Making Data Understand People

Lernout Stephen and **Devos Geert** and **Platteau Frank**[1]

Abstract Title Rebooting Natural Language Processing (NLP).
A Non-Biological AI approach towards Natural Language Understanding

Abstract. The pain Miia is addressing in this paper is that the older generation tools like Natural Language Processing, statistical keyword search and fuzzy logic do not deliver in terms of real text understanding. Their vendors struggle in delivering accurate quality and this results in ill-functioning applications. The newer generation methodologies like Deep Learning and Cognitive Computing are breaking barriers in the (Big Data) fields of Internet of Things, Robotics and Image/Video Recognition but cannot be successfully deployed for text without huge amounts of training and sample data. In the short term, we believe non-biological Artificial Intelligence will produce the best results for text understanding. We applied advanced Linguistic and Semantic Technologies combined with ConceptNet modeling and Machine Learning to cater deep intelligent and cross-language quality to several industries.

METHODOLOGY

Our companies' products are based on proprietary semantic technology - a set of techniques to find meaning in unstructured text, and to use meaning to find other texts. Finding meaning is done by analyzing text linguistically, mapping words and expressions onto a ConceptNet and using powerful semantic pattern matching to combine these concepts into meaningful entities. The basic semantic analysis and matching engine is language and domain independent. So this means that whenever we want the engine to handle a new domain, a new ConceptNet needs to be built. This is obviously a notoriously expensive procedure when this is done manually. The basic approach is to use open-source & domain-specific ontologies and taxonomies. When we add a new language to the domain-dependent application, a dictionary has to be created mapping lexical expressions to the concepts. The ConceptNet then functions as an interlingua, and matching between documents written in different languages becomes possible. Obviously, creating these lexicons manually is also very expensive. The crux in our approach is to find ways to speed up building ConceptNets and mapping lexicons. In order to do this we are developing Machine Learning capabilities to help us create these ConceptNets on the fly , and use machine translation tools to generate lexicons automatically. The following machine learning techniques are being applied.

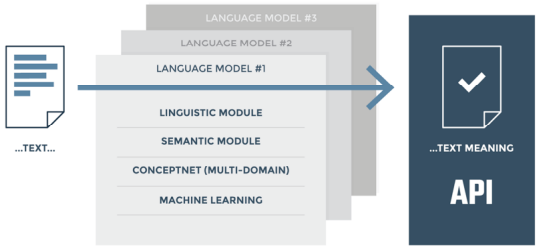
MACHINE LEARNING

- Automatic document classification: when analyzing large disparate sets of documents (e.g. in the legal or administrative domains) it is very important to know to which domain the document belongs.
- Automatic vocabulary extraction of documents in the same domain, with assessment of importance and relevance of the extracted terms, compared to terms from other domains
- Discovery of distance and similarity between concepts. For instance in HR, we can automatically generate similarities between professions, based on competences they do or do not share.
- Parsing of Wikipedia or dictionary articles to automatically extract ontological information ("Gas tungsten arc welding is an arc welding process")
- Online tools that systematically scan the internet looking for new terminology and concepts in a self-learning mode. If a page belongs to a known domain, the terminology on that page is likely to belong to that domain as well.

KEY ADVANTAGES

- language agnostic
- domain independent
- cross-language performance
- self learning

ENGINE VISUAL



WHAT IT REALLY TAKES ...

WHAT IT TAKES FOR DEEP SEMANTIC (LANGUAGE) UNDERSTANDING

Species	ops (Operations per second)	Pattern Recognition	Lexicon	Deep Semantic Language Understanding (DSTU)	Degree of Selfconsciousness
Adult Human	10 ttp 27	superior	unlimited	yes	full
Human Child under 18 months	10 ttp 25	superior	good	no	emerging
Adult Chimp	10 ttp 24	superior	good	no	emerging
Giant bio-mimicking or statistical computers	10 ttp 15-18	superior	unlimited	emerging... at best	no
Dog	10 ttp 20	superior	good	no	zombie
MIIA DSTU on PC server	10 ttp 6	good	unlimited	Yes (rules & concepts preprogrammed)	no

Source: Published literature by Sir Roger Penrose & Stuart Hameroff

[1] Miia Inc. – Text Understanding – San Jose, USA