ESF-NetWords Short Mobility Grant

Grant no.: 4750 – "Distributional models of paradigmatic semantic relations"

Alessandro Lenci
Dipartimento di Linguistica
Università di Pisa
Via S. Maria 36
56126 PISA (Italy)

Email: alessandro.lenci@ling.unipi.it
Webpage: http://www.humnet.unipi.it/linguistica/Docenti/Lenci/index.htm

Actual mobility period: 17 June 2012 – 23 June 2012

Host: Dr. Sabine Schulte im Walde (Institut für Maschinelle Sprachverarbeitung, IMS - Universität Stuttgart,)

1. Purpose of the visit

The purpose of my visit to IMS – Stuttgart with the ESF short mobility grant was to start a long-term collaboration with Dr. Sabine Schulte im Walde to enhance the ability of Distributional Semantic Models (DSMs) to distinguish various types of paradigmatic semantic relations, such as synonymy, antonymy, hypernymy, hyponymy, and co-hyponymy, that greatly differ for their inferential properties (Cruse 1986).

Distributional Semantic Models (DSMs) are a family of computational models that represent word meaning and, even more, measure the distributional similarity between words, under the hypothesis that proximity in distributional space models semantic relatedness (Baroni and Lenci 2010, Schulte im Walde 2006, Turney and Pantel 2010). However, semantically related words differ for the type of relation holding between them: e.g., *dog* is strongly related to both *animal* and *cat*, but with different types of relations (hypernymy and co-hyponymy, respectively). Standard DSMs have difficulties in distinguishing between such relations, because their distributions in text tend to be very similar, cf. *The boy/girl/person loves/hates his cat*, illustrating that the (co-)hyponyms *boy*, *girl*, and *person* as well as the antonyms *love* and *hate* can occur in identical contexts, respectively (Baroni and Lenci 2011, Kotlerman et al. 2010, Lenci and Benotto 2012). The challenge is to understand how corpus-based distributional features can account for the semantic relations. For instance, synonymy and co-hyponymy are reflexive relations, while hypernymy is not, etc. The key role of taxonomic relations in the organization of semantic memory is well-known and fairly uncontroversial (Murphy 2002). Conversely, the extent to which this organization can be derived from syntagmatic co-occurrence statistics is still to be investigated.

Therefore, with this research we expect to contribute to a better understanding of DSMs as psychologically plausible models for meaning acquisition and representation.

In Pisa, I have already started a research line on this topic with Giulia Benotto (PhD student in Linguistics at the University of Pisa) (Lenci and Benotto 2012). Dr. Schulte im Walde is currently the principal investigator of the *DFG Heisenberg Fellowship and Research Grant on Distributional Approaches to Semantic Relatedness*. Therefore, my visit to IMS has provided key opportunities to strengthen the synergies between our two research teams.

2. Description of the work carried out during the visit

The work carried out at IMS-Stuttgart mainly focused on two aspects:

1. exchanging information about the previous and ongoing activities by the Pisa and IMS groups on the analysis of semantic relation with DSMs;
2. planning common research activities on this topic, possibly leading at joint publications.

As for (1), I gave a talk on Monday 18 June 2012, with the title "Semantic Relations in Distributional Semantics: The case of hypernymy", to present the current research carried out in Pisa to develop and test "asymmetric" similarity measures to identify hypernyms with state-of-the-art DSMs. First results of such research has been presented this year at the 1st *SEM Conference (Montreal, Canada, 7-8 June 2012) (cf. Lenci and Benotto 2012). Daily meetings were also held with Dr. Schulte im Walde and members of her research team. In these occasions, the IMS group presented their current work on semantic relations in DSMs. So far, this work has mostly focused on designing and collecting a dataset on German semantic relations. Antonyms, synonyms and hypernyms for 99 German nouns, verbs and adjectives were collected with crowdsourcing methods, with the purpose of using them as a test set to evaluate the ability of distributional methods to discriminate different types of paradigmatic semantic relations. The methodology of data collection was discussed, together with the possibility of its extension to other languages, such as Italian and English.

During this exchange of research information, a strong complementarity emerged between the Pisa and the IMS teams. The former has mainly focused on computational methods for unsupervised semantic relation discrimination (Baroni and Lenci 2011, Lenci and Benotto 2012), while the latter has focused so far on data collection for semantic relation evaluation (Schulte im Walde 2008, Scheible and Schulte im Walde 2012). We have considered this as the ideal situation to foster new synergies between our groups.

As for (2), starting from the current research in Pisa and IMS, we have planned the following common activities:

- Pisa will prepare a dataset for Italian and English, using the same methodology developed by IMS. The latter will provide Pisa with help and support to select the Italian and English stimuli, and to collect the data with Amazon Mechanical Turk. The goal is to create comparable datasets on paradigmatic semantic relations for the three languages, to be used as test set on computational experiments;
- both Pisa and IMS will extend this methodology to collect data about co-hyponyms;
- Pisa will give to IMS the software that has currently developed to implement distributional similarity measures for hypernym identification. So far, these measures have been tested on an existing English dataset (the BLESS dataset; Baroni and Lenci 2011). IMS will apply them to the collected German data. Pisa will also apply them to Italian and English data, after completing their collection;
- Pisa and IMS will jointly work to improve existing methods for hypernym identification, as well as to develop new unsupervised approaches to discriminate other types of paradigmatic relations. In particular, we intend to focus on antonym recognition.

3. Description of the main results obtained

Despite its short length, the ESF visit to IMS was very intense and fruitful. Main results can be summarized as follows:

- we acknowledged the existence of important commonalities in the project goals of Pisa and IMS on the topic of semantic relations and DSMs;
- we exchanged key information about current research carried out at these institutions on such a topic;
- we planned joint work on semantic relations to be carried out together in the following months. In particular, Pisa will complete the data collection for Italian and English by end of September 2012. Then, a phase of computational modeling will start, jointly run by Pisa and IMS;
- we also planned further exchange visits between Pisa and IMS, either by the Principal Investigators, or by PhD students.

These results, as well as the opening of new important research perspectives, have only been possible thanks to my visit to IMS, supported by ESF.

4. Future collaboration with host institution

My visit to IMS has been just the beginning of a promising, long-term collaboration between IMS and Pisa. Existing synergies have been strengthened by the discovery of complementarity of approaches to a common research goal. Intense research exchanges have been planned for the incoming months, to carried out the common research plan that was scheduled during the visit to Stuttgart.

We are also planning to organize a workshop on the topic of semantic relations and distributional semantics, to be submitted to a major computational linguistic conference (e.g., Coling, ACL, *SEM, etc.).

5. Projected publications

Two types of publications are envisaged:

- one conference paper describing the multilingual data collection on semantic relations for German, Italian and English;
- one conference paper on the computational modeling, focusing on the application of DSMs to discriminate the different types of relations and evaluated on the collected datasets.

A more extended journal paper is also planned, describing broad-scale results of the collaboration between Pisa and IMS on this topic.

Since these publications will stem from the work started during the ESF mobility grant, ESF and NetWordS support will explicitly be acknowledged.

References

Baroni, M., and Lenci, A. (2010), “Distributional Memory: A general framework for corpus-based semantics”, *Computational Linguistics*, 36(4): 673-721.

Baroni, M., and Lenci, A. (2011), “How we BLESSed distributional semantic evaluation”, in *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics, EMNLP 2011*, Edinburgh, Scotland, UK: 1-10.

Kotlerman, L., Dagan, I., Szpektor, I., and Zhitomirsky-Geffet, M. (2010), “Directional distributional similarity for lexical inference”, *Natural Language Engineering*, 16(04): 359-389.

Lenci, A. and Benotto, G. (2012), “Identifying hypernyms in distributional semantic spaces”, in *Proceedings of *SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, Montreal, Canada: 75-79.

Murphy, G. L. (2002), *The big book of concepts*, Cambridge, MA: The MIT Press.

Scheible, S., and Schulte im Walde, S. (2012), “Designing a Database of GermaNet-based Semantic Relation Pairs involving Coherent Mini-Networks”, in *Proceedings of the LREC Workshop Semantic Relations II: Enhancing Resources and Applications*. Istanbul, Turkey, May 2012.

Schulte im Walde, S. (2006), “Experiments on the Automatic Induction of German Semantic Verb Classes”, *Computational Linguistics*, 32(2): 159-194.

Schulte im Walde, S., Melinger, A., Roth, M. and Weber, A. (2008), “An Empirical Characterisation of Response Types in German Association Norms”, *Research on Language and Computation* 6(2): 205-238.

Turney, P.D., and Pantel, P. (2010), “From frequency to meaning: Vector space models of semantics”, *Journal of Artificial Intelligence Research (JAIR)*, 37: 141-188.

Pisa, 16 July 2012

Alessandro Lenci