Data Semantics
Final project

The Distributional Hypothesis: semantic models in theory and practice

Stefano Ottolenghi

January 2018

1 What is the Distributional Hypothesis

When it comes to Distributional Semantics and the Distributional Hypothesis, the slogan is often “You shall know a word by the company it keeps” (J.R. Firth [7]).

The idea of the Distributional Hypothesis is that the distribution of words in a text holds a relationship with their corresponding meanings. More specifically, the more semantically similar two words are, the more they will tend to show up in similar contexts and with similar distributions. Stating the idea the other way round may be helpful: given two morphemes with different semantical meaning, their distribution is likely to be different.

For example, *fire* and *dog* are two words unrelated in their meaning, and in fact they are not often used in the same sentence. On the other hand, the words *dog* and *cat* are sometimes seen together, so they may share some aspect of meaning.

Mimicking the way children learn, Distributional Semantics relies on huge text corpora, the parsing of which would allow to gather enough information about words distribution to make some inference. These corpora are treated with statistical analysis techniques and linear algebra methods to extract information. This is similar to the way humans learn to use words: by seeing how they are used (i.e. coming across several examples in which a specific word is used).

The fundamental difference between human learning and the learning a distributional semantic algorithm could achieve is mostly related to the fact that humans have a concrete, practical experience to rely on. This allows them not only to learn the *usage* of a word, but to eventually *understand* its meaning. However, the way word meaning is inferred is still an open research problem in psychology and cognitive science.

1.1 The weak and strong hypothesis

Before going any further, it is important to keep in mind that there are two versions of the distributional hypothesis, as pointed out by A. Lenci [1].

In the weak version of the hypothesis, the distribution of words is believed to be a way of inferring their semantic behaviour. Let us stress the term *behaviour* in the previous sentence. The weak hypothesis does not state at any rate that the distribution of words will, in any way, tell something about the *meaning* of words. In fact, we may only hope (and this would already be a success!) to unveil the ways a word is *used*.

The idea here is that the semantic meaning of words has a specific effect on their distribution. By observing the distribution, we may then hope to infer something about the meaning. The distribution is thus seen as a *latent variable* of the semantic meaning: observing the former may give hints about the latter.

The strong form of the hypothesis, on the other hand, states that words distribution have a *causal relationship* with the way meaning is derived from the text. In fact, this theory entails that the statistical distribution of a word causes the semantic representation humans have of the corresponding idea.

In other words, while the weak version of the distributional hypothesis states that there is a relationship such that

$$\text{meaning} \ \rightarrow \ \text{distribution}$$

but does not say anything about the inverse, the strong version believes that to hold as well.

Perhaps not surprisingly, the weak version is accepted by far more researchers of the (computational) linguistics fields than the strong one. It is not difficult to believe that meaning has a strong influence on the distribution of words, but it is not so easy to believe that distributions shape meaning by themselves.

For example, consider the following sentence:

> *"I hiked 2000 metres uphill in the mountains to have a swim in the lake".*

Knowing the meaning of the word *sea*, humans (hopefully) know that the word *lake* cannot just be replaced with the word *sea* in the sentence above, even if there is the word *swim*, which is often seen together with *sea*. However, if we believed that distribution shapes meaning, then words with supposedly very similar distributions such as *lake* and *sea* could be swapped one with the other. This is not the case, although this is only clear if the meaning of the context is understood.

As we will see, the relationship between distribution and meaning with respect to the distributional hypothesis is a difficult one, and we will discuss it in more depth in Section 3.

2 How to deploy the idea

In order to investigate lexical meaning, we first need to set up a way to measure semantic similarity between words. When are two words similar? What does it even mean for two words to be *similar*? These are important questions, as word similarity is believed to affect the way terms are processed in mental lexicon. We may define synonymy as being related to word interchangeability, but this is not a computationally implementable definition yet.

Thus, a way to make explicit the functional dependance between word distribution and semantic is needed.

Distributional semantics models rely on statistical analysis and linear algebra tools to implement its theory. We will now discuss in more depth Latent Semantic Analysis and hint at further developments.

2.1 Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is arguably **the** mathematical tool of distributional semantics. In its basic form, it allows to parse several texts and analyze similarities between them. It is an unsupervised learning method (i.e. it requires no human-preprocessed data). Although it is not the only possible method, others are roughly based on the same mathematics that LSA employs, so exploring the latter is enough to understand other approaches.

The idea of LSA is to regard words as points in a so-called *distributional space*. The distributional space is simply a vector space (most often $\mathbb{R}^n$), where vectors represent words. The vector contains an average of all possible different words usages deduced from the input text corpora.

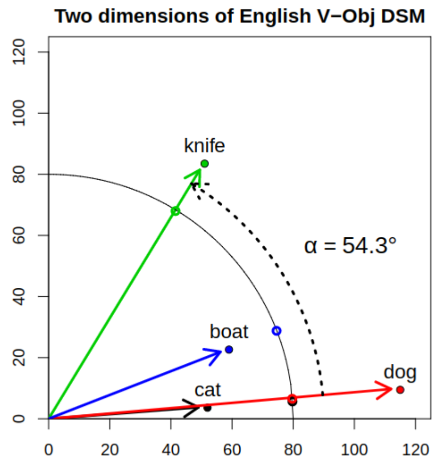

Figure 1: Dot products between words (Evert [2]).

However simple this construction may seem, it has a very nice property: that distances between points correspond to semantic distances between words. In particular, measuring angles between vectors (which is as simple as calculating dot products) is a proxy for measuring the similarity of the corresponding words (see Figure 1).

An example of the result of text corpora processing through Latent Semantic Analysis is shown in Figure 2.

As an aside, notice that the name of this technique comes from the idea that semantic may be thought of as a latent variable responsible for the observed linguistic distributions.

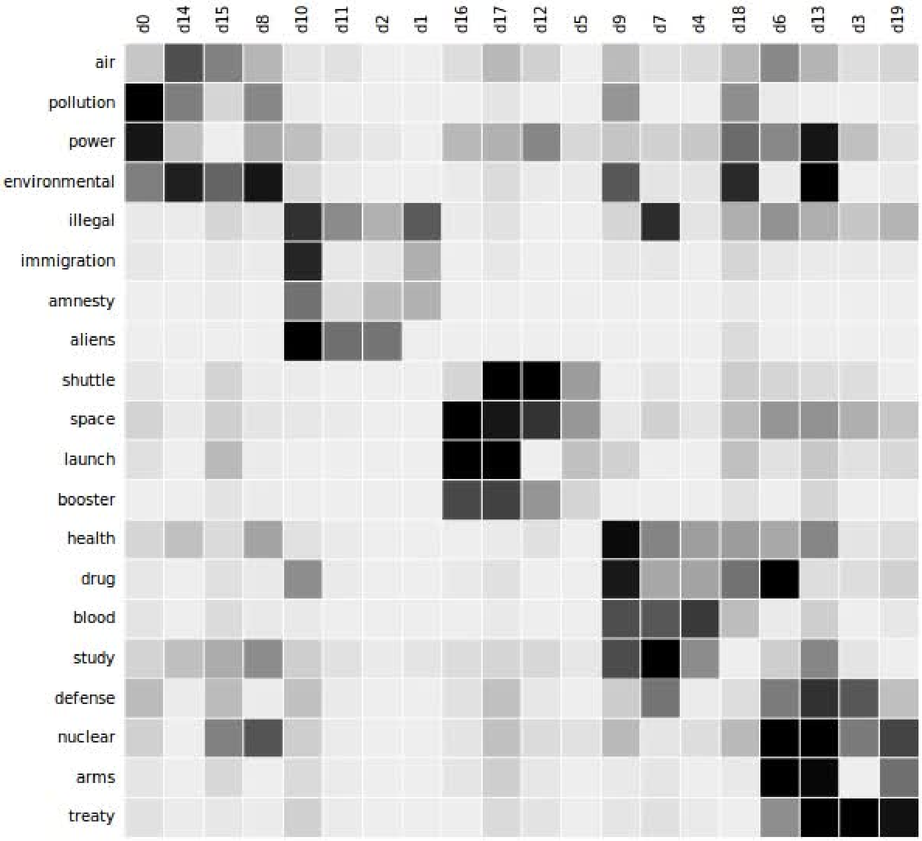

Figure 2: Latent Semantic Analysis-based grouping of documents basing on their keywords (Wikipedia [3]).

2.1.1 How it is done

We briefly go through how Latent Semantic Analysis works from a technical point of view.

Consider n documents, which contain m different words all together. LSA starts by building a matrix of size $m \times n$:

$$\begin{array}{cc}
 & \mathbf{d}_j \\
 & \downarrow \\
\mathbf{t}_i^T \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,j} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & \dots & x_{i,j} & \dots & x_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,j} & \dots & x_{m,n} \end{bmatrix}
\end{array}$$

Each row t_i^T represents a word, while each column d_j represents a document. The matrix entry $x_{i,j}$ holds the occurrence count (or frequency) of the word i in the document j.

It is then possible to calculate similarity both between terms and between documents. In fact, the dot product $t_i^T t_p$ computes the similarity between the word vectors i and p, whereas $d_j d_k^T$ measures similarity between documents j and k.

It is easy to generate matrices of co-occurrence both for words and for documents, through the computation respectively of $X^T X$ and of XX^T. In particular, $X^T X$ will contain information about the similarity of words, basing on the provided corpora; XX^T will contain similar information about the documents.

However, the matrix generated by LSA is likely to be huge in size, and thus computationally difficult to handle. Furthermore, a good part of it may well consist of useless data: noise, synonyms and other information that may be stripped away without any concrete loss.

For these reasons, Singular Value Decomposition is applied to the matrix to extract the most meaningful pieces of information from it. In this way, similarity measures will still be computable, though with less computation effort.

However, the drawback is that the new dimensions are very unlikely to relate to any comprehensible concept: they will just be linear combinations of rows and columns.

To make this concern clearer, consider the following example (from [3]).

(car), (bottle), (flower) → (1.3452 * car + 0.2828 * bottle), (flower)

Here, two words are merged in one new joint dimension, while the third is retained as it is. However, although the interpretation of car and bottle is clear, it is not clear at all how the joint dimension should be thought of.

2.1.2 Critiques and more sophisticated approaches

As we have seen, LSA is able to generate co-occurrence patterns from the input it is provided. However, the idea that meaning may consist of abstractions from purely linguistic co-occurrence patterns has raised many critiques. As already pointed out, the biggest one is that co-occurrence information may only allow to infer the usage of a word, but not its meaning.

Moreover, one concern is understanding whether the current limits and weak points of the distributional semantics approaches are due to the implementations and current models. On the other hand, it may be possible that they are inherent to the distributional hypothesis itself.

When using Latent Semantic Analysis, further rules may be enforced. In fact, although text corpora may be used as they are in their raw form (*naive* approaches), syntactically-savy models are possible as well. In these latter models, the distribution is analysed by keeping in mind the syntactic configurations words should obey to. In this way it should be possible to get closer to the ways terms are used in reality.

It is worth pointing out that LSA-based methods are often the starting point of distributional semantic investigations, but more complex methods have been devised on top of it.

One example of further technique consists on focusing on the notion of property of a concept, to try to capture the essence of its meaning. In this model, *animal* would be identified as property for *dog*, and *wheels* as property for *car*.

Another possible approach is that of semantic association. In this fashion, *swim* would find a relationship with *water*, as the former calls to mind the latter.

3 Critiques to the distributional approach

As we have seen, the distributional hypothesis in its weak form is usually accepted by field experts, but there is strong disagreement over the strong formulation.

One fundamental critique to the distributional hypothesis is that, as far as it could go, distributional semantic models may not go beyond learning word usage. In fact, learning words meaning is totally out of their scope. This is well conveyed by the *Chinese Room* thought experiment, first proposed by Searle in 1980 [5]:

> *Imagine to be locked in a room and to receive batches of Chinese characters. Crucially, you do not know Chinese and for you these characters are just meaningless symbols. You also receive "rules" (in English) on how to combine these characters, and how to respond to Chinese characters with other Chinese characters. Suppose that, after a certain amount of training, you have been able to learn to combine Chinese characters and to reply to Chinese messages in such a way that your answers are de facto indistinguishable from those given by Chinese native speakers. The key point is that even in this case it would still be true that you do not understand Chinese, even though you may be said to "speak it".*

Another point of difficulty is that of composivity. By learning the meaning of the parts (i.e. the single words that make up a sentence), is it possible to learn the meaning of the whole sentence?

As we have seen, technical implementations of the theory (Latent Semantic Analysis) usually rely on basic linear algebra and vector spaces. In these models, each word is conceived as a point. However, the sum of the points representing the words of a sentence is an absolutely inadequate definition for the meaning of the sentence. This is because vector summation is commutative, whereas word summation should not.

In fact, consider the following examples:

1. *The dog ate a cake.*
2. *The cake ate a dog.*
3. *The a cake dog ate.*

If word-vector summation would be enough to define sentence meaning, then the three sentences would be believed to all have the same meaning, although this is clearly not the case. Remarkably, the third sentence does not even have any real meaning at all, although it is believed to be equal to the other two. The point here is that while swapping vectors in a sum is fine, swapping words in a sentence obviously is not.

More sophisticated models that account for non-commutativity in words usage have been developed, but this still looks like an open problem.

One further question, although weakly related to that of non-commutativity, is that of asymmetric relationships. For example, sentences like:

A bus has wheels.

allow to learn that there is a strict relationship between *buses* and *wheels*, but it is not clear how to learn the fact that it is an asymmetric link. Indeed, it is difficult to find a bus without wheels, but it is not difficult at all to find wheels without a bus.

Moreover, experts argue [1] that meaning is most often influenced by factors that cannot be learnt just by the textual representation of content. In fact, the speaker's intentions, the speaker's own way of expression and the context in which the content fits into all shape the meaning of the words it contains. The statistical methods employed by distributional semantics are not able to capture these factors, as they lay beyond its scope.

Also, keep in mind that up to now we have only considered well-formed language instantions, i.e. sentences in which the meaning exactly corresponds to the words used to express it. However, consider the following sentence:

I borrow books and sometimes they do not come back to me.

A human skilled enough with the English language easily understands that the speaker mistaked the word *borrow* with the word *lend*, but would a computer be able to do this?

Ultimately, making inferences given a group of facts is the final goal of learning the meaning of words. However, it is unclear how this could be achieved without a real understanding of semantic meaning. In fact, one of the main critiques to distributional semantics is precisely this: that it cannot provide meaning for words.

Finally, it should be noted that the distributional semantics debate is just an instance of a broader debate. The two constrasting theories in this regard are the Abstract Cognition Hypothesis (ACH) and the Embodied Cognition Hypothesis (ECH).

Standing to Abstract Cognition Hypothesis, concepts and meanings are represented in human cognition by formal symbols. This stance supports the Distributional Hypothesis. On the other hand, the Embodied Cognition Hypothesis believes that meanings are stored in the same perceptual system from which its instance is experienced. This means that knowing the meaning of the word *spider* implies being able to run some sort of internal simulator that re-enacts out concrete experience with spiders. This theory, of course, is in sharp contrast with the distributional hypothesis and its tools, since requires a deeper understanding of words, that goes beyond their usage.

4 Final remarks

As we have discussed at length, even if distributional-based semantic models could tell us something interesting about the language and words usage, it would reveal absolutely nothing about meanings. In fact, one could build graphs, buckets and at any rate classify words in fashionable and complex ways, but it seems pretty unlikely that the meaning of a word could be extracted (just) from those kind of analysis. This is **the** vital problem of distributional semantics.

However, even though philosophical debatings are far from settled, distributional semantics models have allowed good results in real-world applications, and so are used notwithstanding the critiques. Indeed, using vocabulary tests to compare the performance of Latent Semantic Analysis based models with the learning achieved through reading by school children, it was found that they scored similarly [4].

After all, it is important to realize what our goals are. If we would like a machine to understand ideas, concepts and words as a human would do, then there is probably no technique that would allow this in the current state of things. However, if we only require a machine to process and respond to linguistic queries, then distributional models can be used with a good success rate.

To make an analogy with another field, let us notice that we do not know how (artificial) neural networks work, although we use them and rely on them in more and more contexts. Crucially, we can not state that they are able to *think* or *understand* the work they do. For example, when searching for *tieleman orchestra*, Google suggests *thielemann orchestra* instead. Do Google's algorithms know who Thielemann is, or that he is a director? Do they know what a director even is?

All of this is to say: then, why should distributional models not be used, if they have proved to be good enough to infer words usage, which is what matters in practice? Our lives lay on top of automated systems that work, although they do not understand their work. Distributional semantic models may well become one of these tools.

The difficulty of devising computational models that would extract meanings from terms and contexts is mostly related to the fact that we ourselves do not yet understand how we do it. It is unlikely, of course, that we will ever be able to teach a machine to do it without understanding it ourselves first.

At any rate, is seems unlikely (at least as of now) to ever be able to teach stupid machinery to learn meanings and concepts.

The fact that we are able to *understand* the word *circle*, for example, has undoubtedly to do with our past experience and interaction with the real word: we have seen round things, balls and other different geometric shapes, and have an *intuitive idea* of what a circle is. This allows us not only to *use* the word, but to have a real understanding of, which stands at a higher level. In fact, one may argue that the understanding that an 8-year-old could have of *circle* would be at a superior level than any distributional model!

In other words, it is just difficult to believe that a simple algorithm/technique may be able to compete with the human conceptualization of the world. After all, just observing the statistical occurrence of the word *idea* may make it look like it has a lot to do with *lamps* and *lights*.

As Wittgenstein said, "the meaning of a word lies in its use"[6], and looking at usage may be enough to replicate usage patterns. However, grasping and understanding that meaning looks like a totally different matter!

References

[1] Lenci Alessandro 2008. Distributional semantics in linguistic and cognitive research. Rivista di Linguistica 20.1, pp. 1-31

[2] Evert Stefan - Distributional Semantic Models (last accessed: 2018/02/26) http://esslli2016.unibz.it/wp-content/uploads/2015/10/dsm_tutorial_part1.slides.pdf

[3] Wikipedia - Latent Semantic Analysis (last accessed: 2018/02/26) https://en.wikipedia.org/wiki/Latent_semantic_analysis

[4] Landauer Thomas K., Susan T. Dumais 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review CIV/2. 211-240

[5] Searle John 1980. Minds, brains and programs. Behavioural and Brain Sciences III. 417-424

[6] Wittgenstein Ludwig 1953. Philosophical Investigations. Oxford: Blackwell

[7] Firth John R. 1957. Papers in Linguistics. London, Oxford University Press.