Distribution and Inference

Jerry R. Hobbs
USC Information Sciences Institute
Marina del Rey, CA
hobbs@isi.edu

Jonathan Gordon
USC Information Sciences Institute
Marina del Rey, CA
jgordon@isi.edu

We believe that the information encoded in distributional vectors is a lossy projection of an underlying inferential structure. This raises two questions: What's projected and what's lost?

Let's start with the inferential picture (abduction) and derive the distributional picture, and then we can see to what extent we can drive it backwards. In the Interpretation as Abduction framework (Hobbs et al., 1993), the inferential structure is a proof graph for the logical form of a text or utterance. If we understand how this is projected onto distributional vectors, we may be able to discover how to use the latter to recover the former.

The key lies in an important and pervasive property of the inferential structure underlying discourse: its high degree of implicit redundancy.

In the original Quillian paper on semantic nets (Quillian, 1968), he did marker passing from different bits of the content of the text, and when he found an intersection, he knew he had a good reading. In Interpretation as Abduction, back-chaining corresponds to marker passing, and unifying or "factoring" literals corresponds to finding intersections. Unifying literals in an explanation gives you more bang for the buck – a better explanation for fewer assumptions – and it is the primary mechanism for coreference resolution.

To pump intuitions, take an example that was in Hobbs (1978) and was reanalyzed for Hobbs et al. (1993):

> *The plain was reduced by erosion to its present level.*

The sentence contains five content words, whose meanings are encoded in logical axioms we can verbalize as:

1 *reduce*, v: Decrease (change) on a vertical scale.
2 *plain*, n: A flat landform.
3 *erosion*, n: Decrease (change) of a landform on the altitude scale (which is vertical).
4 *present*, a: There's been a change from something else.
5 *level*, n: Position of a flat thing on a vertical scale.

We get the best interpretation, resolving all the coreference issues, if we unify the decreases, the verticals, the flats, the landforms, and the changes. The best interpretation has a huge amount of redundant implicit content.

Our hypothesis is that if two words have high pointwise mutual information distributionally, it is because they have a high amount of redundant implicit content inferentially. Hiding under every high PMI in a word's distributional vector is a predication that the words have in common in their associated axioms.[1]

If distribution is a lossy projection of inference, let's see what is lost. Let's say the logical axioms for *erosion*, *altitude*, and *level* are:

$$(\forall x)[\text{erode}(x) \Leftrightarrow (\exists y, z, s)[\text{altitude}(s) \,\&\, \text{landform}(x) \,\&\, \text{decrease}(x, y, z, s)]]$$

$$(\forall s)[\text{altitude}(s) \Leftrightarrow \text{vertical}(s)]$$

$$(\forall z, x, s)[\text{level}(z, x, s) \Leftrightarrow \text{vertical}(s) \,\&\, \text{flat}(x) \,\&\, \text{at}(x, z, s)]$$

When we note only that *erosion* and *level* overlap in the concept of *vertical*, we are essentially collapsing these axioms to

vertical – erode
vertical – level

[1] More precisely, the shared predication is somewhere in the shared inferential structure for the words. E.g., the axiom for *erosion* uses the predicate *altitude*, whose axiom uses the predicate *vertical*, and it is this predication that is shared with *level* and *reduce*.

That is, we are losing the argument structure, intermediate steps in the inference chain, and all but one proposition in the content of the axioms.

Now suppose we throw away the label *vertical* for the predicate and represent it instead as ⟨*erode, level*⟩, i.e., "that concept by virtue of which *erode* and *level* overlap semantically". Since *vertical* is also the common content of *reduce* and *level*, if we represent this as ⟨*reduce, level*⟩, we are losing more than just a familiar name; we are losing the identity of ⟨*reduce, level*⟩ and ⟨*erode, level*⟩.

So a high PMI in a word's distribution vector tells us that both words somehow involve a common concept, but it does not tell us where else in our knowledge base that concept occurs, how the arguments line up, or what other content there is in the axioms that involve the common concept. That is, a high PMI between words *a* and *b* tells us there is a predicate *p* such that

$$\ldots \,\&\, p(\ldots) \,\&\, \ldots \Rightarrow a(\ldots)$$

and

$$\ldots \,\&\, p(\ldots) \,\&\, \ldots \Rightarrow b(\ldots)$$

are axioms. But it doesn't tell us what *p* is or what's in those dots.

So a word's distributional vector is a coarse-grained approximation of its inferential possibilities. If we knew the right way to compose word vectors into vectors for phrases and sentences, it should tell us how the composition of words and phrases focuses the inferential possibilities of those words and phrases, and the vector for a sentence should be an indication of the bag of predicates in its best interpretation. Distributional semantics may give you cheap inference, and a basis for probabilistic inference.

Now let's try to reverse this picture. Language acquisition researchers argue that distributional information is important in acquisition. This is undoubtedly true, but it can only be a small part of the picture. If all you knew about words were their distribution vectors, the only utterances you could produce would be bags of related words. Learning the meaning of words in the Interpretation as Abduction framework is being able to construct the axioms. So being able to reverse the inference-to-distribution projection is crucial to language acquisition in children and adults and to automatically building knowledge bases for NLP.

Several researchers have used the heuristic that distributionally similar words will have similar axioms.[2] This is a reasonable attempt to jump in one step from distribution to inference. If you've encoded lots of knowledge about physicians, maybe lots of it applies as well to dentists.

Based on the above analysis, we identify three requirements for a more general program of reversing the inference-to-distribution projection:

1 We need a way of determining the missing content in the axioms. It may be that by looking at co-occurrence information with multiple words, we can derive a "bag of predicates" (unlabelled) that the axiom somehow combines.
2 We need a way of determining the predicate–argument structure and the logical structure within the axiom. One may be able to exploit syntactic structure (logical form) of texts to solve this problem. Another optimistic thought: Often from a bag of words we can reconstruct the sentence they come from, and hence its logical form; one may similarly be able to reconstruct axioms from bags of predicates.
3 We need a way of identifying underlying predicates that we introduce. The specific labels (e.g., *vertical*) don't matter, but the identity of two introduced predicates does matter. Examining multiple pairs may help with this.

There has been some success in discovering hypernymy relations from distribution, i.e., axioms of the form $p(x) \Rightarrow q(x)$ (Lenci and Benotto, 2012; Roller et al., 2014). The reason this has been successful is that problems 1 and 2 go away. It's possible this analysis can lead to the identification of other axiom patterns that can be induced from distribution.

We are currently examining the extent to which different models of distributional semantics capture the sort of relations needed in order to explicitly encode lexical and world knowledge. In addition to the approach suggested above, we are exploring methods to derive this knowledge based on analogy (i.e., relational similarity) in vector spaces.

Acknowledgments

This work is supported by Contract W911NF-15-1-0543 with the US Defense Advanced Research Projects Agency (DARPA).

[2] E.g., Beltagy et al. (2013) use distributional similarity to generate weighted inference rules on-the-fly, which allow the use of existing knowledge for similar entities.

References

Islam Beltagy, Cuong Chau, Gemma Boleda, Dan Garrette, Katrin Erk, and Raymond Mooney. 2013. Montague meets Markov: Deep semantics with probabilistic logical form. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics* (**SEM*), pages 11–21, Atlanta, GA, June.

Jerry R. Hobbs, Mark Stickel, Douglas Appelt, and Paul Martin. 1993. Interpretation as abduction. *Artificial Intelligence*, 63(1–2):69–142.

Jerry R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44:311–38.

Alessandro Lenci and Guilia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics* (**SEM*), pages 75–9.

M. Ross Quillian. 1968. Semantic memory. In Marvin Minsky, editor, *Semantic Information Processing*, pages 227–70. MIT Press, Cambridge, Massachusetts.

Stephen Roller, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of the 25th International Conference on Computational Linguistics* (*COLING*), pages 1025–36.