Using cooccurrence statistics and the web to discover synonyms in a technical language

Marco Baroni and Sabrina Bisi

SSLMIT, University of Bologna
Corso della Repubblica 136, 47100 Forlì, Italy
{baroni,sbisi}@sslmit.unibo.it

Abstract

Turney (2001) has shown that computing the mutual information of a pair of words by using cooccurrence counts obtained via queries to the AltaVista search engine performs very effectively in a synonym detection task. Since manual synonym detection is a challenging task for terminologists, we investigate whether the AltaVista-based Mutual Information (AVMI) method can be applied to the task of finding pairs of synonyms in the lexicon of a specialized sub-language. In particular, we experiment with synonyms in the field of nautical terminology. Our results indicate that AVMI is very good at spotting synonym couples among pairs of unrelated terms (with precision close to 90% at 62.5% recall) and that it outperforms more standard methods based on contextual cosine similarity. However, AVMI is not able to distinguish between synonyms and other semantically related terms. Thus, AVMI can be used for synonym mining only if it is combined with techniques to filter out other semantic relations.

1. Introduction

Identifying synonyms is an important step in the development of structured terminological databases (i.e., termbases).

First, the number and nature of the synonyms present in the analyzed domain will affect important design choices, such as whether the termbase should be structured around synonym sets (synsets) or whether each synonym should constitute a separate node of the hierarchy.

Second, if the termbase has a normative function, one must choose a recommended form for each synset.

Third, in the construction of multilingual termbases synsets must be analyzed in order to decide how to connect synonyms across languages.

However, in our experience as terminologist trainers, synonym identification on the basis of subjective intuitions is seen as a daunting task. This is probably due to the fact that synonymy is a hard-to-define property that has no clear-cut boundaries.

For example, the Merriam-Webster On-Line[1] defines a synonym as "one of two or more words or expressions of the same language that have the same or nearly the same meaning in some or all senses." Of course, it is extremely difficult to develop robust intuitions or explicit criteria to decide when the meanings of two words are "near" enough to justify treating them as synonyms.

Thus, an automated procedure that identifies synonyms on the basis of objective distributional grounds would be of great help to terminologists.

Recently, Turney (2001) has shown that a simple algorithm applied to a very large corpus (the web) performs remarkably well on the task of identifying the synonym of a target word, given a set of candidates. The method is to compute the mutual information of a pair of words by using frequency/cooccurrence frequency data extracted from the web via the AltaVista search engine[2], and to rank the pairs on the basis of this score. Pairs with a high AltaVista-based Mutual Information (AVMI) score are more likely to be synonyms.

The AVMI method differs from more traditional corpus-based approaches to synonymy detection in that it does not look at the context in which words occur, but simply at their direct cooccurrence. The simplicity of the approach is counterbalanced by the fact that its statistics are based on a very large corpus (the web).

The fact that AVMI does not require annotated language resources nor specialized NLP tools makes it a particularly attractive method to discover synonyms during rapid multilingual termbase development.

However, Turney studied synonyms belonging to the general language. Automated synonym detection in a technical sub-language is likely to be a harder task, since most terms from a specialized domain will tend, to a certain degree, to be semantically related.

In this study we test the validity of AVMI when applied to the sub-language of English nautical terminology (Bisi, 2003).

Our results are very promising, indicating that the AVMI approach can be extended to the terminological domain. Also, we show that this approach outperforms a more standard method based on contextual cosine similarity, where contextual vectors are computed on smaller specialized corpora.

However we also found that, while AVMI can distinguish between synonym pairs and random combinations of terms very effectively, it is not able to distinguish between synonyms and other terms with strong semantic links, such as hypo-/hypernym pairs or cohyponyms.

Thus, AVMI can be used for synonym mining only if it is combined with techniques to filter out other semantic relations.

The remainder of this paper is organized as follows. In 2. and 3. we present the measures we tested and we shortly review some of the relevant literature. In 4. and 5. we present our experiments. In 6. we draw our conclusions.

[1] http://www.m-w.com/
[2] http://www.altavista.com

2. Cooccurrence-based similarity: AltaVista-based Mutual Information

(Pointwise) Mutual Information (MI) was first introduced to computational linguistics by Church and Hanks (1989). The mutual information between two words w_1 and w_2 is:

$$MI(w_1, w_2) = \log_2 \frac{\mathrm{P}(w_1, w_2)}{\mathrm{P}(w_1)\mathrm{P}(w_2)} \quad (1)$$

There is a vast literature on MI applied to collocation mining (Manning and Schütze, 1999, Ch. 5), and it is also known that MI computed using large cooccurrence windows detects topically related words (Brown et al., 1990).

Turney (2001) has shown that MI computed on a very large corpus (the web) and using a medium-sized cooccurrence window can be used to find synonyms. Evidently (and somewhat surprisingly), synonyms have a tendency to occur in the near of each other.

In particular, Turney uses AltaVista to collect occurrence and cooccurrence frequencies. The latter are computed using the AltaVista NEAR operator, which returns pages in which the two target words occur within 10 words of one another, in either order.

Turney applies the AltaVista-based Mutual Information method to the TOEFL synonym match problem. The task is to choose the synonym of a word from a set of four candidates (e.g., for the target *levied* one has to choose a synonym from *imposed, believed, requested, correlated*). Turney's algorithm picks the candidate that has the highest AVMI with the target as the true synonym.

The AVMI method has a success rate of 72.5% on a test set of 80 synonym match problems. This is particularly impressive given that the average success rate of foreign students taking the TOEFL is reported to be 64.5%, i.e., AVMI performs 8% better than the average test taker.

Terra and Clarke (2003) test MI and other measures on the TOEFL synonym match task by extracting counts from a very large web-derived corpus (53 billion words). Their results confirm the effectiveness of MI which, with the best parameter settings, reaches a success rate of 81.25%.

We compute AVMI using the following formula:

$$AVMI(w_1, w_2) = \log_2 N \frac{\mathrm{hits}(w_1 \,\mathrm{NEAR}\, w_2)}{\mathrm{hits}(w_1)\mathrm{hits}(w_2)} \quad (2)$$

Here, $\mathrm{hits}(w_1 \,\mathrm{NEAR}\, w_2)$ is the number of hits (documents) returned by AltaVista for a query in which the two target terms are connected by the NEAR operator and $\mathrm{hits}(w_n)$ is the number of hits returned for a single term query. We set N, the number of documents indexed by AltaVista, to 350 millions.[3]

[3]This is probably an obsolete estimate. However, the N term, being constant, has no effect on the relative rank of pairs. For our purposes, it functions as a scaling factor.

3. Context-based similarity: The cosine approach

We compared AVMI to the more standard contextual cosine similarity approach.[4] For a detailed discussion of this approach, which is based on the intuition that similar words will tend to occur in similar contexts, see Manning and Schütze (1999, sec. 8.5).

Contextual similarity is computed by building a vector that collects the frequencies of cooccurrence of the target words with all the words in the corpus, or with a subset of these. The cosine of two normalized contextual vectors is given by their dot product:

$$\cos(\overrightarrow{x}, \overrightarrow{y}) = \overrightarrow{x} \cdot \overrightarrow{y} = \sum_{i=1}^{n} x_i y_i \quad (3)$$

The cosine ranges from 1, for perfectly correlated vectors, to 0 for totally uncorrelated vectors, to -1 for perfectly inversely correlated vectors.

Notice that context-based methods, since they require the construction of cooccurrence vectors for each target word and the comparison of such vectors, are harder to scale up to very large corpora than similarity measures based on direct cooccurrence counts.

Latent semantic analysis (LSA) is a particularly sophisticated (and computationally intensive) version of the contextual similarity approach. Through dimensionality reduction techniques, LSA takes into account not only cooccurrence with the same words, but also cooccurrence with words that are similar to each other.

Landauer and Dumais (1997) applied LSA to the TOEFL synonym detection task with a success rate of 64.4%. This is comparable to the performance of the average foreign test taker (64.5%), but considerably lower than the success rates attained by Turney and Terra and Clarke (72.5% and 81.25%, respectively) with a simpler algorithm and a much larger corpus (Landauer and Dumais used a corpus of 4.7 million words).

In our experiments we computed cosine similarity scores by collecting frequency counts from two corpora.[5]

The first corpus (described in detail by Bisi (2003)) contains about 1.2 million words and it is made of documents that were hand-picked from the web and from paper sources because of their representativeness of the domain of nautical terminology.

We also built a larger corpus (4.27 million words) via automated queries for random combinations of nautical terms to the Google search engine.[6]. By informal inspection, most documents in this corpus are valid examples of

[4]It seems fair to compare AVMI to another knowledge-free unsupervised measure. Moreover, while similarity measures relying on knowledge sources such as WordNet perform very well (see, e.g., Budanitsky and Hirst (2001)), the relevant resources are rarely available to terminologists, especially if they work on languages other than English.

[5]We leave it to further research to develop a context-based model based on AltaVista-derived statistics, and possibly to experiment with LSA.

[6]`http://www.google.com`

the specialized language under investigation.[7]

Before collecting context vectors, we removed the 200 words that have the highest document frequency in the Brown corpus (Kučera and Francis, 1967) from both corpora. These are mostly function words. All other words were treated as potential dimensions of the similarity vectors.

Context vectors were collected using two relatively narrow windows: 2 words to either side of the target and 5 words to either side of the target. In both cases, words that cooccurred only once with a target were not counted.

In what follows, we use the following codes for the contextual cosine measures: COST2 (counts from the terminologist's hand-picked corpus, 2-word window), COST5 (terminologist corpus, 5-word window), COSW2 (web corpus, 2-word window) and COSW5 (web corpus, 5-word window).

We also experimented with combinations of direct cooccurrence and context similarity scores. We will only report results for the best performing combination, which was obtained by summing the ranks of the pairs in the lists ordered by AVMI and COSW2.

4. Synonym pairs vs. random pairs

In the first experiment, we use AVMI and cosine similarity to look for synonyms among random pairs of terms from the same domain.

4.1. Test set

The 148 word test set contains the 24 synonyms pairs included in the termbase of Bisi (2003) (pairs such as *bottom/hull*, *frames/ribs*, *displacement/weight*) and 124 non-synonym pairs. 24 of these were created by re-combining the terms in the synonym set (in order to control for possible term-specific effects). The remaining 100 pairs were constructed by forming random couples of 200 nautical terms.

Since all the non-synonym pairs belong to the same domain, they are all, at the very least, topically related, and they often have stronger semantic links. Before running our tests, we classified the random pairs into strongly semantically related vs. not (strongly) related, on the basis of our intuition and knowledge of the domain. We judged 36 pairs (29%) to be formed by strongly semantically related terms. Examples of such pairs include *decks/cockpit*, *awning/stern board*, *install/hatch*, *keel/coated* and *underway/cruising*.[8]

Synonym detection in this setting is clearly a harder task than if the non-synonym pairs had been taken from general vocabulary.

4.2. Results and discussion

Table 1 reports percent precision at eight recall levels for the various similarity measures.[9]

AVMI is by far the best measure, outperforming all the context-based measures and the AVMI/COSW2 combination. At a recall of 62.5% (15 synonym pairs found over 24 present), AVMI has still a precision of 88.2%, i.e., only 2 non-synonym pairs are mixed with the synonyms.

This is a very promising result, confirming that the direct cooccurrence method applied to a very large corpus outperforms the context-based method applied to smaller corpora, and that the web is a large enough corpus that this is true even if we are dealing with terms from a specialized domain.

Among the variants of the context-based approach, the ones using a narrower window perform better, but they have worse data scarcity problems, with big tails of synonym pairs that are at the bottom of the list because they are made of terms with extremely sparse, non-overlapping vectors.

The contextual measures derived from the terminologist corpus outperform those based on the web if we look at the top of the lists, but, as recall increases, their precision drops faster to values close to or below chance level, reflecting the quality/quantity tradeoff that we expect between a smaller hand-built corpus and a larger automatically constructed one.

Interestingly, by inspecting the 31 false positives (non-synonym pairs) that we get at the 75% recall level in the AVMI list, we find that 14 of them (about 45%), are among the pairs that we judged to be strongly semantically related (see previous section). This is 16% higher than the proportion of strongly related pairs in the whole list of non-synonyms (29%).

Thus, a consistent portion of the "rivals" of the synonyms are pairs with strong semantic links of other kinds. This suggests that the MI method is detecting semantic relations in general, and not synonymy in particular.

5. Synonym pairs vs. other semantically related pairs

To further explore the issue raised at the end of the previous section, we ran a second experiment in which the test set was modified to systematically include other types of semantically related words.

5.1. Test set

We added 31 more pairs from the Bisi termbase to the test set described in 4.1. Of these, 19 are cohyponym pairs (e.g., *Bruce anchor/mushroom anchor*, *flexible tank/rigid tank*), 10 are hypo-/hypernym pairs (e.g, *stern platform/sun deck*, *awning/canopy*) and 2 are antonyms (*ahead/astern*, *aboard/overboard*). The cohyponym and hypo-/hypernym pairs include all the instances of these relations in the termbase that share one term with one of the synonym pairs. The antonym pairs are the only two pairs of this type in the termbase.

In order to maintain the same synonym-to-non-synonym ratio as in the first experiment, we removed 31 randomly picked non-synonym pairs from the test set.[10]

[7] The tools we used to build this corpus are now available as the BootCat Tools (Baroni and Bernardini, 2004).

[8] The list did not contain any pair that could be categorized in terms of synonymy, hyponymy or other standard relations recorded in our termbases.

[9] In the case of ties, we treated synonym pairs as if they were ranked *below* non-synonym pairs.

[10] We repeated the experiment by removing different random subsets, and we consistently obtained results similar to those we report.

recall	AVMI	COST2	COST5	COSW2	COSW5	COMB
3 (12.5%)	100	100	60	60	42.9	100
6 (25%)	100	75	60	46.2	46.2	66.7
9 (37.5%)	90	42.9	39.1	40.9	45	52.9
12 (50%)	92.3	17.9	19.4	26.7	25.5	41.4
15 (62.5%)	88.2	10.8	15	19	17.6	27.3
18 (75%)	36.7	12.7	12.7	12.7	13.4	23.4
21 (87.5%)	30.4	14.5	14.5	14.5	14.5	23.3
24 (100%)	16.2	16.2	16.2	16.2	16.2	16.2

Table 1: Synonyms vs. random pairs: Percentage precision at 8 recall levels.

recall	AVMI	COST2	COST5	COSW2	COSW5	COMB
3 (12.5%)	60	42.9	37.5	27.3	20	25
6 (25%)	33.3	46.2	46.2	28.6	27.3	27.3
9 (37.5%)	36	39.1	39.1	29	31	31
12 (50%)	40	19.7	21.1	23.1	22.6	29.3
15 (62.5%)	37.5	10.8	17.4	19.2	18.1	24.6
18 (75%)	26.5	12.7	12.7	12.7	14.1	22.8
21 (87.5%)	25.6	14.5	14.5	14.5	14.5	21.4
24 (100%)	16.2	16.2	16.2	16.2	16.2	16.2

Table 2: Synonyms vs. random and related pairs: Percentage precision at 8 recall levels.

5.2. Results and discussion

Table 2 reports percent precision at eight recall levels with the modified test set.

AVMI is still the best measure (at most recall levels), but its performance has dropped dramatically. This confirms that AVMI is good for finding semantically related terms in general, but it does not single out synonyms when other strongly related terms are present.

The drop in performance is due to the combined effect of hypo-/hypernymy and cohyponymy. For example, among the top 40 pairs in the ranked AVMI list we find 15 synonym pairs, 15 cohyponym pairs and 8 hypo-/hypernym pairs, for a total of 38 "nymic" pairs. The two antonyms are not interfering with synonym detection, as they are at ranks 47 and 146, respectively.

Performance of the cosine-based and combined measures did not drop as much as that of AVMI, but this is probably due to a "flooring" effect.

6. Conclusion

Our study confirms the effectiveness of medium-distance MI measured on a very large corpus as a semantic similarity measure, extending the results of Turney (2001) and Terra and Clarke (2003) to a technical domain, where all terms tend to a certain extent to be related, and thus the task of identifying the most similar terms becomes harder.

However, our results also suggest that AVMI is a good measure to find semantically related pairs in general, not synonyms in particular.

The impressive results of Turney (2001) and Terra and Clarke (2003) are probably due to the fact that in most of their synonym candidate sets the true synonym was the only word to be semantically related to the target.

To turn AVMI into a full-fledged synonym mining technique, it is necessary to find ways to filter out other types of "nyms" from the results. This step can be performed manually by a terminologist (discovery of hyponymy and cohyponymy is not seen as a particularly challenging task). Alternatively, the process of filtering out semantically related non-synonyms could be automated.

Turney (2001) and Lin et al. (2003) collect web-based cooccurrence statistics with special queries that are biased against antonyms. In future research, we plan to extend their approach, by developing special queries that disfavor other nymic relations and/or favor synonym detection.

7. References

M. Baroni and S. Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. *LREC 2004*.

S. Bisi. 2003. *Verso una terminologia descrittiva nel quadro di un approccio apertamente linguistico: ricerca terminologica bilingue italiano-inglese nel campo dello yacht a motore.* Tesi di Laurea, SSLMIT.

P. Brown, V. Della Pietra, P. DeSouza, J. Lai, and R. Mercer. 1990. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

A. Budanitsky and G. Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *Proceedings of Workshop on WordNet and Other Lexical Resources of NAACL 2001*.

K. Church and P. Hanks. 1989. Word association norms, mutual information, and lexicography. *ACL 1989*, 76-83.

H. Kučera and N. Francis. 1967. *Computational analysis of present-day American English*. Brown University Press.

T. Landauer and S. Dumais. 1997. A solution to Plato's problem: A latent semantic analysis theory of the acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–240.

D. Lin, S. Zhao, L. Qin, and M. Zhou. 2003. Identifying synonyms among distributionally similar words. *IJCAI 2003*.

C. Manning and H. Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press.

E. Terra and C. L. A. Clarke. 2003. Frequency estimates for statistical word similarity measures. *HLT-NAACL 2003*, 165–172.

P. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. *ECML 2001*, 491–502.