain Extraction

An input domain vector $\bar{d}_t$, providing text-specific domain probabilities to the ranking algorithm, can be obtained in several ways:

1. as a hardcoded default distribution reflecting a prior likelihood of domains to be encountered in data (e.g., an open government data portal is more likely to contain data about tourism or finance than about astrology);
2. as user input, provided either by the data owner (frequent on open data portals) or by the data scientist supervising the meaning extraction task;
3. using an automated domain extraction method.

In this section we provide an algorithm for the third option. Domain extraction can be seen as a document classification problem for which a large number of solutions exist, e.g., supervised learning-based classifiers. Our method is unsupervised and relies only on the same two resources: MLDB and $\mathbf{W}$.

The inputs of the domain extractor are:

- a set of input texts, each as a series of lemmatised tokens;
- the MLDB providing possible concepts for each lemma;
- conditional probabilities $\mathrm{P}(d_j|c_i)$ providing for a concept c_i the probability of it belonging to domain d_j.

Its output is the domain vector $\bar{d}_t$ that represents the probability of each domain being characteristic of text t.

Note that, because single pieces of text are usually too short to provide meaningful domain estimates, the domain extractor is able to take several pieces of text as input simultaneously. In particular, in an analogous manner to the *one-domain-per-discourse* heuristic, we adopt a *one-domain-vector-per-attribute* hypothesis that a single domain vector can be computed over all values of a given structured data attribute combined together. Thus in the following t represents the concatenation of pieces of short text that we suppose to belong to the same domain.

We define the *domain vector of a concept* c as

$$\bar{d}_c = (\mathrm{P}(d_1|c), \ldots, \mathrm{P}(d_j|c), \ldots, \mathrm{P}(d_{|D|}|c)).$$

Intuitively, $\mathrm{P}(d_j|c)$ is the probability of a domain d_j being representative of a concept c.

The domain extraction algorithm estimates the domain vector $\bar{d}_t$ of the input as the centroid of all domain vectors of all possible concepts of all lemmas in text t:

$$\bar{d}_t = \mathrm{centroid}_{\forall c_i^l} \mathrm{P}(d_j|c_i^l).$$

In section 2.2 we observed a problem specific to structured data: repeating words and phrases are very frequent across attribute values. This is a problem as repetitions introduce a significant bias into the computation of centroids and, in general, into any context-based meaning extraction method. Simply removing repetitions, however, would have the adverse effect of giving rare outliers the same importance as to frequently appearing words. As a compromise, we apply a logarithmic function to the number of repetitions, smoothing differences in frequencies all the while favouring frequent words over rare ones.

We still need to show how to obtain $\mathrm{P}(d_j|c_i)$, that is, the domain vector of concept c_i. It is computed from $\mathbf{W}$ using Bayes' theorem:

$$\mathrm{P}(d_j|c_i) = \mathrm{P}(c_i|d_j)\frac{\mathrm{P}(d_j)}{\mathrm{P}(c_i)}.$$

In turn, we need to provide $\mathrm{P}(d_j)$ and $\mathrm{P}(c_i)$.

The former are considered to be *prior domain probabilities* that are initialisation parameters of our WSD method. The user is expected to initialise each $\mathrm{P}(d_j)$ according to their best judgment of domains to appear in input data sets. In the absence of user input, prior domain probabilities can be set to default values, at the worst case as a uniform distribution.

The latter are again computed from $\mathbf{W}$ simply as

$$\mathrm{P}(c_i) = \sum_{j=1}^{|D|} \mathrm{P}(c_i|d_j)\mathrm{P}(d_j).$$

3.5 Disambiguation

Disambiguation is represented as a separate component from domain-based ranking in order to allow a fusion of WSD methods to be applied. In its current version our disambiguator computes final rankings based on two inputs: the output of domain-based ranking and the output of the POS tagger from multilingual preprocessing. Scores output by the former are modified according to the hints provided by the POS tagger, combining the output of an OpenNLP tagger with prior frequencies of POS tags observed in structured data (nouns and adjectives being much more frequent than verbs and adverbs).

Name	Lang.	Content Type	Nb. texts	Nb. tokens
Prodotti Tradizionali Trentini	IT	short and long text from attribute values	108	4,952
Esercizi Alberghieri	IT	short text from attribute values	15	81
Esercizi Alberghieri	IT	attribute names	30	59
Esercizi Alberghieri	IT	short text from metadata values	30	146
Esercizi Alberghieri	EN	short text from metadata values	30	126
Public Infrastructures	EN	attribute names	94	24
Total			**307**	**5,388**

Figure 3. Evaluation data sets taken from open data in Trentino, Italy, and the UK.

4 EVALUATION

4.1 Evaluation Corpora and Method

Our evaluation data sets are open government data from Trentino, Italy[3] and from the UK[4] (fig. 3). The languages tested were English and Italian while the domains covered were food, tourism, and government. Four types of text were analysed:

- long, conventional text as attribute values from the food domain in Italian (*Prodotti Tradizionali*);
- short text as attribute values in Italian from the food domain (*Prodotti Tradizionali*) and from the tourism domain (*Esercizi Alberghieri*);
- attribute names in Italian (*Esercizi Alberghieri*) and in English from the construction domain (*Public Infrastructures*);
- metadata values in Italian from the tourism domain (*Esercizi Alberghieri*) and in English from the construction domain (*Public Infrastructures*).

The corpora were hand-annotated on all words with concepts from the UKC (the multilingual database we used for our evaluations). The English and Italian contents of the UKC were imported from Princeton WordNet 2.1 (110K synsets) and the Italian MultiWordNet (34K synsets), respectively.

Because results were biased by an occasionally incorrect tokenisation, we discarded such tokens from the computation of results. This way we could evaluate the WSD method independently of the rest of the NLP pipeline.

The input domain vectors (i.e., the domains relevant to the data sets) were provided by automated domain extraction, therefore results reflect the performance of the domain extractor and of the disambiguator together, in other words, of the 'fully automated' mode of disambiguation without user intervention.

Precision and recall were computed using multi-class evaluation, each concept considered as a class in itself. In order to combine scores of all classes we used both micro- and macro-averaging:

$$P_\mu = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=0}^{|C|} TP_i + FP_i}; R_\mu = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=0}^{|C|} TP_i + FN_i}$$

[3] `http://dati.trentino.it`
[4] `http://data.gov.uk`

$$P_M = \frac{1}{|C|}\sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FP_i}; R_M = \frac{1}{|C|}\sum_{i=1}^{|C|} \frac{TP_i}{TP_i + FN_i}$$

Macro-averaging is a simple arithmetic mean over each class, ignoring repeating words (if the same meaning is erroneously disambiguated a hundred times it still counts as one mistake). Micro-averaging, instead, is computed by occurrence and is thus heavily biased by repetitions in the data. Finally, we combined precision and recall for each type of averaging to obtain the F-scores that are shown in the results below.

4.2 Evaluation Results

Our results are shown in fig. 4. There are three pairs of bars shown for each data set, each pair corresponding to the micro- and macro-averaged F-scores:

DB (Domain-Based): our method described above;
Freq (Frequency-Based): as it is common for evaluations of WSD methods, we provide as comparison a baseline frequency-based disambiguator that is based on prior meaning frequencies or ranks, always selecting the most frequent meaning independently of the surrounding text;
KB (Knowledge-Based): still as comparison, a classic knowledge-based WSD method that is designed for longer pieces of text. Using the IS-A hierarchy of concepts in the MLDB, it computes LCA (least common ancestor) distances between the meanings of the word being disambiguated and the meanings of surrounding contextual words. The hypothesis behind this method is that shorter LCA distances correspond to 'closer' and thus more probable meanings.

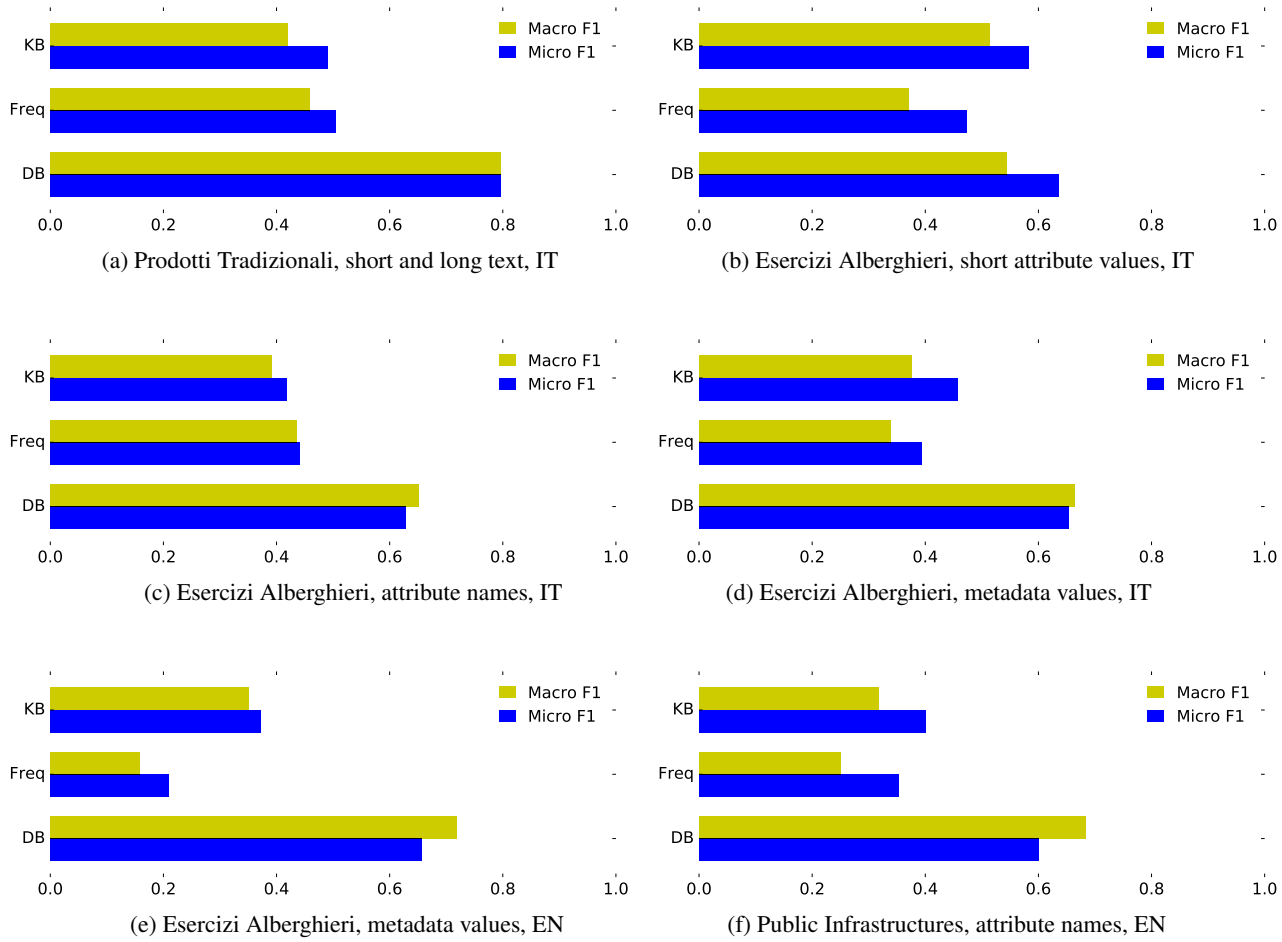
Note that the frequency-based baseline method is context-independent and language-specific, while the knowledge- and the domain-based method share the property of being language-independent as they both operate on concepts. They both rely on context, although in significantly different ways: the knowledge-based disambiguator uses surrounding words (co-text) while the domain-based one uses structural context solely for the purpose of domain extraction.

The following observations can be made about the results:

- for all data sets, the scores obtained by our method are superior to the knowledge-based method and the baseline, and except for one data set (b), the difference is consistently higher than 20%;
- there is no significative difference in performance between Italian (a–d) and English (e–f), although results are not directly comparable as the data sets evaluated are different;
- results are the best (80%) on the data set containing large quantities of long text (a): this is not surprising as this data set contains vocabulary that very clearly belongs to the food domain; moreover, larger quantities of text allow the domain extractor to be more precise;
- on short text the micro-F scores are in the 60–65% range while the macro-F scores are more spread out in the 55–72% range.

5 DISCUSSION

Our results seem to confirm that our approach—based on the principles of domain-based operation, a language-independent WSD algorithm, and a structural delineation of context—can suitably support natural language understanding tasks over structured data. The F-scores obtained usually fall in the 60%–70% range (with a lower



(a) Prodotti Tradizionali, short and long text, IT

(b) Esercizi Alberghieri, short attribute values, IT

(c) Esercizi Alberghieri, attribute names, IT

(d) Esercizi Alberghieri, metadata values, IT

(e) Esercizi Alberghieri, metadata values, EN

(f) Public Infrastructures, attribute names, EN

Figure 4. Evaluation results. KB: knowledge-based disambiguation included for comparison; Freq: frequency-based baseline disambiguation included for comparison; DB: domain-based disambiguation.

outlier of 55% and a higher outlier of 80%) which makes our current WSD implementation usable in real-world applications, including those requiring multilingual and cross-lingual support.

To put our results into perspective, *inter-tagger agreement* and baseline methods provide upper and lower bounds on the scores that can reasonably be targeted by WSD systems. Inter-tagger agreement on all-words tasks (where all meaningful words are sense-annotated) using fine-grained senses (such as those provided by WordNet) was reported in [1, section 1.6] to vary between 70% and 90%. On the other end of the spectrum, a baseline method that always chooses the statistically most frequent meaning reportedly ([1, section 1.6]) reaches 57% on average for the English all-words task on long text. Interestingly, our evaluation of frequency-based baseline disambiguation on our own corpora performed considerably worse, in the 20–40% range. We attribute this somewhat surprising outcome to the domain-specificity of our corpora where the distribution of meanings may be radically different from the corpora used to compute frequency data. The conclusion we draw from this comparison is that the classic baseline method tends to provide scores on structured data that are below the barrier of usability, providing a further argument for more sophisticated WSD methods. Let us also note that, in practice, meaning frequencies are not easily available for languages other than English.

Despite the promising results, we still consider our method to be essentially an early-stage proof of concept. From a research perspective, several of our hypotheses need further verification, as are some rudimentary techniques in need of a more solid theoretical backing.

In particular, a line of research we wish to pursue is a statistically backed-up linguistic analysis of the lexical categories and of the syntax used in block language typically present in structured data. We wish to investigate to what extent this language (or these languages) can be systematically characterised, and to what extent such characterisations may be exploited for WSD, e.g., through statistical parsing. Such a work would be the continuation of successful prior research on *descriptive phrases*, the language of classification labels that has already been described in [7]. We consider descriptive phrases as a special case of block language, and thus a subset of the language we are interested in interpreting.

Another hypothesis that we have not thoroughly investigated so far is the cross-lingual applicability of domain information. The domain resource we exploited—XWND [10], itself derived from WND [11]—was developed using semi-automated knowledge-based techniques on top of Princeton WordNet. It is therefore necessarily biased towards the English language to some extent. While our successful application of it to WSD on Italian does provide a certain evidence towards cross-lingual usability from a practical perspec-

tive, it is hard to draw theoretical conclusions from such high-level quantitative comparisons. For instance, a lot depends on the degree of polysemy present in the lexical databases of the languages being compared: a lower number of polysemous words makes the disambiguation task easier. Specifically in the case of English and Italian wordnets the degree of polysemy turns out to be almost identical,[5] so we do not consider our results to be heavily biased by the underlying lexical databases. Still, the linguistic specificity of domain resources remains a question to be investigated, in our view much related to the more general problem of transferring knowledge resources across languages and cultures.

6 RELATED WORK

Word sense disambiguation is a mature research area with a wide range of solutions proposed [1]. While the problem in its generality is considered AI-hard, its actual difficulty is largely dependent on the targeted coverage (lexical-sample vs all-words), granularity of meaning distinctions (homonymy vs polysemy), corpora, etc.

Most research efforts and evaluations, including those reported in [1], were performed on conventional long text. Statistics derived from those results cannot directly be compared to ours, obtained on structured data. Unfortunately, there is very little published research on sense-disambiguation of structured data or block language, making the comparison of our results difficult. Works we are aware of are only concerned with the disambiguation of data schemas, most frequently for the purpose of ontology matching. [8], for example, analyses labels of tree-shaped classifications and proposes a structure-based disambiguation technique taking ancestor and descendant nodes as context. [16] is a survey on similar techniques. In our view, however, ontology matching is not among the tasks that greatly benefit from WSD, as the goal of ontology matching is to find correct matches between ontology elements, regardless of whether their textual contents are correctly disambiguated or not. For this reason, techniques proposed specifically for ontology matching tasks tend not to generalise well to other use cases.

The article [12] provides a detailed analysis on context extraction and disambiguation from data schemas. From our perspective, its main contribution is the generic and adaptable process by which context can be extracted from diverse schema types and depending on the underlying use case. Its difference with respect to our work is that it is aimed at English only, it does not address the disambiguation of data values, and it uses different WSD techniques.

Previous results on domain-driven WSD heavily inspired our work. We adapted some techniques put forth in works by Magnini et al., such as [11], and we reused resources provided by González-Agirre et al. [10]. While these works were developed for the processing of conventional text in English, we wished to show that they could successfully be adapted to structured data and to text in multiple languages.

ACKNOWLEDGMENTS

We acknowledge the *ESSENCE*, *Smart Society*, *Open Data Trentino*, *Healthcare Data Safe Havens*, and *Digital University* projects for having provided us with resources and data. We also acknowledge the authors of the *Extended WordNet Domains* project for having published their resource free of use.

[5] We evaluate the *degree of polysemy* by the proportion of monosemous words, the average number of synsets per word, and the average number of words per synset.

REFERENCES

[1] Eneko Agirre and Philip Edmonds. *Word Sense Disambiguation: Algorithms and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2007.

[2] Gábor Bella, Fausto Giunchiglia, Ahmed Ghassan Tawfik AbuRa'ed, and Fiona McNeill. A Multilingual Ontology Matcher. In *Proceedings of OM-2015 located at ISWC 2015, CEUR-WS vol. 1545*.

[3] D. Biber, S. Johansson, Geoffrey Leech, S. Conrad, and E. Finegan. *Longman Grammar of Spoken and Written English.* Longman, 1999.

[4] D. Crystal. *Dictionary of Linguistics and Phonetics*. The Language Library. Wiley, 2011.

[5] Fausto Giunchiglia et al. Faceted Lightweight Ontologies. In *Conceptual Modeling: Foundations and Applications*, volume 5600. Springer Berlin Heidelberg, 2009.

[6] Maud Ehrmann et al. Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*

[7] Fausto Giunchiglia, Uladzimir Kharkevich, and Ilya Zaihrayeu. Concept search. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC 2009 Heraklion, pages 429–444, Berlin, Heidelberg, 2009. Springer-Verlag.

[8] Fausto Giunchiglia, Mikalai Yatskevich, and Pavel Shvaiko. Semantic Matching: Algorithms and Implementation. *J. Data Semantics*, 9:1–38, 2007.

[9] Alfio Gliozzo and Carlo Strapparava. *Semantic Domains in Computational Linguistics*. Springer-Verlag Berlin Heidelberg, 2009.

[10] Aitor González-Agirre, German Rigau, and Mauro Castillo. *A Graph-Based Method to Improve WordNet Domains*, pages 17–28. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[11] Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo. Using Domain Information for Word Sense Disambiguation. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, SENSEVAL '01, pages 111–114, Stroudsburg, PA, USA, 2001. Association for Computational Linguistics.

[12] Federica Mandreoli and Riccardo Martoglia. Knowledge-based sense disambiguation (almost) for all structures. *Inf. Syst.*, 36(2):406–430, April 2011.

[13] Fiona McNeill, Paolo Besana, Juan Pane, and Fausto Giunchiglia. *Service Integration through Structure-Preserving Semantic Matching*, pages 64–82. IGI Global, 2010.

[14] George A. Miller. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41, November 1995.

[15] Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 21–25, 2002.

[16] Joe Tekli. An overview on xml semantic disambiguation from unstructured text to semi-structured data: Background, applications, and ongoing challenges. *IEEE Trans. on Knowl. and Data Eng.*, 28(6):1383–1407, June 2016.

[17] Piek Vossen, editor. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA, 1998.