Text Tokenization for Knowledge-free Automatic Extraction of Lexical Similarities

Aristomenis Thanopoulos, Nikos Fakotakis and George Kokkinakis

Electrical and Computer Engineering Department, University of Patras
26500, Rion, Greece
aristom@wcl.ee.upatras.gr

Abstract

Previous studies on automatic extraction of lexical similarities have considered as semantic unit of text the *word*. However, the theory of contextual lexical semantics implies that larger segments of text, namely *non-compositional multiwords*, are more appropriate for this role. We experimentally tested the applicability of this notion applying automatic collocation extraction to identify and merge such multiwords prior to the similarity estimation process. Employing an automatic WordNet-based comparative evaluation scheme along with a manual evaluation procedure, we ascertain improvement of the extracted similarity relations.

Keywords

Automatic methods, lexical similarity extraction, collocation extraction, automatic evaluation.

1 Introduction

Lexico-semantic knowledge is highly important for many NLP applications, such as language modeling, word sense disambiguation, construction of ontologies and thesauri, information extraction, machine translation, language generation, etc. Moreover, it is language and domain dependent. Since existing electronic lexico-semantic resources, e.g. WordNet (Miller, 1990) lack full coverage across both directions, automatic acquisition of such knowledge from corresponding text corpora is an attractive and economic solution.

A type of semantic knowledge is lexical similarity, referred to as *semantic similarity* or *word similarity* as well. The majority of Automatic Lexico-semantic Similarity Extraction (ALSE) techniques employ the notion that semantic properties of words can be defined "by their actual and potential context" (Cruse, 1986). Therefore, semantic similarity between two words (the *focus* or *target words*) can be estimated by the similarity of their contextual environments, that is the words (the *parameter words*) they habitually co-occur with. For example, the pairs of expressions *a bottle of wine* - *a bottle of sangria*, *drinking wine* - *drinking sangria*, etc. suggest that *wine* and *sangria* are lexically and conceptually similar.

Variations of ALSE approaches regard the contextual environment, which is either the local context of the target word (window methods) or its syntactic dependencies (syntactic methods). In both cases the contextual scope may vary. Nevertheless, all previous approaches considered as basic unit of text the *word*. In this paper we introduce the idea that a more efficient preprocessing for ALSE is to identify non-compositional collocations, and consider each one of them as a single lexical unit. We base our opinion on the theoretical foundations of contextual lexical semantics and we tested it conducting experiments within a knowledge-free ALSE framework. We employed existing semantic resources to provide automatic comparative evaluation along with manual evaluation.

We adopt a knowledge-free approach to ALSE, for the sake of simplicity and portability across languages and domains. That is, apart from raw text corpora, no other knowledge sources or sophisticated linguistic tools are employed. In order to avoid the need for a parser, we employed local context adjacency. Moreover, Grefenstette (1994) found that for the less frequent words, that is, for most words, the window method outperformed the syntax-based method. Knowledge-free approaches for semantic similarity estimation consider as textual environment the content words contained in a text window centered on the target word (Grefenstette, 1994; Schütze, 1998; Martin et al, 1998). Approaches using large context windows are computationally expensive and their output indicates topic-similarity rather than semantic similarity, being therefore effective for word sense discrimination (Schütze, 1998). Having confirmed this by comparative experiments, we employed only next and previous word adjacency. In order to exploit information about precedence and succession in local context, we employed an information theoretic similarity measure proposed by Lin (1998a), which distinguishes between different types of dependence relations. Functional words were disregarded.

2 The notion of Lexico-Semantic Unit

Cruse (1986), defining the notion of the textual entity *word* from the perspective of contextual lexical semantics, describes it as “the lexical element which is typically the smallest element of a sentence which has positional mobility and the largest unit which resists interruption by the insertion of new material between its constituent parts”. Although the word is indeed the lexical element which *typically* complies with both requirements, there are plenty of word sequences which satisfy them as well, such as *New York* and *kick the bucket*. Cruse describes them as “minimal semantic constituents which consist of more than one word”; we call them *non-compositional multiwords* (NCMs).

Consideration of NCMs as single semantic units assists in two directions. At first, some spurious contextual data are eliminated; in particular co-occurrences of the parts of non-compositional expressions with adjacent words outside the collocation. For example, considering *mutual* as a contextual argument of *fund* we are likely to erroneously correlate *fund* with nouns often co-occurring with *mutual*, such as *agreement*, *respect*, etc. Secondly, the useful contextual data for classifying the multiword entities themselves are increased. Considering, for example, the pairs of phrases *capital Beijing – capital Hong Kong* and *Beijing government – Hong Kong government*, it is obvious that in order to completely exploit the common contextual elements it is necessary to identify first that *Hong_Kong* should be considered as a single lexical item.

On the other hand, it is important to keep constituents of compositional multiwords as distinct lexical tokens. Failure on this matter generates two degrading factors as far as the extracted knowledge is concerned. At first, the syntactic relations and thus the semantic bonds included in every compositional sequence are not exploited. For example, handling the collocations *President Reagan* and *President Bush* as single lexical units, we miss *President* as the common contextual element of *Reagan* and *Bush*, helpful for ascertaining their semantic similarity. Second, the corpus vocabulary is increased because the constituents of such collocations occur independently in the text as well and the contextual data become sparser. That is, both *President Reagan* and *Reagan* would be maintained in the corpus.

3 Extraction of multiwords

Statistical metrics employed for knowledge-free corpus-based extraction of collocations, are the t-score, χ^2 score, likelihood ratio or pointwise mutual information (Manning and Schütze, 1999). They deduce collocationhood comparing observed with expected-by-chance n-gram frequencies. In fact they don't guarantee non-compositionality; it is rather a likely consequence of strong collocationality. More confident techniques for the extraction of NCMs require other resources, such as WordNet (Pearce, 2001) or syntactically analyzed textual data (Sekine et al, 1992). However, since they fall outside our knowledge-free discipline, we did not employ them in the present study.

3.1 The measures

Likelihood Ratio (LR) is a mostly employed measure of statistical significance. However, although it reveals statistically strong correlations, it is not an equally good indicator of non compositionality. For example, its 10-best list from WSJC includes the bigrams *years earlier* and t*his year*, due to their very high frequencies and despite their compositionality, which is apparent in the corpus as well; e.g. consider {*week, day, month, year*} *earlier*.

Mutual dependency, a measure derived from pointwise mutual information (Thanopoulos et al., 2002), promotes multiwords occurring more often than not tied together:

$$D(w_1, w_2) = \log_2 \frac{P^2(w_1 w_2)}{P(w_1) \cdot P(w_2)} = k + \log_2(\lambda_1 \cdot \lambda_2), \text{ where } \lambda_1 \equiv \frac{c_{12}}{c_1}, \ \lambda_2 \equiv \frac{c_{12}}{c_2}.$$

Setting $\lambda_{min}=\min(\lambda_1,\lambda_2)$, the maximum value of D is obtained when $\lambda_{min} = 1$. In this case the constituent words are unbrokenly tied together, e.g. *Ku Klux Klan*. In order to allow for higher coverage (i.e. more multiwords to be identified) we relax this condition (i.e $\lambda_{min} = 1$) allowing for lower values of λ_{min}. For example, the condition $\lambda_{min} > \lambda^T = 0.5$ is quite safe, allowing only bigrams with both their elements occurring together more often than not.

3.2 The algorithm

Since measures of statistical significance of serial events can be applied directly only to pairwise data, we applied an agglomerative algorithm, similar to Smadja's (1993), which constructs statistically significant n-grams by iteratively merging significant bigrams. Let ***m(x,y)*** a measure of statistical significance, expressing the degree of statistical correlation between tokens **x** and **y** in corpus **C**. The algorithm is formally described in Figure 1. The operator $\oplus$ represents string concatenation. Step 2 represents the fact that the *word* is unbreakable. In step 3 the list of new significant bigrams is constructed. For every n-gram the minimum value of **m** calculated during its construction. E.g., the bigrams *New York* and *Stock Exchange* were extracted on the 1st iteration and the 4-gram *New York Stock Exchange* on the 2nd.

```
1. Set a cut-off threshold $\mathbf{m_T}$.
2. Initialize: $V_m = \{(w_i,M): w_i \in \text{vocabulary}(\mathbf{C}), M \rightarrow \infty\}$
3. Set $V' = \{(w_i \oplus w_j, m_{ij}): (w_i,m_i),(w_j,m_j) \in V , m(w_i,w_j) > m_T, m_{ij} = \min(m_i,m_j,m(w_i,w_j))\}$
4. $\forall (w_i,w_j) \in V'$: Obtain new $\mathbf{C}$ by replacing bigrams $(w_i,w_j)$ with $w_i \oplus w_j$ throughout $\mathbf{C}$.
5. Set new $V = V + V'$
6. Collect co-occurrence statistics from the new corpus.
7. Repeat from Step-3 **R** times or until no new significant n-grams are formed
8. Keep the **N**-best pairs (considering $\mathbf{m_{ij}}$)
```

Figure 1. The algorithm for the N-best n-gram collocation extraction

Observation of the extracted multiword lists indicated that numerous errors (compositional collocations) regard expressions consisting of a named entity and a common word, (*Reagan administration*, *Java programmers*). In order to improve the performance of multiword extraction we keep only the n-grams in which all the words have the case of the initial character in common. We should note that this rule is ineffective for languages in which common nouns appear with the initial character in upper case, such as German.

4 Comparative Evaluation

Several researchers have performed lexicographic evaluation of the outcome of their work on lexical acquisition (Smadja, 1993; Schütze, 1998). Such approaches guarantee a high accuracy and coverage, but they are expensive, time-consuming and not easily repeatable. This is a notable drawback in an algorithm comparison task. For this reason, Grefenstette (1994) and Lin (1998b) exploited existing lexico-semantic resources, such as WordNet and Roget thesaurus to provide comparative evaluation. Indeed, automatic evaluation, besides faster and cheaper than manual evaluation, allows repeatability of the experiments and hence comparison between alternate approaches. Therefore, we followed both evaluation approaches separately: Automatic evaluation against WordNet and additional lists of entities and manual inspection by domain experts in order to deal with economic terminology.

4.1 Automatic evaluation

The main body of our gold standard is WordNet. Although several named entities are included, there has been no systematic effort of storing such information (Miller, 1990). Since newswire text abounds in named entities, it is purposeful to utilize relevant semantic information. Considering the nature of WSJC, we used lists of entity names obtained from publicly available Internet databases; specifically American locations, companies and organizations. A similarity relation is considered correct if the lemmas of the related words are found in the same set of entities, or they appear to be synonyms or antonyms in WordNet, or the similarity of their respective concepts in the WordNet hierarchy, according to the information theoretic measure proposed in (Lin, 1998b) exceeds a certain threshold T_{sim}. The concept probabilities were calculated in a similar way as in (Resnik, 1999).

4.2 Manual evaluation

Although manual evaluation of the extracted long list of relations is a burdensome task, it is quite easier to perform comparative evaluation between two different but largely overlapping resources. Since the largest portion of the extracted resources is common, we confine the evaluation to the different parts. That is, if $S_J^{(N)}$ is the set of the N-best relations produced, then it is sufficient to evaluate the sets $S_1^{(N)}$-$S_2^{(N)}$ and $S_2^{(N)}$-$S_1^{(N)}$. A domain expert evaluates every relation as correct, incorrect, or undecided. The latter regards relations between distant concepts but not irrelevant.

5 Experimental results

We performed the described algorithm on a 42 million words portion of the Wall Street Journal corpus (years 1987, 1988, 1989). The corpus was tokenized according to two alternative approaches: Word-based segmentation (A1) and merging of the N-best (N_1=1000, N_2=3000) multiwords extracted, according to either LR or λ_{min} (A2). In all the cases we applied the aforementioned automatic evaluation process, employing WordNet (we set T_{sim} = 0.6), enhanced with named entity information. The accuracy of ALSE considering the N-best similarity relations across N is depicted on Figure 2. Furthermore, keeping the 10000-best relations for strategies (A1) and (A2, m=D, N_1=1000), we subjected the relations occurring only to the one and not the other to expert judgment (see Table 1).

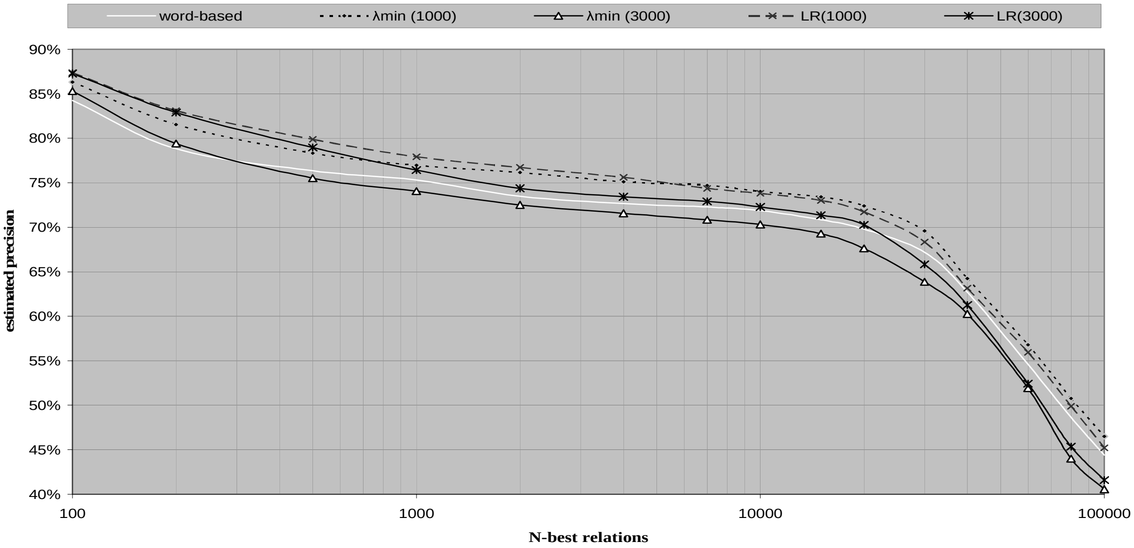

Figure 2. Comparative evaluation results.

The results show that multiword-based tokenization performs consistently better than word-based, using the 1000-best multiwords for both metrics, while λ_{min} performs better than LR, projecting mainly NCMs. However, the performance is low in the case of the 3000-best collocations, due to allowance of many compositional expressions in the extracted multiwords. It should be noted though that in newswire corpora such as the WSJC numerous compositional expressions occur almost invariably because they are used to describe events repeated on a daily basis (e.g. “spokesman said”); therefore collocation-based text segmentation is probably more efficient in a balanced corpus.

Corpus origin ***Evaluation***	Plain & ┐ tokenized	Tokenized & ┐ plain
Correct	141	252
Incorrect	118	35
Undecided	220	192

Table 1. Expert evaluation for the 10000-best relations.

Conclusion

Considering knowledge-free techniques for the automatic extraction of lexical similarities crucial for the sake of portability of NLP systems, we investigated, using both theoretic and experimental means, the effect of multiword-based segmentation prior to lexical similarity extraction. In order to achieve reliable and inexpensive comparative evaluation, we employed both WordNet-based automatic and manual evaluation. Although non-compositionality of the collocations was derived simply by raw corpus statistics, based on the evidence of strong collocationality, ALSE performance increases. This exhibits that the proposal of semantic tokenization is of practical usefulness and can be exploited more efficiently, provided that more accurate detection of non-compositional multiwords is achieved, with the usage of higher level NLP tools and resources.

References

Cruse D.A. (1986) Lexical Semantics, Cambridge University Press.

Grefenstette G. (1994) Explorations in Automatic Thesaurus Discovery, Boston, Kluwer.

Lin D. (1998a). An Information-Theoretic Definition of Similarity. Proceedings of International Conference on Machine Learning.

Lin D. (1998b). Automatic retrieval and clustering of similar words, Proc. of COLING-ACL.

Manning C. and Schütze H. (1999) Foundations of Statistical Natural Language Processing, MIT Press, Cambridge.

Martin S., Liermann J., Ney H. (1998) Algorithms for bigram and trigram word clustering, Speech Communication, Vol.24, pp.19-37.

Miller G. (1990) Wordnet: An on-line lexical database, International Journal of Lexicography, Vol.3.

Pearce, D. (2001) Synonymy in collocation extraction, Proceedings of the NAACL'01 Workshop on WordNet and other Lexical Resources.

Schütze H. (1998) Word Sense Discrimination. Computational Linguistics, Vol.24, pp.97-124.

Sekine, S., Carroll J. J., Ananiadou S. and Tsujii J. (1992) Automatic Learning for Semantic Collocation, Proceedings of the 3rd Conference on Applied NLP, ACL, pp.104-110.

Smadja F. (1993) Retrieving Collocations from text: Xtract, Computational Linguistics vol. 19.

Thanopoulos A., Fakotakis N. and Kokkinakis G. (2002) Comparative Evaluation of Collocation Extraction Metrics. Proceedings of the LREC2002 Conference, pp.609-613.