A Universal Language Engine for Machine Reading

Giuseppe Attardi and Maria Simi

joint work with Niladri Chatterjee, Stefano Dei Rossi, Zahurul Islam, Haoyuan Li
Dipartimento di Informatica
Università di Pisa

Context. Traditionally, research in Natural Language Processing has divided the activity of language analysis into a series of simpler tasks, in order to tackle such a complex problem by splitting it into more affordable separate sub-tasks. In particular, tasks such as Part of Speech Tagging or Chunking purportedly can be carried out in isolation using only shallow linguistic information, for instance superficial aspects related to the morphological structure of words (prefix, suffix, capitalization, etc.). Processing can be done with Markov models, assuming a behavior that only depends on the current state and is independent from the past and the future. Exploiting techniques of Machine Learning, tools have been developed that perform such tasks with good accuracy (close to 98%) and high efficiency.

Sentence analysis requires performing a sequence of tasks, using linguistic pipelines consisting of tools such as: Sentence Splitter, Tokenizer, Lemmatizer, POS tagger, Chunker, Parser. Additional tools for semantic analysis include Named Entity Recognizer, Super Sense Tagger, Semantic Role Labeling, Coreference Resolution [20] and Machine Translation [13].

It is questionable though whether the human mind performs linguistic analysis in separate stages: in fact in many situations it would appear quite beneficial to exploit information produced by one component while performing another task.

It has been recently shown [9] that a dependency parser can subsume other tasks such as chunking with no loss of efficiency, by exploiting new algorithms and deterministic parsing techniques [1, 3, 4, 5, 6]. Semantic Role labeling can be done while parsing with minor extensions to the parser algorithm. We also showed that a Named Entity Recognizer can avoid the use of POS tags [7], extracting directly simple features from words.

Hence it possible to reconsider the architecture of Natural Language Processing systems, reducing the number of pipeline stages, as some recent research on multitask systems [18] suggests.

Improving the effectiveness of text analytics technologies may have significant impact in the way information is processed to extract knowledge, in the way computers can support higher level of interaction with humans and in the way they can assist in performing complex tasks.

Research Goals. Machine Reading [19] involves the ability of capturing knowledge from naturally occurring texts and transforming it into suitable representations for use by reasoning systems, analysis systems or any other processing tools.

Our main research goal is to design and develop the architecture for a *Universal Integrated Natural Language Analytics Engine for Machine Reading*, i.e. capable of performing complex semantic analysis of texts, exploiting multiple layers of features extracted simultaneously from the input documents. Such engine can be trained jointly on all tasks and produce a *common model* consisting of shared weights that allows obtaining a *reading of a document* from multiple perspectives. In particular the engine could be trained to perform parsing, semantic role labeling, information extraction and syntax-based translation.

Linguistic knowledge will be incorporated into special document indexes that will enable semantic search. The *semantic index* will represent linguistic information in a redundant way, spread across each occurrence of a word [2], in order to be quickly accessible where needed for search, while exploiting compression techniques for space saving. Such redundancy is the opposite of database normalization and is akin to *denormalization*.

The language model and semantic index will represent billions of linguistic knowledge items stored in a form resembling an *artificial brain* with a huge number of connections. The language engine will exploit this structure and provide effective linguistic abilities.

Research Challenges. A relevant aspect of the integrated architecture will be the ability of creating *a single model encompassing all knowledge about a language*, instead of relying on half a dozen models, one for each task.

The unified model instead will be learned automatically by absorbing large quantities of text and it will be enriched and grow continuously by the addition of new textual documents, obtained from the many available sources both public and private (e.g. mail).

Suitable machine learning techniques will be required, for instance *Deep Learning* techniques for building the engine models, extending those used in the DeSR parser [6].

A second essential aspect of a language engine should be the ability of *continuously learning* from additional sources. While an initial model can be learned from labeled corpora, the engine should be capable of absorbing new information from unlabeled data that it reads subsequently. A suitable approach for learning from unlabeled data is by means of *Self Training*. Self Training is a semi-supervised learning method where the training corpus is extended with selected unannotated data that the system itself has tagged and that are considered as sufficiently accurate.

The research would aim to develop an efficient *incremental model representation* so that the model can be updated with new evidence, rather than having to retrain a model from scratch.

Achievements. Our team has developed state of the art techniques and solutions for text analytics, information extraction, indexing and search and question answering.

Our systems have achieved top scores in competitions such as CoNLL 2006, 2007, 2008, TREC Question Answering 2000, 2001, 2002, TREC Terabyte 2004, 2005, TREC Blog Mining 2006, Evalita 2007, 2009.

In particular we mention:

- The DeSR dependency parser.
- The Tanl linguistic pipeline, including POS, NER, Super Sense taggers and Coreference resolution.
- The DeepSearch semantic search engine.
- The semantically annotated Wikipedia and applications using it, e.g. Yahoo! Correlator.
- Linguistically directed browsing: Yahoo! Quest.

Demos of these systems are available from http://www.di.unipi.it/~attardi.

References

1. G. Attardi. Experiments with a Multilanguage Non-projective Dependency Parser. In Proc. of the Tenth CoNLL. 2006.
2. G. Attardi, M. Simi, Blog Mining Through Opinionated Words, *Proc. of The Fifteenth Text Retrieval Conference (TREC 2006)*, NIST, Gaithersburg (MD), 2006.
3. G. Attardi, A. Chanev, M. Ciaramita, F. Dell'Orletta and M. Simi. Multilingual Dependency Parsing and Domain Adaptation using DeSR. *Proc. the CoNLL Shared Task Session of EMNLP-CoNLL 2007*. 2007.
4. G. Attardi, M. Simi. DeSR at the Evalita Dependency Parsing Task *Proc. of Workshop Evalita 2007. Intelligenza Aritificiale*, 4(2). 2007.
5. G. Attardi, F. Dell'Orletta. Reverse Revision and Linear Tree Combination for Dependency Parsing. *Proc. of NAACL HLT 2009*. 2009.
6. G. Attardi, F. Dell'Orletta, M. Simi, J. Turian. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. *Proc. of Workshop Evalita 2009*. 2009.
7. G. Attardi, S. Dei Rossi, F. Dell'Orletta, E.M. Vecchi. The Tanl Named Entity Recognizer at Evalita 2009. *Proc. of Workshop Evalita 2009*. 2009.
8. G. Attardi, S. Dei Rossi, F. Dell'Orletta, E.M. Vecchi. Experiments in tagger combination: arbitrating, guessing, correcting, suggesting. *Proc. of Workshop Evalita 2009*. 2009.
9. G. Attardi, F. Dell'Orletta. Chunking and Dependency Parsing. *Proc. of LREC 2008 Workshop on Partial Parsing*, Marrakech, 2008.
10. C. Bosco, A. Mazzei, V. Lombardo, G. Attardi, A. Corazza, A. Lavelli, L. Lesmo, G. Satta, M. Simi. Comparing Italian parsers on a common treebank: the Evalita experience. *Proc. of LREC 2008*, Marrakech, 2008.
11. M. Ciaramita, G. Attardi, F. Dell'Orletta and M. Surdeanu. DeSRL: A Linear-Time Semantic Role Labeling System. *Proceedings the Twelfth Conference on Natural Language Learning*, Manchester, 2008.
12. H. Zaragoza, J. Atserias, M. Ciaramita, and G. Attardi. Semantically annotated snapshot of the English Wikipedia, *Proceedings of LREC 2008*, Marrakech, 2008.
13. Attardi, G., et al.: Tanl (Text Analytics and Natural Language Processing): Analisi di Testi per il Semantic Web e il Question Answering. http://medialab.di.unipi.it/wiki/SemaWiki. 2008.

14. G. Attardi, S. Dei Rossi, G. Di Pietro, A. Lenci, S. Montemagni, M. Simi. A Resource and Tool for Super-sense Tagging of Italian Texts. *Proceedings of LREC 2010*, Malta. 2010.
15. J. Atserias, G. Attardi, M. Simi, H. Zaragoza. Active Learning for Building a Corpus of Questions for Parsing. LREC 2010.
16. G. Attardi, D. Li. Extending a Dependency Treebank with Self-Training. Submitted.
17. G. Attardi, S. Dei Rossi, G. Di Pietro, A. Lenci, S. Montemagni, M. Simi. A Resource and Tool for Super-sense Tagging of Italian Texts. LREC 2010.
18. R. Collobert, J. Wetson. A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. Proc. of the 25th Int. Conference on Machine Learning, Helsinki, Finland. 2008.
19. DARPA. DARPA-BAA-09-03 Machine Reading Broad Agency Announcement (BAA), http://www.darpa.mil/IPTO/solicit/baa/BAA-09-03_PIP.pdf. 2009.
20. G. Attardi, S. Dei Rossi, M. Simi. The Tanl Coreference Tagger at SemEval 2010. *Proc. of SemEval 2010*, 2010.