Combining Distributional and Paradigmatic information in a lexical substitution task

Diego De Cao and Roberto Basili

University of Roma Tor Vergata, Roma, Italy
{decao,basili}@info.uniroma2.it

Abstract. This paper describes an unsupervised approach to the Lexical Substitution task as applied to the Evalita 2009 competition. The applied approach builds on an *extended Latent Semantic Analysis* model that combines distributional and paradigmatic evidence in a unified vector space. All the systems employed in the ART laboratory are based on two main steps: 1) selection of a set of candidate substitute words and 2) ranking of each candidate according to combination of similarity metrics in *extended LSA* spaces. Comparative validation is reported in this paper.

Keywords: Lexical Substitution, Unsupervised Learning, Distributional Lexical Semantics, Latent Semantic Analysis.

1 Introduction

Lexical Substitution (LexSub) aims to find alternatives that can substitute a given word (the target) in an input context. This task may have a significant impact on several applications, such as Question Answering, Lexical Acquisition [1], or Paraphrasing [2]. One of the key problems in lexical substitution is the selection of proper candidate substitutes. The selection of candidates for a LexSub task requires the availability of a knowledge base for computing the potentially useful substitutes. This makes the LexSub task challenging. In general LexSub is more complex than Word Sense Disambiguation (WSD) as the lexical substitute words do not come from a closed set of candidates, i.e. the synonyms. Generally candidates for lexical substitution may have more general or more specific meanings. Moreover, they do not even share any common meaning with the target word in several cases. So the complexity of determining a candidate word set (i.e. to choose the set of words with a related meaning) does not only derive from sense disambiguation but also in the identification of the proper level of generalization or specification.

In this paper, we describe an unsupervised approach to the selection of the proper substitutes from a given candidate set in a LexSub task. It is based on a semantic similarity measure modeled in an *extended Latent Semantic Space* that combines distributional and paradigmatic information (such as synonymy in WordNet [3]).

The system developed for Evalita 2009 is mainly based on two steps. First, the list of candidate words as derived from a lexical resource (i.e. WordNet)

according to different involved relations (e.g. synonymy, hyperonymy) is built. The role of WordNet in this step implies a number of strong assumptions that will be better discussed in section 4. Then, the system ranks each candidate word according to a combination of semantic evidence as computed over different *extended LSA* spaces. In Section 2 these *extended LSA* models are described, while in Section 3 the system will be presented in deeper details. In Section 4 the results and conclusions are reported.

2 Extending distributional evidence through structured spaces

In standard, Latent Semantic Analysis (LSA) [4], the source term-by-document matrix M is factorized by the Singular Value Decomposition (SVD) into $M = USV^T$, where U is an orthonormal matrix of left singular vectors, S is a diagonal matrix of singular values, and V is an orthonormal matrix of right singular vectors. In line with [5], another way to obtain the SVD is to compute an eigenvalue decomposition of the 2-by-2 block matrix:

$$B = \begin{bmatrix} 0 & M \\ M^T & 0 \end{bmatrix} \tag{1}$$

The eigenvalues of B are the singular values of M and the left and right singular vectors are contained within the eigenvectors of this composite matrix B. Notice that the 0 matrices are of different orders in Eq. 1, as M has a dimension of $n \times m$ wherever n and m are the number of terms and documents, respectively. They express linear independence between term vectors and document vectors respectively. However, different assumptions can be done, as terms (and documents) may be done linearly dependent according to prior knowledge, e.g. synonymy information. Let D_1 and D_2 denote two symmetric matrices, e.g. expressing term and document similarities respectively. It is thus possible to extend the block matrix B by redefining it as:

$$B = \begin{bmatrix} D_1 & M \\ M^T & D_2 \end{bmatrix} \tag{2}$$

D_1 and D_2 can host further lexical information and express similarity models that modify the effects of Singular Value Decomposition. They in fact modify the resulting lower dimensional space: this approaches are elsewhere called *embeddings* ([6]). Matrix D_i may have a nature depending on the task, such as pairs of topically equivalent words in a specific domain or translation pairs in a multilingual lexicon (see [5] for an example). In this work we use D_1 to encode paradigmatic information, whereas the zeroes corresponding to pairs of synonym words are replaced by their *paradigmatic score*. This models synonymy and ambiguity of the involved words: two monosemic words in a synonymy relation receive a grater score than a pair in which one (or both) are polisemic. The *paradigmatic score* accounts for this as hereafter described.

Given a term t as found in a corpus and let $S_t \subset S$ be the subset of lexical senses of t in WordNet, for each synset $s \in S$ we compute $\sigma_{t,s}$ as:

$$\sigma_{t,s} = \begin{cases} \frac{1}{card(s)} \cdot log\left(\frac{|S|}{|S_t|}\right) & \textbf{if } s \in S_t \\ 0 & \textbf{otherwise} \end{cases} \quad (3)$$

where $card(s)$ represents the number of terms covered by the synset s.
Let ω_{ij} as *paradigmatic score* for two terms t_i and t_j, this is defined as:

$$\omega_{ij} = \sum_{s \in S} \sigma_{t_i,s} \cdot \sigma_{t_j,s} \quad (4)$$

Notice that in equation 4 the only valuable contribution is given by the senses $S_{t_i} \cap S_{t_j}$ shared between t_i and t_j, so that faster computation of ω_{ij} scores is allowed. Hereafter we will refer to the eigenvalue decomposition of matrix B in equation 2 as the *extended LSA* space.

3 The Lexical Substitution System

The proposed LexSub experimental platform proceeds through three steps: 1) the extraction of the lexical substitution sets for the target words, 2) the acquisition of domain models for candidates and 3) the ranking of candidate lexical substitutes over individual sentences according to the acquired domain models. A further step 4) back-off is included to deal with test cases for which the step 1) produces an empty candidate set.

1) Extract the Lexical Substitution set
At first a candidate lexical substitution set LS_t for all test cases of a target word t is created including possible substitution candidates for t. This is based on the WordNet sense hierarchy. We consider a term $t_{ls} \in LS_t$ as a candidate lexical substitute for t, if it satisfies one of these constraints: 1) t and t_{ls} are synonyms 2) t_{ls} is direct hypernym or hyponym of t 3) t and t_{ls} have a common hypernym. Hereafter we consider the *global lexical substitution set* LS as the union of all LS_t for all the target word t.

2) Domain Models acquisition
In order to account for contextual information about candidate terms, we consider three different vector spaces in this work. The first one is a global document space, where a term is described through the combination of documents in which it occurs. The second and the third models are window based spaces: every term t_{ls} is represented through its co-occurences in windows of a fixed size. A term is represented by separated matrices for the left and right co-occurrences in order to model a small amount of syntactic information.

The input of the *extended LSA* process for the global document space is a term by document matrix weighted according to the classical $tf \cdot idf$ score. The output of *extended LSA* process is a k dimensional space D, where k is the

dimensionality cut like in classical LSA approach. For the window based models, the *extended LSA* process takes in input a term by context matrix in which the context is represented by the ten words at left and at right, respectively. Also here the matrix is weighted according to the classical $tf \cdot idf$ score in which the context was seen as a document. As in the global space, the output of *extended LSA* process applied to context spaces is a k dimensional space (hereafter C_L and C_R respectively for the left and right context). Notice that the global space can be made independent from the target sentences, but the two contextual spaces are much more sparse[1].

3) Ranking candidate Lexical Substitution set

Given a target word t and the set of terms in the left and right contexts, C_L^t and C_R^t respectively, we extract the candidate LS_t for t as previously described in step 1. Let $\overrightarrow{t}$ be the vector representation of t in the global space D, we can compute the similarity $\gamma_{t,tc}$ as the standard *cosine similarity* between $\overrightarrow{t}$ and $\overrightarrow{tc}$, $\gamma_{t,tc} = sim(\overrightarrow{t}, \overrightarrow{tc})$. For both window based contexts, we can define two context vectors $\overrightarrow{c_L^t}$ and $\overrightarrow{c_R^t}$ as linear combinations of contextual terms, in C_L^t and C_R^t respectively. Let $\overrightarrow{tc_L}$ and $\overrightarrow{tc_R}$ be the vector representations of the candidate term tc into the two contextual spaces respectively. We define the left and right similarity through the cosine measure as $\chi_{t,tc}^L = sim(\overrightarrow{tc_L}, \overrightarrow{c_L^t})$ and $\chi_{t,tc}^R = sim(\overrightarrow{tc_R}, \overrightarrow{c_R^t})$.

For each candidate term tc, the final ranking score depends on the three scores $\gamma_{t,tc}$, $\chi_{t,tc}^L$ and $\chi_{t,tc}^R$ through a scaling factor $\phi(tc)$ defined by:

$$\phi(tc) = \frac{log(f(tc))}{max_{w \in LS_t}(log(f(w)))} \tag{5}$$

where $f(tc)$ and $f(w)$ are the number of occurrences of tc and w in the corpus. Different combinations are available, in particular we adopted the two following ones:

$$comb_{max}(t, tc) = \alpha_1 \max(\chi_{t,tc}^L, \chi_{t,tc}^R) + \alpha_2 \gamma_{t,tc} + \alpha_3 \phi(tc) \tag{6}$$

$$comb_{sum}(t, tc) = \alpha_1 (\chi_{t,tc}^L + \chi_{t,tc}^R) + \alpha_2 \gamma_{t,tc} + \alpha_3 \phi(tc) \tag{7}$$

where the three parameters are such that $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

4) Back-off model

Unfortunately in our system the coverage of the LS set depends strongly on the lexical resources. Some times may be that the LS_t of a target term is empty due the lack of coverage of the employed resources. To prevent this we implemented a distributional *back-off* model. This model works only when the LS_t for a target word t is empty. It simply finds the most similar words tc with the same Part-of-Speech tag of t according to the global space D. The similarity metric here is the standard *cosine similarity*, and the number of distributionally similar words considered for every test case is 100.

[1] In order to limit complexity in the experiments, we built window-based spaces only for terms $tc \in LS_t$, for some test t

4 Evaluation

In order to carry out the experiments for `Evalita 2009` we adopted two lexical resources: ItalWordNet [7] and MutiWordNet [8]. The first one has a larger coverage for verbs and nouns, while the second one better covers adjectives and adverbs. The domain corpus was made by about 50K articles extracted from Repubblica[2] with a dictionary of 200K terms. The dimensionality cut for all spaces in the *extended LSA* models was 100. We submitted four different systems, (as described in Table 1) resulting from different choices for the ranking model, the back-off model and the employed lexical resources for the different POS tags.

Table 1. Different systems according to the models and resources adopted.

	Models		Resources	
System	Noun & Verb	Adjective & Adverb	Noun & Verb	Adjective & Adverb
sys1	$comb_{max}$, back-off	$comb_{max}$, back-off	ItalWordNet	MultiWordNet
sys2	$comb_{sum}$, No_back-off	$comb_{sum}$, No_back-off	ItalWordNet	MultiWordNet
sys3	$comb_{max}$, back-off	$comb_{sum}$, No_back-off	ItalWordNet	MultiWordNet
sys4	$comb_{max}$, back-off	$comb_{max}$, back-off	MultiWordNet	MultiWordNet

There are basically two scoring methodologies[3]: (i) *BEST*, which scores the best substitute for a given target term, and (ii) *OOT*, which scores for the best 10 substitutes for a given target term, and systems do not benefit from providing less responses. *Mode precision* and *mode recall* calculate precision and recall, respectively, against the synonym chosen by the majority of annotators (if there is a majority). In Tables 2 and 3 the results of different POS tags for the *OOT* and *Best* scores are reported. As shown, our model works slightly better for verbs and nouns with respect to adjectives and adverbs considering the *OOT* metric. This is due mainly to the poor coverage of adjectives and adverbs of the employed lexical resources. The *Best* metric achieves high values for adverbs, as for the lower cardinality of *LS*, that increases the accuracy at the first hit.

The main problems for our system are related to the weak coverage of the lexical substitution sets. The *LS* set covers only the 55% of the test instances. In Table 4 we computed the scores narrowing the test cases at only those sentences that have at least one entry into our candidate Lexical Substitution set.

As shown, the system reach an accuracy comparable with the best system at `Evalita 2009`. The presented models are totally unsupervised, as they do not require neither the availability of sense tagged data, nor they consider WordNet information for building the *extended LSA* space models. Although we introduced a back-off model to improve coverage, it seems also to add too much noise.

[2] http://www.repubblica.it

[3] The scoring measures are fully described in the document at http://evalita.fbk.eu/doc/Guidelines_evalita09_lexical_substitution.pdf

Table 2. OOT scores achieved for different POS tags

		Verb		Noun		Adjective		Adverb		Global	
		P	R	P	R	P	R	P	R	P	R
sys1	std	**25.47**	**25.47**	**22.67**	**22.67**	11.11	**11.11**	19.98	19.98	20.09	**20.09**
	mode	**35.46**	**35.46**	**28.14**	**28.14**	**15.54**	**15.54**	30.3	**25.76**	**27.74**	**27.74**
sys2	std	25.04	25.04	21.73	21.73	**16.68**	9.96	**32.33**	15.63	23.00	18.72
	mode	**35.46**	**35.46**	27.14	27.14	13.99	13.99	**32.38**	**25.76**	26.32	26.32
sys3	std	**25.47**	**25.47**	**22.67**	**22.67**	**16.68**	9.96	**32.33**	15.63	**23.48**	19.11
	mode	**35.46**	**35.46**	**28.14**	**28.14**	13.99	13.99	**32.38**	**25.76**	26.58	26.58
sys4	std	17.43	16.72	17.21	16.75	**16.68**	9.96	**32.33**	15.63	18.62	14.78
	mode	22.71	22.71	20.6	20.6	13.99	13.99	**32.38**	**25.76**	20.52	20.52

Table 3. The BEST score achieved for different POS tags

		Verb		Noun		Adjective		Adverb		Global	
		P	R	P	R	P	R	P	R	P	R
sys1	std	**2.62**	**2.62**	**2.27**	**2.27**	3.39	3.39	5.41	**5.41**	3.16	3.16
	mode	**6.77**	**6.77**	4.52	4.52	**6.22**	**6.22**	12.12	**12.12**	**6.97**	**6.97**
sys2	std	2.58	2.58	2.17	2.17	**6.28**	**3.75**	**10.63**	5.14	3.90	3.17
	mode	**6.77**	**6.77**	**5.03**	**5.03**	**6.22**	**6.22**	**12.38**	9.85	6.71	6.71
sys3	std	**2.62**	**2.62**	**2.27**	**2.27**	**6.28**	**3.75**	**10.63**	5.14	**3.95**	**3.21**
	mode	**6.77**	**6.77**	4.52	4.52	**6.22**	**6.22**	**12.38**	9.85	6.58	6.58
sys4	std	1.87	1.79	1.72	1.67	**6.28**	**3.75**	**10.63**	5.14	3.52	2.80
	mode	3.19	3.19	3.02	3.02	**6.22**	**6.22**	**12.38**	9.85	5.03	5.03

Table 4. Scores achieved over sentences with at least one solution in LS

	OOT				Best			
	P	R	mode P	mode R	P	R	mode P	mode R
sys1	**45.41**	39.81	45.23	45.23	6.69	6.69	**11.62**	**11.62**
sys2	44.52	39.03	44.61	44.61	7.58	6.64	11.41	11.41
sys3	41.77	**41.77**	**46.68**	**46.68**	**7.67**	**6.72**	11.2	11.2
sys4	33.64	28.63	32.78	32.78	6.61	5.62	8.09	8.09

References

1. McCarthy, D.: Lexical substitution as a task for wsd evaluation. In: Proceedings of the ACL-02 workshop on Word sense disambiguatio. Morristown, NJ, USA (2002)
2. Dagan, I., Glickman, O., Gliozzo, A., Marmorshtein, E., Strapparava, C.: Direct word sense matching for lexical substitution. In: Proceedings of the 21st COLING and 44th Annual Meeting of ACL. Sydney, Australia (2006)
3. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller., K.: Introduction to wordnet: An on-line lexical database. International Journal of Lexicography, vol. 13, issue 4, pp. 235–312 (1990)
4. Landauer, T., Dumais, S.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Review, vol. 104, pp. 211–240 (1997)
5. Bader, B.W., Chew, P.A.: Enhancing multilingual latent semantic analysis with term alignment information. In: Proceedings of the 22nd International Conference on Computational Linguistics, pp. 49–56. Morristown, NJ, USA (2008)
6. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science, vol. 290, pp. 2323–2326 (2000)
7. Alonge, A., Bertagna, F., Calzolari, N., Roventini, A.: The Italian Wordnet, EuroWordNet Deliverable D032D033 part B5. Technical report (1999)
8. Pianta, E., Bentivogli, L., Girardi, C.: Multiwordnet: developing an aligned multilingual database. In: Proceedings of the First International Conference on Global WordNet. Mysore, India (2002)