FBK-irst: Lexical Substitution Task Exploiting Domain and Syntagmatic Coherence

Claudio Giuliano and **Alfio Gliozzo** and **Carlo Strapparava**
FBK-irst, I-38050, Povo, Trento, ITALY
{giuliano, gliozzo, strappa}@itc.it

Abstract

This paper summarizes FBK-irst participation at the lexical substitution task of the SEMEVAL competition. We submitted two different systems, both exploiting synonym lists extracted from dictionaries. For each word to be substituted, the systems rank the associated synonym list according to a similarity metric based on Latent Semantic Analysis and to the occurrences in the Web 1T 5-gram corpus, respectively. In particular, the latter system achieves the state-of-the-art performance, largely surpassing the baseline proposed by the organizers.

1 Introduction

The lexical substitution (Glickman et al., 2006a) can be regarded as a subtask of the lexical entailment, in which for a given word in context the system is asked to select an alternative word that can be replaced in that context preserving the meaning. Lexical Entailment, and in particular lexical reference (Glickman et al., 2006b)[1], is in turn a subtask of textual entailment, which is formally defined as a relationship between a coherent text T and a language expression, the hypothesis H. T is said to entail H, denoted by $T \rightarrow H$, if the meaning of H can be inferred from the meaning of T (Dagan et al., 2005; Dagan and Glickman., 2004). Even though this notion has been only recently proposed in the computational linguistics literature, it attracts more and more attention due to the high generality of its settings and to the usefulness of its (potential) applications.

[1] In the literature, slight variations of this problem have been also referred to as *sense matching* (Dagan et al., 2006).

With respect to lexical entailment, the lexical substitution task has a more restrictive criterion. In fact, two words can be substituted when meaning is preserved, while the criterion for lexical entailment is that the meaning of the thesis is implied by the meaning of the hypothesis. The latter condition is in general ensured by substituting either hyperonyms or synonyms, while the former is more rigid because only synonyms are in principle accepted.

Formally, in a lexical entailment task a system is asked to decide whether the substitution of a particular term w with the term e in a coherent text $H_w = H^l w H^r$ generates a sentence $H_e = H^l e H^r$ such that $H_w \rightarrow H_e$, where H^l and H^r denote the left and the right context of w, respectively. For example, given the source word 'weapon' a system may substitute it with the target synonym 'arm', in order to identify relevant texts that denote the sought concept using the latter term.

A particular case of lexical entailment is recognizing synonymy, where both $H_w \rightarrow H_e$ and $H_e \rightarrow H_w$ hold. The lexical substitution task at SEMEVAL addresses exactly this problem. The task is not easy since lists of candidate entailed words are not provided by the organizers. Therefore the system is asked first to identify a set of candidate words, and then to select only those words that fit in a particular context. To promote unsupervised methods, the organizers did not provide neither labeled data for training nor dictionaries or list of synonyms explaining the meanings of the entailing words.

In this paper, we describe our approach to the Lexical Substitution task at SEMEVAL 2007. We developed two different systems (named IRST1-lsa and IRST2-syn in the official task ranking), both exploiting a common lists of synonyms extracted from dictionaries (i.e. WordNet and the Oxford Dictio-

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 145–148,
Prague, June 2007. ©2007 Association for Computational Linguistics

nary) and ranking them according to two different criteria:

Domain Proximity: the similarity between each candidate entailed word and the context of the entailing word is estimated by means of a cosine between their corresponding vectors in the LSA space.

Syntagmatic Coherence: querying a large corpus, the system finds all occurrences of the target sentence, in which the entailing word is substituted with each synonym, and it assigns scores proportional to the occurrence frequencies.

Results show that both methods are effective. In particular, the second method achieved the best performance in the competition, defining the state-of-the-art for the lexical substitution task.

2 Lexical Substitution Systems

The lexical substitution task is a textual entailment subtask in which the system is asked to provide one or more terms $e \in E \subseteq syn(w)$ that can be substituted to w in a particular context $H_w = H^l w H^r$ generating a sentence $H_e = H^l e H^r$ such that both $H_w \rightarrow H_e$ and $H_e \rightarrow H_w$ hold, where $syn(w)$ is the set of synonyms lemmata obtained from all synset in which w appears in WordNet and H^l and H^r denote the left and the right context of w, respectively.

The first step, common to both systems, consists of determining the set of synonyms $syn(w)$ for each entailing word (see Section 2.1). Then, each system ranks the extracted lists according to the criteria described in Section 2.2 and 2.3.

2.1 Used Lexical Resources

For selecting the synonym candidates we used two lexical repositories: *WordNet 2.0* and the *Oxford American Writer Thesaurus* (1^{st} Edition). For each target word, we simply collect all the synonyms for all the word senses in both these resources.

We exploited two corpora for our systems: the *British National Corpus* for acquiring the LSA space for ranking with domain proximity measure (Section 2.2) and the *Web 1T 5-gram Version 1* corpus from Google (distributed by Linguistic Data Consortium)[2] for ranking the proposed synonyms according to syntagmatic coherence (Section 2.3).

[2]Available from http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13.

No other resources were used and the sense ranking in WordNet was not considered at all. Therefore our system is fully unsupervised.

2.2 Domain Proximity

Semantic Domains are common areas of human discussion, such as Economics, Politics, Law (Magnini et al., 2002). Semantic Domains can be described by DMs (Gliozzo, 2005), by defining a set of term clusters, each representing a Semantic Domain, i.e. a set of terms having similar topics. A DM is represented by a $k \times k'$ rectangular matrix $\mathbf{D}$, containing the domain relevance for each term with respect to each domain.

DMs can be acquired from texts by exploiting term clustering algorithms. The degree of association among terms and clusters, estimated by the learning algorithm, provides a domain relevance function. For our experiments we adopted a clustering strategy based on Latent Semantic Analysis (LSA) (Deerwester et al., 1990), following the methodology described in (Gliozzo, 2005).

The input of the LSA process is a Term by Document matrix $\mathbf{T}$ of the frequencies in the whole corpus for each term. In this work we indexed all lemmatized terms. The so obtained matrix is then decomposed by means of a Singular Value Decomposition, identifying the principal components of $\mathbf{T}$.

Once a DM has been defined by the matrix $\mathbf{D}$, the Domain Space is a k' dimensional space, in which both texts and terms are associated to Domain Vectors (DVs), i.e. vectors representing their domain relevance with respect to each domain. The DV $\vec{t_i'}$ for the term $t_i \in \mathcal{V}$ is the i^{th} row of $\mathbf{D}$, where $\mathcal{V} = \{t_1, t_2, \ldots, t_k\}$ is the vocabulary of the corpus. The DVs for texts are obtained by mapping the document vectors $\vec{d_j}$, represented in the vector space model, into the vectors $\vec{d_j'}$ in the Domain Space, defined by

$$\mathcal{D}(\vec{d_j}) = \vec{d_j}(\mathbf{I^{IDF}}\mathbf{D}) = \vec{d_j'} \quad (1)$$

where $\mathbf{I^{IDF}}$ is a diagonal matrix such that $i_{i,i}^{IDF} = IDF(w_i)$ and $IDF(w_i)$ is the *Inverse Document Frequency* of w_i. The similarity among both texts and terms in the Domain Space is then estimated by the cosine operation.

To implement our lexical substitution criterion we ranked the candidate entailed words according to their domain proximity, following the intuition that if two words can be substituted in a particular context, then the entailed word should belong to the

same semantic domain of the context in which the entailing word is located.

The intuition above can be modeled by estimating the similarity in the LSA space between the pseudo document, estimated by Equation 1, formed by all the words in the context of the entailing word (i.e. the union of H^l and H^r), and each candidate entailed word in $syn(w)$.

2.3 Syntagmatic Coherence

The syntagmatic coherence criterion is based on the following observation. If the entailing word w in its context $H_w = H^l w H^r$ is actually entailed by a word e, then there exist some occurrences on the WEB of the expression $H_e = H^l e H^r$, obtained by replacing the entailing word with the candidate entailed word. This intuition can be easily implemented by looking for occurrences of H_e in the Web 1T 5-gram Version 1 corpus.

Figure 1 presents pseudo-code for the synonym scoring procedure. The procedure takes as input the set of candidate entailed words $E = syn(w)$ for the entailing word w, the context H_w in which w occurs, the length of the n-gram ($2 \leqslant n \leqslant 5$) and the target word itself. For each candidate entailed word e_i, the procedure $ngrams(H_w, w, e_i, n)$ is invoked to substitute w with e_i in H_w, obtaining H_{e_i}, and returns the set Q of all n-grams containing e_i. For example, all 3-grams obtained replacing "bright" with the synonym "intelligent" in the sentence "He was *bright* and independent and proud." are "He was intelligent", "was intelligent and" and "intelligent and independent". The maximum number of n-grams generated is $\sum_{n=2}^{5} n$. Each candidate synonym is then assigned a score by summing all the frequencies in the Web 1T corpus of the so generated n-grams[3]. The set of synonyms is ranked according the so obtained scores. However, candidates which appear in longer n-grams are preferred to candidates appearing in shorter ones. Therefore, the ranked list contains first the candidate entailed words appearing in 5-grams, if any, then those appearing in 4-grams, and so on. For example, a candidate e_1 that appears only once in 5-grams is preferred to a candidate e_2 that appears 1000 times in 4-grams. Note that this strategy could lead to an output list with repetitions.

[3]Note that n-grams with frequency lower than 40 are not present in the corpus.

```
Given E, the set of candidate synonyms
Given H, the context in which w occurs
Given n, the length of the n-gram
Given w, the word to be substituted
E' ← ∅
for each e_i in E do
    Q ← ngrams(H, w, e_i, n)
    score_i ← 0
    for each q_j in Q do
        Get the frequency f_j of q_j
        score_i ← score_i + f_j
    end for
    if score_i > 0 then add the pair {score_i, e_i} in E'
end for
Return E'
```

Figure 1: The synonym scoring procedure

3 Evaluation

There are basically two scoring methodologies: (i) BEST, which scores the best substitute for a given item, and (ii) OOT, which scores for the best 10 substitutes for a given item, and systems do not benefit from providing less responses[4].

BEST. Table 1 and 2 report the performance for the domain proximity and syntagmatic coherence ranking. Please note that in Table 2 we report both the official score and a score that takes into account just the *first* proposal of the systems, as the usual interpretation of BEST score methodology would suggest[5].

OOT. Table 4 and 5 report the performance for the domain proximity and syntagmatic coherence ranking, scoring for the 10 best substitutes. The results are quite good especially in the case of syntagmatic coherence ranking.

Baselines. Table 3 displays the baselines respectively for the BEST and OOT using WordNet 2.1 as calculated by the task organizers. They propose many baseline measures, but we report only the

[4]The task proposed a third scoring measure MW that scores precision and recall for detection and identification of multiwords in the input sentences. However our systems were not designed for this functionality. For the details of all scoring methodologies please refer to the task description documents.

[5]We misinterpreted that the official scorer divides anyway the figures by the number of proposals. So for the competition we submitted the oot result file without cutting the words after the first one.

	P	R	Mode P	Mode R
all	8.06	8.06	13.09	13.09

Table 1: BEST results for LSA ranking (IRST1-lsa)

	P	R	Mode P	Mode R
all	12.93	12.91	20.33	20.33
all (official)	*6.95*	*6.94*	*20.33*	*20.33*

Table 2: BEST results for Syntagmatic ranking (IRST2-syn)

WordNet one, as it is the higher scoring baseline. We can observe that globally our systems perform quite good with respect to the baselines.

4 Conclusion

In this paper we reported a detailed description of the FBK-irst systems submitted to the Lexical Entailment task at the SEMEVAL 2007 evaluation campaign. Our techniques are totally unsupervised, as they do not require neither the availability of sense tagged data nor an estimation of sense priors, not considering the WordNet sense order information. Results are quite good, as in general they significantly outperform all the baselines proposed by the organizers. In addition, the method based on syntagmatic coherence estimated on the WEB outperforms, to our knowledge, the other systems submitted to the competition. For the future, we plan to avoid the use of dictionaries by adopting term similarity techniques to select the candidate entailed words and to exploit this methodology in some specific applications such as taxonomy induction and ontology population.

Acknowledgments

Claudio Giuliano is supported by the X-Media project (`http://www.x-media-project.org`), sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978. Alfio Gliozzo is supported by FIRB-Israel research project N. RBIN045PXH.

	P	R	Mode P	Mode R
WN BEST	9.95	9.95	15.28	15.28
WN OOT	29.70	29.35	40.57	40.57

Table 3: WordNet Baselines

	P	R	Mode P	Mode R
all	41.23	41.20	55.28	55.28

Table 4: OOT results for LSA ranking (IRST1-lsa)

	P	R	Mode P	Mode R
all	69.03	68.90	58.54	58.54

Table 5: OOT results for Syntagmatic ranking (IRST2-syn)

References

I. Dagan and O. Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. In *proceedings of the PASCAL Workshop on Learning Methods for Text Understanding and Mining*, Grenoble.

I. Dagan, O. Glickman, and B. Magnini. 2005. The pascal recognising textual entailment challenge. Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment.

I. Dagan, O. Glickman, A. Gliozzo, E. Marmorshtein, and C. Strapparava. 2006. Direct word sense matching for lexical substitution. In *Proceedings ACL-2006*, pages 449–456, Sydney, Australia, July.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*.

O. Glickman, I. Dagan, M. Keller, S. Bengio, and W. Daelemans. 2006a. Investigating lexical substitution scoring for subtitle generation tenth conference on computational natural language learning. In *Proceedings of CoNLL-2006*.

O. Glickman, E. Shnarch, and I. Dagan. 2006b. Lexical reference: a semantic matching subtask. In *proceedings of EMNLP 2006*.

A. Gliozzo. 2005. *Semantic Domains in Computational Linguistics*. Ph.D. thesis, ITC-irst/University of Trento.

B. Magnini, C. Strapparava, G. Pezzulo, and A. Gliozzo. 2002. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8(4):359–373.