Measuring semantic content in distributional vectors

Aurélie Herbelot
EB Kognitionswissenschaft
Universität Potsdam
Golm, Germany
aurelie.herbelot@cantab.net

Mohan Ganesalingam
Trinity College
University of Cambridge
Cambridge, UK
mohan0@gmail.com

Abstract

Some words are more contentful than others: for instance, *make* is intuitively more general than *produce* and *fifteen* is more 'precise' than *a group*. In this paper, we propose to measure the 'semantic content' of lexical items, as modelled by distributional representations. We investigate the hypothesis that semantic content can be computed using the Kullback-Leibler (KL) divergence, an information-theoretic measure of the relative entropy of two distributions. In a task focusing on retrieving the correct ordering of hyponym-hypernym pairs, the KL divergence achieves close to 80% precision but does not outperform a simpler (linguistically unmotivated) frequency measure. We suggest that this result illustrates the rather 'intensional' aspect of distributions.

1 Introduction

Distributional semantics is a representation of lexical meaning that relies on a statistical analysis of the way words are used in corpora (Curran, 2003; Turney and Pantel, 2010; Erk, 2012). In this framework, the semantics of a lexical item is accounted for by modelling its co-occurrence with other words (or any larger lexical context). The representation of a target word is thus a vector in a space where each dimension corresponds to a possible context. The weights of the vector components can take various forms, ranging from simple co-occurrence frequencies to functions such as Pointwise Mutual Information (for an overview, see (Evert, 2004)).

This paper investigates the issue of computing the semantic content of distributional vectors. That is, we look at the ways we can distributionally express that *make* is a more general verb than *produce*, which is itself more general than, for instance, *weave*. Although the task is related to the identification of hyponymy relations, it aims to reflect a more encompassing phenomenon: we wish to be able to compare the semantic content of words within parts-of-speech where the standard notion of hyponymy does not apply (e.g. prepositions: see *with* vs. *next to* or *of* vs. *concerning*) and across parts-of-speech (e.g. *fifteen* vs. *group*).

The hypothesis we will put forward is that semantic content is related to notions of relative entropy found in information theory. More specifically, we hypothesise that the more specific a word is, the more the distribution of the words co-occurring with it will differ from the baseline distribution of those words in the language as a whole. (A more intuitive way to phrase this is that the more specific a word is, the more information it gives us about which other words are likely to occur near it.) The specific measure of difference that we will use is the Kullback-Leibler divergence of the distribution of words co-ocurring with the target word against the distribution of those words in the language as a whole. We evaluate our hypothesis against a subset of the WordNet hierarchy (given by (Baroni et al, 2012)), relying on the intuition that in a hyponym-hypernym pair, the hyponym should have higher semantic content than its hypernym.

The paper is structured as follows. We first define our notion of semantic content and motivate the need for measuring semantic content in distributional setups. We then describe the implementation of the distributional system we use in this paper, emphasising our choice of weighting measure. We show that, using the compo-

nents of the described weighting measure, which are both probability distributions, we can calculate the relative entropy of a distribution by inserting those probability distributions in the equation for the Kullback-Leibler (KL) divergence. We finally evaluate the KL measure against a basic notion of frequency and conclude with some error analysis.

2 Semantic content

As a first approximation, we will define semantic content as informativeness with respect to denotation. Following Searle (1969), we will take a 'successful reference' to be a speech act where the choice of words used by the speaker appropriately identifies a referent for the hearer. Glossing over questions of pragmatics, we will assume that a more informative word is more likely to lead to a successful reference than a less informative one. That is, if Kim owns a cat and a dog, the identifying expression *my cat* is a better referent than *my pet* and so *cat* can be said to have more semantic content than *pet*.

While our definition relies on reference, it also posits a correspondence between actual utterances and denotation. Given two possible identifying expressions e_1 and e_2, e_1 may be preferred in a particular context, and so, context will be an indicator of the amount of semantic content in an expression. In Section 5, we will produce an explicit hypothesis for how the amount of semantic content in a lexical item affects the contexts in which it appears.

A case where semantic content has a direct correspondence with a lexical relation is hyponymy. Here, the correspondence relies entirely on a basic notion of extension. For instance, it is clear that *hammer* is more contentful than *tool* because the extension of *hammer* is smaller than that of *tool*, and therefore more discriminating in a given identifying expression (See *Give me the hammer* versus *Give me the tool*). But we can also talk about semantic content in cases where the notion of extension does not necessarily apply. For example, it is not usual to talk of the extension of a preposition. However, in context, the use of a preposition against another one might be more discriminating in terms of reference. Compare a) *Sandy is with Kim* and b) *Sandy is next to Kim*. Given a set of possible situations involving, say, Kim and Sandy at a party, we could show that b) is more discriminating than a), because it excludes the situations where Sandy came to the party with Kim but is currently talking to Kay at the other end of the room. The fact that *next to* expresses physical proximity, as opposed to just being in the same situation, confers it more semantic content according to our definition. Further still, there may be a need for comparing the informativeness of words across parts of speech (compare *A group of/Fifteen people was/were waiting in front of the town hall*).

Although we will not discuss this in detail, there is a notion of semantic content above the word level which should naturally derive from composition rules. For instance, we would expect the composition of a given intersective adjective and a given noun to result into a phrase with a semantic content greater than that of its components (or at least equal to it).

3 Motivation

The last few years have seen a growing interest in distributional semantics as a representation of lexical meaning. Owing to their mathematical interpretation, distributions allow linguists to simulate human similarity judgements (Lund, Burgess and Atchley, 1995), and also reproduce some of the features given by test subjects when asked to write down the characteristics of a given concept (Baroni and Lenci, 2008). In a distributional semantic space, for instance, the word 'cat' may be close to 'dog' or to 'tiger', and its vector might have high values along the dimensions 'meow', 'mouse' and 'pet'. Distributional semantics has had great successes in recent years, and for many computational linguists, it is an essential tool for modelling phenomena affected by lexical meaning.

If distributional semantics is to be seen as a general-purpose representation, however, we should evaluate it across all properties which we deem relevant to a model of the lexicon. We consider semantic content to be one such property. It underlies the notion of hyponymy and naturally models our intuitions about the 'precision' (as opposed to 'vagueness') of words.

Further, semantic content may be crucial in solving some fundamental problems of distributional semantics. As pointed out by McNally (2013), there is no easy way to define the notion of a function word and this has consequences for theories where function words are *not* assigned a distributional representation. McNally suggests that the most appropriate way to separate function

from content words might, in the end, involve taking into account how much 'descriptive' content they have.

4 An implementation of a distributional system

The distributional system we implemented for this paper is close to the system of Mitchell and Lapata (2010) (subsequently M&L). As background data, we use the British National Corpus (BNC) in lemmatised format. Each lemma is followed by a part of speech according to the CLAWS tagset format (Leech, Garside, and Bryant, 1994). For our experiments, we only keep the first letter of each part-of-speech tag, thus obtaining broad categories such as N or V. Furthermore, we only retain words in the following categories: nouns, verbs, adjectives and adverbs (punctuation is ignored). Each article in the corpus is converted into a 11-word window format, that is, we are assuming that context in our system is defined by the five words preceding and the five words following the target.

To calculate co-occurrences, we use the following equations:

$$freq_{c_i} = \sum_t freq_{c_i,t} \quad (1)$$

$$freq_t = \sum_{c_i} freq_{c_i,t} \quad (2)$$

$$freq_{total} = \sum_{c_i,t} freq_{c_i,t} \quad (3)$$

The quantities in these equations represent the following:

$freq_{c_i,t}$	frequency of the context word c_i with the target word t
$freq_{total}$	total count of word tokens
$freq_t$	frequency of the target word t
$freq_{c_i}$	frequency of the context word c_i

As in M&L, we use the 2000 most frequent words in our corpus as the semantic space dimensions. M&L calculate the weight of each context term in the distribution as follows:

$$v_i(t) = \frac{p(c_i|t)}{p(c_i)} = \frac{freq_{c_i,t} \times freq_{total}}{freq_t \times freq_{c_i}} \quad (4)$$

We will not directly use the measure $v_i(t)$ as it is not a probability distribution and so is not suitable for entropic analysis; instead our analysis will be phrased in terms of the probability distributions $p(c_i|t)$ and $p(c_i)$ (the numerator and denominator in $v_i(t)$).

5 Semantic content as entropy: two measures

Resnik (1995) uses the notion of information content to improve on the standard edge counting methods proposed to measure similarity in taxonomies such as WordNet. He proposes that the information content of a term t is given by the self-information measure $-\log p(t)$. The idea behind this measure is that, as the frequency of the term increases, its informativeness decreases. Although a good first approximation, the measure cannot be said to truly reflect our concept of semantic content. For instance, in the British National Corpus, *time* and *see* are more frequent than *thing* or *may* and *man* is more frequent than *part*. However, it seems intuitively right to say that *time, see* and *man* are more 'precise' concepts than *thing, may* and *part* respectively. Or said otherwise, there is no indication that more general concepts occur in speech more than less general ones. We will therefore consider self-information as a baseline.

As we expect more specific words to be more informative about which words co-occur with them, it is natural to try to measure the specificity of a word by using notions from information theory to analyse the probability distribution $p(c_i|t)$ associated with the word. The standard notion of entropy is not appropriate for this purpose, because it does not take account of the fact that the words serving as semantic space dimensions may have different frequencies in language as a whole, i.e. of the fact that $p(c_i)$ does not have a uniform distribution. Instead we need to measure the degree to which $p(c_i|t)$ differs from the context word distribution $p(c_i)$. An appropriate measure for this is the Kullback-Leibler (KL) divergence or relative entropy:

$$D_{\mathrm{KL}}(P\|Q) = \sum_i \ln(\frac{P(i)}{Q(i)})P(i) \quad (5)$$

By taking $P(i)$ to be $p(c_i|t)$ and $Q(i)$ to be $p(c_i)$ (as given by Equation 4), we calculate the relative entropy of $p(c_i|t)$ and $p(c_i)$. The measure is clearly informative: it reflects the way that t modifies the expectation of seeing c_i in the corpus. We hypothesise that when compared to the distribution $p(c_i)$, more informative words will have a

more 'distorted' distribution $p(c_i|t)$ and that the KL divergence will reflect this.[1]

6 Evaluation

In Section 2, we defined semantic content as a notion encompassing various referential properties, including a basic concept of extension in cases where it is applicable. However, we do not know of a dataset providing human judgements over the general informativeness of lexical items. So in order to evaluate our proposed measure, we investigate its ability to retrieve the right ordering of hyponym pairs, which can be considered a subset of the issue at hand.

Our assumption is that if X is a hypernym of Y, then the information content in X will be lower than in Y (because it has a more 'general' meaning). So, given a pair of words $\{w_1, w_2\}$ in a known hyponymy relation, we should be able to tell which of w_1 or w_2 is the hypernym by computing the respective KL divergences.

We use the hypernym data provided by (Baroni et al, 2012) as testbed for our experiment.[2] This set of hyponym-hypernym pairs contains 1385 instances retrieved from the WordNet hierarchy. Before running our system on the data, we make slight modifications to it. First, as our distributions are created over the British National Corpus, some spellings must be converted to British English: for instance, *color* is replaced by *colour*. Second, five of the nouns included in the test set are not in the BNC. Those nouns are *brethren, intranet, iPod, webcam* and *IX*. We remove the pairs containing those words from the data. Third, numbers such as *eleven* or *sixty* are present in the Baroni et al set as nouns, but not in the BNC. Pairs containing seven such numbers are therefore also removed from the data. Finally, we encounter tagging issues with three words, which we match to their BNC equivalents: *acoustics* and *annals* are matched to *acoustic* and *annal*, and *trouser* to *trousers*. These modifications result in a test set of 1279 remaining pairs.

We then calculate both the self-information measure and the KL divergence of all terms included in our test set. In order to evaluate the system, we record whether the calculated entropies match the order of each hypernym-hyponym pair. That is, we count a pair as correctly represented by our system if w_1 is a hypernym of w_2 and $KL(w_1) < KL(w_2)$ (or, in the case of the baseline, $SI(w_1) < SI(w_2)$ where SI is self-information).

Self-information obtains 80.8% precision on the task, with the KL divergence lagging a little behind with 79.4% precision (the difference is not significant). In other terms, both measures perform comparably. We analyse potential reasons for this disappointing result in the next section.

7 Error analysis

It is worth reminding ourselves of the assumption we made with regard to semantic content. Our hypothesis was that with a 'more general' target word t, the $p(c_i|t)$ distribution would be fairly similar to $p(c_i)$.

Manually checking some of the pairs which were wrongly classified by the KL divergence reveals that our hypothesis might not hold. For example, the pair *beer – beverage* is classified incorrectly. When looking at the *beverage* distribution, it is clear that it does not conform to our expectations: it shows high $v_i(t)$ weights along the *food, wine, coffee* and *tea* dimensions, for instance, i.e. there is a large difference between $p(c_{food})$ and $p(c_{food}|t)$, etc. Although *beverage* is an umbrella word for many various types of drinks, speakers of English use it in very particular contexts. So, distributionally, it is *not* a 'general word'. Similar observations can be made for, e.g. *liquid* (strongly associated with *gas*, presumably via coordination), *anniversary* (linked to the verb *mark* or the noun *silver*), or again *projectile* (co-occurring with *weapon*, *motion* and *speed*).

The general point is that, as pointed out elsewhere in the literature (Erk, 2013), distributions are a good representation of (some aspects of) intension, but they are less apt to model extension.[3] So a term with a large extension like *beverage* may have a more restricted (distributional) intension than a word with a smaller extension, such as

[1] Note that KL divergence is not symmetric: $D_{\mathrm{KL}}(p(c_i|t)\|p(c_i)))$ is not necessarily equal to $D_{\mathrm{KL}}(p(c_i)\|p(c_i|t))$. The latter is inferior as a few very small values of $p(c_i|t)$ can have an inappropriately large effect on it.

[2] The data is available at `http://clic.cimec.unitn.it/Files/PublicData/eacl2012-data.zip`.

[3] We qualify 'intension' here, because in the sense of a mapping from possible worlds to extensions, intension cannot be said to be provided by distributions: the distribution of *beverage*, it seems, does not allow us to successfully pick out all beverages in the real world.

beer.[4]

Contributing to this issue, fixed phrases, named entities and generally strong collocations skew our distributions. So for instance, in the *jewelry* distribution, the most highly weighted context is *mental* (with $v_i(t)$ = 395.3) because of the music album *Mental Jewelry*. While named entities could easily be eliminated from the system's results by preprocessing the corpus with a named entity recogniser, the issue is not so simple when it comes to fixed phrases of a more compositional nature (e.g. *army ant*): excluding them might be detrimental for the representation (it is, after all, part of the meaning of *ant* that it can be used metaphorically to refer to people) and identifying such phrases is a non-trivial problem in itself.

Some of the errors we observe may also be related to word senses. For instance, the word *medium*, to be found in the pair *magazine – medium*, can be synonymous with *middle*, *clairvoyant* or again *mode of communication*. In the sense of *clairvoyant*, it is clearly more specific than in the sense intended in the test pair. As distributions do not distinguish between senses, this will have an effect on our results.

8 Conclusion

In this paper, we attempted to define a measure of distributional semantic content in order to model the fact that some words have a more general meaning than others. We compared the Kullback-Leibler divergence to a simple self-information measure. Our experiments, which involved retrieving the correct ordering of hyponym-hypernym pairs, had disappointing results: the KL divergence was unable to outperform self-information, and both measures misclassified around 20% of our testset.

Our error analysis showed that several factors contributed to the misclassifications. First, distributions are unable to model extensional properties which, in many cases, account for the feeling that a word is more general than another. Second, strong collocation effects can influence the measurement of information negatively: it is an open question which phrases should be considered 'words-with-spaces' when building distributions. Finally, distributional representations do not distinguish between word senses, which in many cases is a desirable feature, but interferes with the task we suggested in this work.

To conclude, we would like to stress that we do not think another information-theoretic measure would perform hugely better than the KL divergence. The point is that the nature of distributional vectors makes them sensitive to word usage and that, despite the general assumption behind distributional semantics, word usage might not suffice to model all aspects of lexical semantics. We leave as an open problem the issue of whether a modified form of our 'basic' distributional vectors would encode the right information.

[4] Although it is more difficult to talk of the extension of e.g. adverbials (*very*) or some adjectives (*skillful*), the general point is that text is biased towards a certain usage of words, while the general meaning a competent speaker ascribes to lexical items does not necessarily follow this bias.

Acknowledgements

This work was funded by a postdoctoral fellowship from the Alexander von Humboldt Foundation to the first author, and a Title A Fellowship from Trinity College, Cambridge, to the second author.

References

Baroni, Marco, and Lenci, Alessandro. 2008. Concepts and properties in word spaces. In Alessandro Lenci (ed.), *From context to meaning: Distributional models of the lexicon in linguistics and cognitive science* (Special issue of the Italian Journal of Linguistics 20(1)), pages 55–88.

Baroni, Marco, Raffaella Bernardi, Ngoc-Quynh Do and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL2012)*, pages 23–32.

Baroni, Marco, Raffaella Bernardi, and Roberto Zamparelli. 2012. Frege in Space: a Program for Compositional Distributional Semantics. Under review.

Curran, James. 2003. *From Distributional to Semantic Similarity*. Ph.D. thesis, University of Edinburgh, Scotland, UK.

Erk, Katrin. 2012. Vector space models of word meaning and phrase meaning: a survey. *Language and Linguistics Compass*, 6:10:635–653.

Erk, Katrin. 2013. Towards a semantics for distributional representations. In *Proceedings of the Tenth International Conference on Computational Semantics (IWCS2013)*.

Evert, Stefan. 2004. *The statistics of word cooccurrences: word pairs and collocations*. Ph.D. thesis, University of Stuttgart.

Leech, Geoffrey, Roger Garside, and Michael Bryant. 1994. Claws4: The tagging of the british national corpus. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*, pages 622–628, Kyoto, Japan.

Lund, Kevin, Curt Burgess, and Ruth Ann Atchley. 1995. Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th annual conference of the Cognitive Science Society*, Vol. 17, pages 660–665.

McNally, Louise. 2013. Formal and distributional semantics: From romance to relationship. In *Proceedings of the 'Towards a Formal Distributional Semantics' workshop*, 10th International Conference on Computational Semantics (IWCS2013), Potsdam, Germany. Invited talk.

Mitchell, Jeff and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429, November.

Resnik, Philipp. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 448–453.

Searle, John R. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge University Press.

Turney, Peter D. and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.