Distributional Semantic Features as Semantic Primitives – or not

Gemma Boleda
Department of Translation and Language Sciences
Universitat Pompeu Fabra, Barcelona
gemma.boleda@upf.edu

Katrin Erk
Linguistics Department
University of Texas at Austin
katrin.erk@mail.utexas.edu

Introduction

In the past, approaches to meaning that relied on semantic primitives were popular both in Linguistics (Jackendoff 1990) and Artificial Intelligence (Schank and Tesler 1969; Wilks 1975). For instance, a representation for the sentence (1) could be as in (2) (Dorr 1992, 255):

(1) John went to the store

(2) $[_{Event}GO_{Loc}([_{Thing}John],$
$[_{Position}TO_{Loc}([_{Thing}John], [_{Location}Store])])]$

Word meanings have also long been represented in terms of semantic primitives (Fodor et al. 1980). For instance, the meaning of *man* can be analyzed as in (3):

(3) $[+HUMAN, +MALE]$

Semantic primitives have traditionally served several purposes (Jackendoff 1990; Schank and Tesler 1969; Wilks 1975; Winograd 1978). We will focus on the following ones:

a) to capture **conceptual /** real-world aspects of word meaning, for instance the HUMAN feature applies to objects described by *man* and *woman*, while MALE applies to *man*, *boy*, and *stallion*;

b) to formalize what is common to nearly **synonymous expressions** within a single language (*John gave a book to Mary*, *Mary was given a book by John*, and *Mary received a book from John*), or in different languages (*John went to the store* and Spanish *John fue a la tienda*), e.g. for Interlingua approaches to Machine Translation (Dorr 1992);

c) to account for **inferences** that a speaker is able to make: From the representation in (2), given the adequate specifications, we can induce that John's location at the end of the event expressed by sentence (1) was the store. Also, from (3), again given some processing mechanism, we can account for the fact that *John is male* follows from *John is a man*.

While semantic primitives have been criticized both on philosophical (Fodor et al. 1980) and psychological grounds (Fodor, Fodor, and Garrett 1975), and have been largely abandoned in AI, the problems that primitive-based knowledge representations were meant to address remain challenging. For that reason, we think it makes sense to use primitive-based approaches as a point of comparison to shed light on a currently popular approach to computationally representing word and phrase meaning: distributional semantics.

Distributional models use large corpora to learn a meaning representation for a target word based on context items occurring in close proximity of the target. This representation is a vector. In the simplest case (called *count*-based in (Baroni, Dinu, and Kruszewski 2014)), its dimensions stand for context items observed to co-occur with the target (Turney and Pantel 2010). It can also be *prediction*-based, a word embedding with dimensions that are latent classes (Mikolov, Yih, and Zweig 2013). Our discussion encompasses both types of distributional models, as both compute some form of vector representation based on observed co-occurrence with context items.

Like the semantic primitive representation in (3), the distributional representation for the word *man* consists of a collection of features. In distributional models, there are typically more features –from a few dozen to hundreds of thousands–, their values are continuous rather than binary, and the representation is automatically induced rather than manually constructed. But does this make a distributional representation fundamentally different from a primitive-based one? Unlike primitive-based approaches, we will argue, distributional semantic models do not try to find features that are more primitive than the lexical items they are used to describe, nor do they try to identify individual features that give rise to inferences. Instead, they use a large number of features that may individually be weak, but that in the aggregate allow us to make inferences similar to those licensed by semantic primitives.

Semantic primitives and distributional semantic features

We explore some crucial characteristics of semantic primitives (Wilks 1975; Winograd 1978; Fodor et al. 1980; Geeraerts 2010), and ask whether they also match distributional semantics.

Basic units of semantic representation and cognition. Primitives are supposed to express minimal units for semantic representation that are at the basis of cognition. Words are decomposed into primitives, which represent the word meanings. So as not to enter infinite loops, primitives should be **irreducible**[1] and **non-linguistic**. Indeed, if, say, we define *bachelor* as *unmarried man*, we need to further define *unmarried* and *man*. At some point, either some atomic level will be reached, or a circular system will emerge (as it happens in general-purpose dictionaries). Also, the atomic elements should not be linguistic in nature but *grounded* in some other type of information, otherwise they are not semantic representations, but simply translations themselves in need of an interpretation (Searle 1984). This set of basic elements should be significantly **smaller** than the set of words they need to define; else, trivially, every word could be one primitive, and then no explanation would be achieved, as all meaning representations would be completely disjoint from one another (Fodor, Fodor, and Garrett 1975). However, they should still be **comprehensive**, that is, they should jointly express the same meaning the word expresses (Wilks 1975; Winograd 1978).

These characteristics are problematic in multiple respects. First, if not linguistic, what can primitives be? A reasonable source are sensory-motor properties, but it is doubtful that all words can be reduced to sensory-motor properties (Fodor et al. 1980): Think of *grandmother* in *This is my cat's grandmother*. Alternative non-linguistic anchorings of primitives are hard to come by. Second, it is far from clear that it is possible to determine empirically what the set of basic elements of cognition is, in a similar way as the periodic table was established in chemistry (Winograd 1978) — or indeed whether they exist in the first place. This drastically affects the possibility of building practical applications based on primitives. Third, and relatedly, if semantic primitives were cognitively real, you would expect effects. Words whose decomposition includes negation (*bachelor* as *man who has never been married*) should be more complex to process than words that do not – but no experimental evidence was found for this in (Fodor, Fodor, and Garrett 1975). Fourth, if there are significantly fewer semantic primitives than words, meaning nuances will inevitably be lost. For instance, a common analysis of *to kill* can informally be expressed as *to cause someone to become not alive*. There are many causing-to-die situations that do not correspond to killing situations.

Distributional features are not necessarily primitives in any cognitive, perceptual, or other way: They are simply summaries of contexts that words appear in. However, crucially, they do capture information that is analogous to the information primitives were designed to capture. If they didn't capture conceptual information, it would be hard to explain why distributional models can reproduce human judgements and behavior on a wide range of semantic tasks: Word similarity (*string-cord* are similar, *professor-cucumber* are not), synonym identification both for words (TOEFL task; (Landauer and Dumais 1997)) and phrases (*personal appeal-charisma*, (Dinu and Baroni 2014)), categorization (Baroni and Lenci 2010), property identification for known objects (Bruni et al. 2012) and for unknown objects (Lazaridou, Bruni, and Baroni 2014; Erk 2014), among many others. Even information that has been traditionally represented as a semantic feature, such as ±MALE, can be captured in distributional semantics: Recent work (Mikolov, Yih, and Zweig 2013) showed that $King - Man + Woman$ (where $King$ etc. are vectors for the corresponding words) produces a vector that is very close to the $Queen$ vector.

Distributional features are also not irreducible: Any item that functions as a dimension (context) can also be a target that gets a distributional representation in turn. This implies that there is no reduction of words to a set of simple features, since everything can be a target and a feature (Baroni and Lenci 2010). Moreover, the set of distributional features can be very large, and even relatively small distributional models provide more flexible representations than a typical semantic primitive representation, because they consist of more features – Mikolov, Yih, and Zweig (2013) use 50, Bruni et al. (2012) use 30,000, and 300 is a typical number for SVD-reduced models –, with values that are continuous as opposed to binary. This allows distributional representations to express many nuances of meaning, that is, to be fairly comprehensive.

Finally, distributional features are not committed to a specific source for features. They can be linguistic (textual), but also non-linguistic (such as visual, with information automatically extracted from images), and in fact it has been shown that the combination of different modalities improves semantic representations (Bruni et al. 2012; Roller and Schulte im Walde 2013). This allows distributional models to be *grounded* in sensory properties, while at the same time retaining higher-level information uniquely expressed through language. For instance, in Bruni et al. (2012), textual models accounted better for race-related uses of the terms *black* and *white*, including modification of abstract nouns (*black project*), while visual models fared better for more directly perceptual uses (*white car*).

Inference-able. In a semantic primitives system, as discussed above, the +MALE feature value allows for the inference *John is male* from the statement *John is a man*. Given the right features, inference on relations such as hypernymy (*man* IS-A *person*) is straightforward. Distributional models have a different notion of inference, or maybe a different set of notions (as there is no uniform framework that everybody uses), often based on the idea that an item is inferred if its

[1] Here a caveat is in order: The irreducibility property holds at one given level of description (Winograd 1978). For instance, in chemistry we find a table of atomic or primitive elements, with specific rules as to how to combine them, with which any physical substance can be described. The fact that these elements can be analyzed in terms of sub-atomic particles does not alter their functioning as primitives with respect to chemistry. In semantics, taking MALE as primitive means that it works as such in our system, and indeed it does account for relevant facts, from gender systems in language to categorization decisions in psycholinguistic experiments.

representation is close by, for example a word is inferred to be a synonym if its vector is close to the vector of the target. This notion of inference is inherently graded and weighted, so it fits better in a probabilistic framework than in a framework of hard inference. A type of inference that distributional models arguably are particularly suited to is determining whether two words are near-synonyms in a given sentence context (Erk and Padó 2008), because they can capture meaning, and the way that it is affected by context, through a large amount of nameless features, as opposed to trying to capture all the nuances of word meaning through an explicitly given set of senses. Hypernymy inferences have until recently been thought to be beyond what distributional models can do, but several recent papers indicate that it may be possible to extract hypernymy judgments through dedicated similarity measures (Lenci and Benotto 2012; Roller et al. 2014), though it is too early to tell how robust this inference can be. Another recent example of distributional inference is Lazaridou, Bruni, and Baroni (2014), who did image labeling based on the hypothesis that distributionally similar terms will look similar in images. All these types of inferences can be viewed as special cases of property inference, where properties of a concept are inferred based on the hypothesis that distributionally similar concepts have similar properties (Erk 2014).

Discussion and conclusion

If we assume semantic features that do not have to be irreducible and that are allowed to take on continuous values, they could in principle provide all the functionality that distributional representations offer, *if they were sufficiently fine-grained* – but one of the core points that we want to make is that it is too hard to determine such a fine-grained set of semantic primitives. In contrast, it is doable to collect a large number of features automatically that are not (and do not need to be) individually inference-enabling, just in the aggregate. These features can be textual or non-textual.

This highlights two major assets of distributional semantics as a model of natural language meaning. First, the fact that, because its features are not meant to be primitives and it is not the individual features but their aggregate that is relevant, we do not have to care to select individual features correctly; it suffices to choose an overall good class of features. Thus, unlike primitive-based systems, distributional semantic models are robust, and their performance degrades gracefully (if, for a given task, 300 dimensions yields the maximum score, a move to 250 dimensions will not make the system crash, only decrease its performance). Second, distributional semantics comes with a well-defined learning mechanism to induce semantic representations from naturally occurring data. Typically, the experimenter chooses a definition for *context* (say, two lemmas to the left and two lemmas to the right of the target word) and a set of data (say, a dump of the Wikipedia), and both the features and their values will be automatically induced from data naturally produced by humans: Sentences, documents, image labels, etc.[2] This is not only an advantage from an engineering point of view, but is also something that makes these models more plausible from a cognitive perspective (Landauer and Dumais 1997; Baroni et al. 2010; Murphy et al. 2011), since all humans learn from language in context. Retaking the list of purposes provided above, distributional semantics

(a) provides a good proxy for conceptual features;

(b) is able to provide similar representations for synonymous expressions, not only for words, but also for phrases;

(c) supports some forms of inference.

Thus we argue that distributional semantics can serve as the basis for a semantic representation of words and phrases that serves many of the purposes semantic primitives were designed for, without running into many of their philosophical, empirical, and practical problems. Distributional models license graded, similarity-based kinds of inferences that differ fundamentally from inferences over distinct categories.

Still, distributional approaches face several critical problems that need to be addressed. First, though we have many individual examples of distributional inference, a more general characterization of what distributional inference is and what purposes it can serve remains to be done. A second core problem is compositionality. When semantic primitives are used as predicate symbols, as in (2), they inherit the usual notion of compositionality from logic. But how can a representation of a phrase be constructed from distributional representations of its part? There has been some work on compositional distributional phrase representations (Coecke, Sadrzadeh, and Clark 2011; Baroni and Zamparelli 2010; Socher et al. 2012). An alternative option is to combine distributional information at the lexical level with logical form (and its usual notion of compositionality) at the sentence level (Beltagy et al. 2013). Incidentally, a compositionality-related problem that is not solved yet for either distributional or primitive-based approaches is how composition of feature representations can address polysemy by selecting appropriate features. (Zeevat et al. (2014) constitute a recent approach using primitives.)

So should distributional representations be viewed as an *alternative* to primitives? In fact, there is a better option, namely to integrate the two. Using prominent human-defined features when they are available will clearly make for a model with better fit, and using distributional features will make for a model that is more robust and that captures facets of lexical meaning even when it is not easily explicitly specifiable. At a technical level, first steps in this direction have been taken by multi-modal distributional approaches that integrate distributional features with human-defined features (Andrews, Vigliocco, and Vinson 2009; Johns and Jones 2012; Roller and Schulte im Walde 2013).

Acknowledgements

The first author acknowledges support by the Spanish Ministerio de Economía y Competitividad under project FFI2013-41301-P. The second author acknowledges support by the National Science Foundation under grant 0845925.

[2]Different definitions of context can be combined in various ways, see e.g. (Roller and Schulte im Walde 2013).

References

Andrews, M.; Vigliocco, G.; and Vinson, D. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review* 116(3):463–498.

Baroni, M., and Lenci, A. 2010. Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics* 36(4):673–721.

Baroni, M., and Zamparelli, R. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Baroni, M.; Murphy, B.; Barbu, E.; and Poesio, M. 2010. Strudel: a corpus-based semantic model based on properties and types. *Cognitive science* 34(2):222–54.

Baroni, M.; Dinu, G.; and Kruszewski, G. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, 238–247. Baltimore, Maryland: Association for Computational Linguistics.

Beltagy, I.; Chau, C.; Boleda, G.; Garrette, D.; Erk, K.; and Mooney, R. 2013. Montague meets Markov: Deep semantics with probabilistic logical form. In *Proc. of the Second Joint Conference on Lexical and Computational Semantics (*SEM)*.

Bruni, E.; Boleda, G.; Baroni, M.; and Tran, N. K. 2012. Distributional Semantics in Technicolor. In *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics*, 136–145.

Coecke, B.; Sadrzadeh, M.; and Clark, S. 2011. Mathematical foundations for a compositional distributed model of meaning. *Linguistic Analysis* 36(1-4):345–384. A Festschrift for Joachim Lambek.

Dinu, G., and Baroni, M. 2014. How to make words with vectors: Phrase generation in distributional semantics. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, 624–633.

Dorr, B. J. 1992. A Two-Level Knowledge Representation for Machine Translation: Lexical Semantics and Tense/Aspect. In *Proc. of the First SIGLEX Workshop on Lexical Semantics and Knowledge Representation*, 269–287. London, UK, UK: Springer-Verlag.

Erk, K., and Padó, S. 2008. A structured vector space model for word meaning in context. In *Proc. of the 2008 Conference on Empirical Methods in Natural Language Processing*, 897–906.

Erk, K. 2014. What do you know about an alligator when you know the company it keeps? Draft.

Fodor, J. a.; Garrett, M. F.; Walker, E. C.; and Parkes, C. H. 1980. Against definitions. *Cognition* 8(3):263–7.

Fodor, J.; Fodor, J. A.; and Garrett, M. F. 1975. The Psychological Unreality of Semantic Representations. *Linguistic Inquiry* 6(4):515–531.

Geeraerts, D. 2010. *Theories of Lexical Semantics*. Oxford University Press.

Jackendoff, R. S. 1990. *Semantic structures*. Cambridge, MA (etc.): The MIT Press.

Johns, B. T., and Jones, M. N. 2012. Perceptual inference through global lexical similarity. *Topics in Cognitive Science* 4(1):103–120.

Landauer, T. K., and 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.

Lazaridou, A.; Bruni, E.; and Baroni, M. 2014. Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, 1403–1414.

Lenci, A., and Benotto, G. 2012. Identifying hypernyms in distributional semantic spaces. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, 75–79.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proc. of NAACL-HLT 2013*, 746–751.

Murphy, B.; Poesio, M.; Bovolo, F.; Bruzzone, L.; Dalponte, M.; and Lakany, H. 2011. EEG decoding of semantic category reveals distributed representations for single concepts. *Brain and Language* 117:12–22.

Roller, S., and Schulte im Walde, S. 2013. A Multimodal LDA Model integrating Textual, Cognitive and Visual Modalities. In *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1146–1157.

Roller; S.; Erk, K.; and Boleda, G. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proc. of CoLing 2014*, 1025–1036.

Schank, R. C., and Tesler, L. 1969. A Conceptual Dependency Parser for Natural Language. In *Proc. of COLING '69*, 1–3.

Searle, J. 1984. *Minds, Brains and Science*. Harvard University Press, Cambridge, MA.

Socher, R.; Huval, B.; Manning, C. D.; and Ng, A. Y. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proc. of the joint meeting of the Conference on Empirical Methods in Natural Language Processing and the Conference on Computational Natural Language Learning (EMNLP-CoNLL)*.

Turney, P. D., and Pantel, P. 2010. From Frequency to Meaning : Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37:141–188.

Wilks, Y. 1975. Preference Semantics. In Keenan, E. L., ed., *Formal Semantics of Natural Language*. Cambridge: Cambridge University Press. 329–348.

Winograd, T. 1978. On primitives, prototypes, and other semantic anomalies. In *Proc. of the 1978 workshop on Theoretical issues in natural language processing (TINLAP'78)*, 25–32.

Zeevat, H.; Grimm, S.; Hogeweg, L.; Lestrade, S.; and Smith, E. A. 2014. Representing the lexicon. Draft.