Distributional Semantics beyond Concrete Concepts

Yves Peirsman (yves.peirsman@arts.kuleuven.be)
Research Foundation – Flanders & QLVL, University of Leuven
Blijde-Inkomststraat 21 PO Box 3308, 3000 Leuven, Belgium

Yannick Versley (yannick.versley@unitn.it)
Center for Mind/Brain Sciences, University of Trento
Palazzo Fedrigotti – Corso Bettini 31, I-38068 Rovereto, Italy

Tim Van de Cruys (t.van.de.cruys@rug.nl)
Rijksuniversiteit Groningen
Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, the Netherlands

Central Concern

In the last decade, corpus-based distributional models of semantic similarity and association have slipped into the mainstream of cognitive science and computational linguistics. On the basis of the contexts in which a word is used, they claim to capture certain aspects of word meaning and human semantic space organization. In computational linguistics, these models have been used to automatically retrieve synonyms (Lin, 1998) or to find the multiple senses of a word (Schütze, 1998), among other tasks. In cognitive science, they have been applied to the modelling of semantic priming (Burgess, Livesay, & Lund, 1998; Landauer & Dumais, 1997), semantic dyslexia (Buchanan, Burgess, & Lund, 1996), categorization and prototypicality (Louwerse, Hu, Cai, Ventura, & Jeuniaux, 2005), and many other phenomena. Yet, despite their claims to model human language behaviour, relatively little is known about the precise relationship between these distributional models and human semantic knowledge. While they offer a credible account of (thematic or general) similarity of unary predicates such as concrete nouns, the question remains if and how more complex knowledge can be modelled using distributional information. This workshop therefore wants to focus on new challenges to distributional approaches that lie beyond the traditional modelling of concrete concepts.

Challenges to distributional models

Three current challenges to distributional semantics take a central position in our workshop. These are the discovery of verb meaning, the modelling of different aspects of semantics and the combination of different types of data. We want to address each of these, with specific attention to the relationship between distributional models and human semantic cognition.

Modelling verb meaning

A first challenge is the modelling of verb meaning. The results of a related workshop at the European Summer School in Logic, Language and Information (ESSLLI-2008), showed that verb clustering is a much more difficult task than noun clustering (Peirsman, Heylen, & Geeraerts, 2008; Van de Cruys, 2008; Versley, 2008). This may have a number of reasons. One is the fact that, in order to model the meaning of a verb, we may require information about all the possible arguments that verb can have. These arguments can be realized in different ways in text, and need not always be close to their governing verb. This situation differs markedly from that of nouns, whose meaning can often be modelled reasonably well on the basis of the adjacent adjectives. It is thus an open question what precise information distributional models of verb semantics should take into account (Schulte im Walde, 2008). One possible source of inspiration is the Featural and Unitary Semantic Space (FUSS) model (Vigliocco, Vinson, Lewis, & Garrett, 2004), which models objects and actions in the same semantic space and explicitly claims to mirror semantic representations in humans. A second reason for the difficulty distributional models experience with verb semantics is the absence of an uncontroversial Gold Standard. WordNet, FrameNet and other sources all present different and fuzzy classifications that humans may not always agree on. Different Gold Standards may thus favour different sources of information. In short, when it comes to verb semantics, we are faced with a double question: *what* do we want to model, and *how* can this be done?

Modelling different aspects of semantics

In spite of this difficulty, the discovery of verb semantics is still a 'traditional' task for distributional models, which are typically used for the discovery of semantic relations like similarity (between *plane* and *airplane*) or association (between *plane* and *airport*). We may thus ask ourselves what other information, apart from these two types of relations can be collected from linguistic data. One current endeavour in this field is the automatic generation of concept properties from corpora. The StruDEL system (Baroni & Lenci, forthc.), for instance, aims to retrieve from a corpus the semantic features of a given concept, like like *flies* or *lays_eggs* for the concept ROBIN (McRae, Cree, Seidenberg, & McNorgan, 2005). It does so by combining the traditional distributional approach with pattern-based search (Hearst, 1992). StruDEL indeed gives a better fit to human norms, but still differs significantly in a number of respects, both from the traditional distributional approach and from the properties given by people. This

suggests that the two approaches might model different aspects of semantic cognition. Again, it raises the issue of what dimensions of semantics are captured by what type of linguistic information.

Combining different types of data

Despite these positive results, distributional models of lexical semantics can only make a partial claim to the modelling of human semantic cognition. After all, when children learn the meaning of words, they probably make use of much more information than just the linguistic context that the words occur in. Andrews, Vigliocco, and Vinson (in press) argue that in order to arrive at realistic models of semantic cognition, the distributional approach therefore has to be complemented with experiential data that captures our experience with the physical world. This combination of data types provides a solution to the problems of the individual approaches: the distributional perspective offers a more realistic view of learning about abstract concepts than purely experiential models, while the experiential perspective provides the grounding in the physical world that is lacking in distributional approaches. This study thus shows how the combination of different types of information gives a more robust and realistic model of human semantic cognition. More generally, it illustrates that researchers of distributional semantics should be open towards other, possibly competing, approaches to lexical semantics that may enrich our knowledge about semantic cognition.

Conclusion

Central to all these challenges is the relationship of distributional models to human semantic cognition. What are the main differences between the conceptualization of verbs and that of nouns? What aspects of human property spaces are found in large corpora, and how? And how does distributional learning relate to other types of learning? These are the main questions our workshop would like to address. To this goal, we would like to attract researchers from both cognitive science and computational linguistics that have been concerned with distributional models. We believe that uniting these two perspectives can lead to a fruitful discussion about the present and future of distributional semantics. In this way, we would like to reconcile different directions in research on distributional semantics, and outline relevant paths for future research.

References

Andrews, M., Vigliocco, R., & Vinson, D. (in press). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*.

Baroni, M., & Lenci, A. (forthc.). Concepts and properties in word spaces. *From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science (Special issue of the Italian Journal of Linguistics)*.

Buchanan, L., Burgess, C., & Lund, K. (1996). Overcrowding in semantic neighborhoods: Modeling deep dyslexia. *Brain and Cognition*, *32*, 111–114.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, *25*, 211–257.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING 1992)* (pp. 539–545).

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL98* (pp. 768–774). Montreal, Canada.

Louwerse, M. M., Hu, X., Cai, Z., Ventura, M., & Jeuniaux, P. (2005). The embodiment of amodal symbolic knowledge representations. In I. Russell & Z. Markov (Eds.), *Proceedings of the 18th International Florida Artificial Intelligence Research Society* (pp. 542–547). Menlo Park, CA: AAAI Press.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods, Instruments, and Computers*, *37*, 547–559.

Peirsman, Y., Heylen, K., & Geeraerts, D. (2008). Size matters: Tight and loose context definitions in English word space models. In M. Baroni, S. Evert, & A. Lenci (Eds.), *Proceedings of the ESSLLI workshop on distributional lexical semantics. Bridging the gap between semantic theory and computational simulations* (pp. 34–41).

Schulte im Walde, S. (2008). Human associations and the choice of features for semantic verb classification. *Research on Language and Computation*, *6*(1), 79–111.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, *24*(1), 97–124.

Van de Cruys, T. (2008). A comparison of bag of words and syntax-based approaches for word categorization. In M. Baroni, S. Evert, & A. Lenci (Eds.), *Proceedings of the ESSLLI workshop on distributional lexical semantics. Bridging the gap between semantic theory and computational simulations* (pp. 47–54).

Versley, Y. (2008). Decorrelation and shallow semantic patterns for distributional clustering of nouns and verbs. In M. Baroni, S. Evert, & A. Lenci (Eds.), *Proceedings of the esslli workshop on distributional lexical semantics. bridging the gap between semantic theory and computational simulations* (pp. 55–62).

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, *48*(4), 422–488.