Distributional Semantics for IR

Eduard Hovy
CMU

From its beginning, IR had the good fortune of using a form of distributional semantics, albeit a rather weak one. In fact, one might say that this is the reason IR works. But the time has come to upgrade IR to a more advanced semantics, since that is likely to deliver better results.

Distributional semantics, as stated by Firth, is usually anchored on single words: "you shall know a word by the company it keeps". A word signature vector, created for every word and using a variety of scoring formulas to compute the association strengths of words that co-occur with it, is the traditional implementation. Such vectors have been used over the past decade in NLP to perform word sense disambiguation, find good translation words, identify paraphrases and coreferences, and other tasks. However, in this form distributional semantics is a misnomer: to be a true semantics it is required that a formal operation of compositionality be defined, allowing two word vectors to be composed into a third that expresses the (distributional) semantics of the composite. In not-truly-semantic distributional semantics, there is no composition. So although we think of "the capital of New Jersey" as somehow a composite of the vectors for "capital" and "New Jersey", there is no representation of the fact that "of" signals a specific relationship, or that the non-city senses of "capital" (as in money) and associated words should be removed. Thus the resulting vector has little overlap with the vector for "Trenton".

IR adopted the same form of not-truly-semantic distributional semantics, by representing a document as a distribution of the unstructured bag of [content] words that comprise it, and a query similarly (perhaps extended through relevance feedback), both of which can be implemented by a distributional vector (albeit a sparse one for a query). In the sense of distributional semantics, IR treats these distributions as the 'semantics' of the document or query, and relies upon the mutually disambiguating effects of words to pinpoint the semantics more of less as a side effect: "bank"+"deposit"+"money" is a different vector from "bank"+"deposit"+"mud", and even without making systems try to differentiate the semantics of the senses of "bank" and "deposit", "money" and "mud" are different enough to separate out the meanings and reinforce the intended one.

Recent research on distributional semantics has devoted quite a lot of attention to the problem of compositionality. It is required for a semantics that smaller meanings can be composed into larger ones that properly express the 'larger' meaning. How large is feasible —a paragraph? a document?— is not yet known. Despite a lot of fancy work, some wild claims about tensors and quantum mechanics, and brave attempts to construct corpora and evaluation tests that would help guide research, it doesn't seem that a satisfactory solution has been found. It is possible today to create distributional structures to represent words and other units

and to compose them in various ways for various effects, sometimes surprisingly successful, but there is no clear overall model for how it works.

An approach that we have been pursuing in my group at CMU resembles work by Baroni and Lenci and others. Rather than creating a one-dimensional distributional vector from all the words that co-occur with a target word, we separate out the (syntactic) contexts and create a two-dimensional matrix; for "dog", for example, there is one dimension for actions (verbs) it participates in as the subject (such as "run" and "bark"), another for the actions for which it is the direct object (such as "call" and "feed"), another for the physical attributes (adjectives such as "happy" and "devoted"), and so on. Adherents of this type of Structured Distributional Semantics model (SDSM) argue that the differentiation by syntactic context helps distinguish "man bites dog" from "dog bites man", and in general produces more specific association lists (though at the cost of increased sparsity of the vectors).

Today, still, there is no satisfactory solution to compositionality; there is no way to measure how well a vector or matrix represents a unit larger than a word or short phrase, there is no clearly articulated model of the component dimensions of the matrix. In other words, there is a lot of work to be done. IR offers an ideal experimental environment to investigate the semantics of larger units such as queries and documents.

If there is any community who should know about building word association vectors with very large corpora, and using various methods of comparing them, clustering them, and otherwise manipulating them, it is the IR community. Interestingly, the IR community has long resisted any use of the word 'semantics' — probably a residue of the emotional response of Salton long ago— and any research that too overtly smacks of semantics. This fastidiousness has held it back. It is good to see the title of this workshop, and the willingness of a new generation of IR researchers to openly discuss the Forbidden Topics.

In my talk I will briefly outline the SDSM model that we have built at CMU and some uses we have put it to, and describe our successes and failures in attacking compositionality. I will then argue that compositionality is precisely what the IR community is lacking when it cannot find the "Trenton" and "capital of New Jersey" are the same thing. I will suggest some ways in which the IR community can go about rectifying this problem, albeit at the cost of actually working with (a weak) form of that dreaded thing, semantics.