Dr. Strangestats or: How I Learned to Stop Worrying and Love Distributional Semantics

Marco Baroni

I was a teenage generativist. I was raised in fairly observant Chomskyan schools, and I still abide by the program for linguistics as the algorithmic study of human language competence Chomsky laid out 60 years ago. Then, how did I become an adult werelinguist, theoretical semanticist by day, corpus-based, statistics-driven computationalist at night?

I don't do corpus-based, statistics-driven distributional semantics because I am, in principle, attracted by or sympathetic towards usage-based, non-symbolic, inductive approaches to language. I do distributional semantics because at a certain point I discovered that it is the only semantic formalism allowing me to do my job as a linguist.

I first felt the need for "semantics" while writing my master thesis about derivational morphology, where the salience of morpheme boundaries predicts phenomena such as the likelihood that an affix undergoes phonetic reduction, blocking of phonological rules, morphemic-route access in lexical retrieval, etc. But one of the main factors determining, in turn, the salience of morpheme boundaries is *semantic transparency*, that is, the extent to which the meaning of a derived word is related to the meaning of its stem, (cf. *re-decorate* vs. *recollect*).

I then started looking around for an approach to semantics that would (i) provide large-scale coverage of the lexicon and (ii) make quantitative predictions about degrees of similarity (or relatedness). The first requirement came from the fact that I needed to account for the often semantically arbitrary sets of stems and derived forms that were subject to specific morpheme salience phenomena. The second requirement derived from the fact that, in all phenomena I looked into, the effect of semantic transparency was never "all or nothing", but rather a fuzzy phenomenon with many intermediate cases, so I needed a theory making graded predictions.

Formal approaches to semantics, even those that paid attention to lexical meaning, failed both requirements. The "functionalist" stuff, while in principle sympathetic to the idea of degrees of similarity, was too awfully fuzzy, not explicit enough to make quantitative predictions, and in any case failing the coverage requirement.

Unfortunately, I discovered distributional semantics too late to use it in my morphology work, where I just gave up the idea of accounting for semantic transparency effects, but from when, years later, I discovered LSA and its cousins, I never found a reason to go back to other approaches to semantics, simply because, from a practical point of view, I still see no alternative to the distributional approach.

Something I've learned along the way is that being able to quantify degrees of semantic similarity is not only good for tasks such as assessing the semantic transparency of derived forms or finding near synonyms. Distributional semanticists (including some that will attend this seminar) came up with clever and elegant ideas to account, in terms of semantic similarity, for complex linguistic phenomena such as predicting the selectional preferences of verbs, capturing argument alternation classes or accounting for co-composition effects.

And there is ongoing and very promising work (that, I think, will be discussed at the seminar) on dealing with fundamental challenges for distributional semantics such as polysemy or scaling up to phrase and sentence meaning.

So, while there is a lot of hard work ahead of us, I'm confident that in a few years we will have empirically successful models of distributional semantics that are not limited to single words in isolation, and, equipped with these new models, we will be able to account for many more linguistic phenomena in terms of semantic similarity.

Still, current distributional semantics is entirely prisoned inside a linguistic cage: all it can tell us (and that's not little!) is how similar words, phrases and sentences are to each other. Without a hook into the "outside world", all we will be able to do is to measure how similar, say, the sentence *A boy is laughing* is to *A girl is crying*, but we will never be able to tell whether either sentence can be truthfully asserted of the current state of the world.

While I understand that there is much more to the "outside world" than this, I think that one first, reasonable step we can take is to explore whether we can connect distributional semantic representations with our visual perception of the world. In concrete, we should aim for a system that, given a

picture depicting a scene with, say, a laughing boy, could tell us that *A boy is laughing* is an appropriate statement describing the scene.

Interestingly, state-of-the-art image analysis systems represent images not unlike distributional semantics represents words – that is, images are represented by vectors that record the distribution of a set of discrete feature occurrences in them. So, there is hope, and I think a central goal for distributional semantics in the next few years should be to work on how to develop a common semantic space, where vector-based representations of linguistic expressions, on one side, and objects and scenes, on the other, can be mapped and compared.

Given such shared linguistic-visual semantic space, the same similarity scoring techniques we are already using in distributional semantics might be extended to account for referential aspects of meaning: The sentence *A boy is laughing* is truthfully stated of (a picture depicting) a scene if the vector representing the sentence and the vector representing the scene are above a certain threshold of similarity.

My colleagues and I are currently working on methods to build the proposed shared linguistic-visual semantic space (and other researchers are also making good progress in this direction). At the seminar, I would like to discuss (among many other things, of course!) both concrete ideas about how to construct the common space, and what are linguistically interesting scenarios in which we could make use of it.