UNIVERSITY OF TRENTO - Italy

Faculty of Cognitive Science

Master's Degree in Cognitive Science

Using Twitter To Build And Update A Time-Sensitive Semantic Resource

Tutor
Marco Baroni

Student
Marco Milano

Academic Year 2010/2011

UNIVERSITY OF TRENTO - Italy
Faculty of Cognitive Science

Summary

In linguistics, the distributional hypothesis of semantics states that we *shall know the meaning of a word by the company it keeps* (Firth, 1957).

The distributional properties of words as they appear in a selected text within a fixed context window are thought to be a reliable indicator of their meaning. Under this hypothesis it is also applicable from the opposite, or that the topic of a text can be inferred by the most frequent words that appear in it.

When Firth expressed his intuition more than fifty years ago, it was practically impossible to test it. Today, thanks to the digital nature of information, as well as to the improvements in computational power and memory space, things have changed.

Within the broader computational linguistics framework, which seeks to investigate language by means of computational models, computational semantics attempts to represent the meaning of words in a computable manner.

The most reliable and efficient application of the distributional hypothesis of semantics, so far, is the Word-Space Model (WSM), also known as the Vector-Space Model (VSM), as the former is usually referred to in its original form (Salton, 1971).

The VSM is a mathematical tool that represents words as vectors in a multidimensional geometrical space, where each vector's elements correspond to the frequency statistics of a target word within a document or a context – i..e., its neighbour words.

The present study will focus on that particular VSM in which the number of dimensions of the resulting space reflects the number of context words of a set of target words.

Representing words as points on a Cartesian plane allows us to calculate the distance between such points and to consider it as a good estimator of their semantic similarity in a supposed corresponding mental space: the closer the

points on the plane, the more similar the corresponding words with respect to their meaning (Widdows, 2004).

For what regards the actual processing of semantic relations, words co-occurrence statistics is extracted from a collection of texts and is used to build a computational representation of their meaning; the final output is a frequency matrix.

Once words are arranged as vectors, computing the semantic similarity between two given words is just a matter of choice, the cosine distance measure usually being the preferred option, as it happens to carry a very nice processing feature: it is, in fact, the dot product of two normalised vectors.

The cosine measure is also what we have adopted in this present work, in order to account for word pairs meaning similarity in the semantic space we have constructed.

The cosine distance measure is easily implementable on a computer and whilst machines cannot understand what that means, the output that is produced makes so much sense to humans that it is considered to be a plausible representation of words meaning (McDonald and Ramscar, 2001).

The VSM of semantics is currently the standard computational model for almost any research conducted in computational semantics (Turney and Pantel, 2010; Baroni and Lenci, 2010; Mitchell and Lapata, 2010; Sandin et al. 2011).

Importantly, the VSM allows the unsupervised extraction of semantic relations from large corpora, requiring only a simple pre-processing step of the entry dataset (Baroni 2010).

On the other hand, the main drawback of the VSM representation of words as we intend it, is its computational cost: by considering each co-occurring word in a text as a possible context of its neighbour words, the resulting frequency matrix becomes as computationally unsustainable as the number of words grows, to the extent of becoming an impossible enterprise when the whole vocabulary is considered (Sahlgren, 2006).

To solve this critical issue, many matrices reduction techniques have been developed, namely the Latent Semantic Analysis (LSA) technique which adopts the truncated Singular Value Decomposition algorithm (often just SVD) to reduce a vector space's initial number of dimensions (Landauer and Dumais, 1997).

Notably, the LSA is also a linguistic model that captures the underlying semantic relations between words that might otherwise be hard to notice in the standard frequency matrix representation, given its sparse nature. As proved by Landauer and Dumais, reducing one frequency matrix dimensionality is very likely to lead to a more precise representation of word pairs semantic similarity.

Truncated SVD is a reliable dimensionality reduction technique but, unfortunately, it sometimes suffers from serious flaws: it is not scalable, which means that it cannot be applied to increasingly larger datasets and it's a one-step operation which has to be repeated from scratch each time one is willing to modify or update an entry matrix due to the occurrence of new data (Sahlgren, 2005).

Consequently, a new approach to the matrices reduction issue, called Random Indexing, was proposed in the early 2000s, which overcomes the SVD limited applicability and allows the processing of larger datasets without losing the representational power of its predecessor (Kanerva, 2000; Karlgren and Sahlgren, 2001).

The random indexing technique has the key advantage of reducing the number of dimensions of the final geometrical space built out of words co-occurrence statistics, by randomly projecting what would be the original frequency vectors onto a smaller, dimension-fixed vector space. It is only in a second processing step that the random space is updated with actual words frequency statistics, leading to a frequency matrix.

Our work has been inspired by the dynamic features and representational power of the random indexing approach and aims to model meaning

representation over time by using the random indexing method to build a reliable and easy-to-update frequency matrix.

The reason for the application of the random indexing technique is also due to the non-standard nature of the target corpus to which we decided to apply it.

We chose, in fact, to test it against a corpus of 140-characters-long messages (tweets) collected from an online micro-blogging platform, Twitter. A tweet is a message that Twitter users share with their followees, or the people that have access to their 'blog' feeds.

Essentially, Twitter is hub of social information which delivers all sorts of content.

As I will report, there has been a notable surge in research conducted on Twitter-generated datasets over the last three years, for many different reasons; I will try to address the more relevant ones in detail.

For a computational linguist, Twitter represents a non-stop streaming of data about real-world usage of language. Of course, the built-in limitation of maximum 140 characters per message doesn't necessarily reflect a real conversation or exchange of information, but it somewhat assures that someone is spontaneous. For such reasons, Twitter could be thought of as one of the best sources of social information about almost anything that really matters. Consequently, it has become critical for businesses as well as for academics and Institutions to be able to mine such information in order to extract what is really relevant.

This relatively new field is called text mining (or information extraction) and as I will show, it makes it possible to identify an event just by looking at the distribution of words and how this changes over a specific time span.

The ease to apply a relatively simple semantic analysis even to large datasets has encouraged us to investigate whether the creation and evaluation of

an incremental version of a standard semantic resource, could be applied to events detection in mini-blogs streams.

An incremental semantic resource is a semantic resource generated by processing words co-occurrence statistics extracted from a large dataset and subsequently updating it whenever new data becomes available.

Our objective was to update an entry semantic resource with time-limited datasets in order to be able to account for the expected shift in words semantic similarity patterns.

Twitter somewhat mirrors what happens in the real world. If something is enough relevant for Twitter users to be consistently shared, it should be possible to identify its relevance by accounting for shifts in word pairs semantic similarity patterns.

By making a semantic resource virtually incremental it could in theory be possible to identify changes in word pairs semantic content by looking at their cosine distance measure in the updated vector spaces.

As I will show, at this stage, we have simply built updated versions of the entry semantic resource, a process which results in as many semantic spaces as there are time-limited corpora available – in our experiment: five updated semantic spaces plus the entry one.

Each new, updated semantic space represents frequency information of word pairs from the previous semantic space updated with information from the new specific corpus.

As I mentioned, we decided to focus on Twitter because of the raising interest that such information sharing platform had gained between computationally linguists over the last few years but it was also the consequence of a lucky coincidence: a large sample of time-limited corpora was made available to us from Herdağdelen, as part of his work on common sense knowledge extraction from the Web under the supervision of Baroni (Herdağdelen and Baroni, 2011 – to appear).

UNIVERSITY OF TRENTO - Italy
Faculty of Cognitive Science

Whereas constructing a semantic resource is nowadays an established natural language processing task, building reliable semantic resources from the Web is still a working progress.

Creating a reliable semantic resource out of a micro-blogging service platform has proved quite challenging. The nature of a very short and colloquial format is itself a serious issue as Twitter users tend to shrink or truncate words as well as to abuse the correct usage of punctuation.

The collection and pre-processing of tweets has required a sort of normalisation of most frequent and clearly identifiable 'distortions' of standard word forms. Only English tweets were made part of the processed corpora.

This pre-filtering was done by using a rough and very basic modified version of the original language identifier SLIde developed by Fabio Celli, a PhD student at the CIMeC.

The actual Twitter dataset was downloaded and made available to us by Amaç Herdağdelen, whilst he was a PhD student at the CIMeC.

Text pre-processing software and the random indexing algorithm program to generate the entry frequency matrix were developed by Marco Baroni.

The program to update the entry matrix (a modified version of Baroni's software) along with a script to normalise English tweets and the fine-tuned modified version of SLIde, were developed by the author.