Probabilistic Models of semantics as compared to previous approaches

Ingmar Schuster

August 25, 2014

> It may be presumed that any two morphemes A and B having different meanings, also differ somewhere in distribution. *Harris (1951)*

While the terms *Vector Space Semantics* and *Distributional Semantics* (and many others) are used interchangeably in the literature for models inferring a vector representation of lexical meaning from co-occurrence information. While all of these models use vectors or higher order tensors in one way or the other to represent the latent meaning of words (lexical semantics), very few if any use distributions in the statistical sense. This is surprising given the fact that researchers acknowledge the central role statistics plays in acquiring latent meaning representations (Lenci, 2008).

From the perspective of probability theory, all data is generated from some underlying probability distribution. Probabilistic models are often constructed by careful combination of simple probability distributions to explain a more complex data-generating process, possibly by assuming latent variables (such as vectors representing latent word meaning). This is the position from which I develop a new approach to Distributional Semantics: I model word meaning and the composition of words in phrases in the probabilistic framework. Probability theory has the advantage of being a very flexible framework for model building which has been and is being extensively researched and has solid mathematical foundations. A very natural approach of inferring the probabilities for values of latent variables and parameters in probabilistic model is the application of Bayes Theorem Because of this, I use Markov Chain Monte Carlo algorithms whenever the data-generating distribution resulting from a model cannot be derived analytically, which is typically the case.

One of the main open issues in distributional semantics is a model of compositionality, i.e. a model of how the meaning of individual words (such as the words *aggressive* and *dog*) combines in multi-word phrases (such as *aggressive dog*). Previous approaches tend to do this by inferring a representation for phrases, which often are new vectors (or higher-order tensors) representing phrase meaning (Baroni & Zamparelli, 2010; Guevara, 2010; Clarke, Weir, & Lutz, 2011; Baroni & Lenci, 2010; Clark & Coecke, 2008). This is a very interpretable solution for adjective-noun phrases: the latent meaning of a noun (represented as a vector over $\mathbb{R}$) is simply modified by the latent meaning of an adjective (sometimes represented as a matrix over $\mathbb{R}$, which can be applied as a linear function to

modify noun vectors). There are several problems with this. For one, interpretability breaks down as soon as other types of phrases, such as noun-verb phrases are considered (Given a latent vector representation for *the man*, how does the verb *run* modify this in *The man runs*?). Also, consider transitive verbs such as *give*, which result in phrases made up of more than two words (e.g. *Lucy give me a book*). If adjectives are represented as matrices, a consequent step would be to represent transitive verbs as a higher-order tensor. However, some verbs can be used in a transitive and intransitive way (consider for example *I told you the story* vs. *I told you so*), so more several representations for a word would become necessary.

I do away with all of these difficulties, and some that I am not discussing here, by not trying to represent the meaning of phrases. Instead, I propose to construct models which assign probabilities to phrases. This works with adjective-noun phrases: a fitted model could, for example, assign a higher probability to the adjective-noun phrase *aggressive dog* as compared to the probability *green dog*. Probabilities are inherently interpretable: one is more likely to see an occurrence of *aggressive dog* then an occurrence of *green dog*. Also, transitive verbs do not pose a problem: instead of trying to represent a phrase containing a transitive verb, we simply assign a probability to it. Phrases with transitively used verbs can easily be compared to phrases with the same verb used intransitively, just by comparing the probability assigned by a fitted model.
My first approach to a model of semantics based on probability theory uses linear probabilistic matrix factorization to model two-word phrases. In the fitted model, given latent variables z_{t_1}, z_{t_2} for two types t_1, t_2, the probability of seeing a phrase $t_1\ t_2$ is proportional to $z_{t_1}^T W z_{t_2}$ for some weight matrix W. While this model meets the goal of assigning probabilities to phrases, it does not in principle assign probabilities to phrases consisting of more than two words. This problem is solved by my compositionality model based on Gaussian Processes. Given latent variables $z_{t_1}, z_{t_2}, \ldots, z_{t_k}$ (where k is variable), a Gaussian Process (GP) is a model for a function f assigning a scalar output to the vector $(z_{t_1}^T, z_{t_2}^T, \ldots, z_{t_k}^T)^T$. By fitting the latent variables (z_{t_l} for type l) and the GP to the actual frequency of phrases in the corpus, we can assign a probability to any multi-word phrase (up to proportionality) by evaluating $f((z_{t_1}^T, z_{t_2}^T, \ldots, z_{t_k}^T)^T)$ for the phrase $t_1, t_2, \ldots, t_k$.

My GP-based model has the advantage of being a straight-forward statistical model of semantic compositionality. While vectors representing the meaning of words are mostly inferred based on statistics, an odd tendency in literature on compositionality in distributional semantics is the use of non-statistical models. Some papers use tensor products to represent compositional meaning (Widdows, 2008; Clark & Pulman, 2007; Clark & Coecke, 2008), others are basically reviews of linear algebra (Clarke et al., 2011). While some of the previous literature uses a statistical approach to semantic compositionality, to our knowledge only the case of two word phrases has been tackled to date (Guevara, 2010; Baroni & Zamparelli, 2010).

Acknowledgements

This position paper gained considerably by consulting the literature reviews of Turney and Pantel (2010) and especially Lenci (2008).

References

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673–721. Retrieved from http://dl.acm.org/citation.cfm?id=1945049

Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1183–1193). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1870773

Clark, S., & Coecke, B. (2008). A compositional distributional model of meaning. *Proceedings of the Second Quantum Interaction Symposium (QI-2008)*, 133–140. Retrieved from http://www.pps.jussieu.fr/ mehrs/AAAI_2008.pdf

Clark, S., & Pulman, S. (2007). Combining Symbolic and Distributional Models of Meaning. In *Aaai spring symposium: Quantum interaction.* Retrieved from http://www.aaai.org/Papers/Symposia/Spring/2007/SS-07-08/SS07-08-008.pdf

Clarke, D., Weir, D., & Lutz, R. (2011). Algebraic Approaches to Compositional Distributional Semantics. In *Proceedings of the ninth international conference on computational semantics* (pp. 325–329).

Guevara, E. (2010). A Regression Model of Adjective-Noun Compositionality in Distributional Semantics. In *Proceedings of the 2010 workshop on geometrical models of natural language semantics.* (pp. 33–37).

Harris, Z. S. (1951). *Methods in structural linguistics.* University of Chicago Press.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, *1*, 1–31. Retrieved from http://linguistica.sns.it/RdL/20.1/ALenci.pdf

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

Widdows, D. (2008). Semantic vector products: Some initial investigations. In *Proceedings of the second aaai symposium on quantum interaction.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.116.4752&rep=rep1&t