The Lexical Substitution task at EVALITA 2009

Antonio Toral

Istituto di Linguistica Computazionale
Consiglio Nazionale delle Ricerche, Pisa, Italy
antonio.toral@ilc.cnr.it

Abstract. This paper describes the Italian Lexical Substitution task organised for EVALITA 2009. In this task, given a word in a specific context, the participant is asked to provide the synonyms which best fit in that context. The motivation behind the task, its objectives, the data prepared and distributed to participants, the baselines developed and the evaluation measures used are introduced. The results obtained both by the participating systems and by the baselines are presented. Finally, the different methodologies and resources used by the participants' systems and the results obtained by each of them are discussed.

Keywords: Lexical Substitution, Word Sense Disambiguation, Evaluation.

1 Motivation

Word Sense Disambiguation (WSD) is a fundamental step in the pursuit of Natural Language Understanding. Due to its important role, WSD has been present in relevant evaluation contests in recent years (e.g. Senseval/Semeval[1], EVALITA [1]). The evaluation of WSD has typically consisted of disambiguating the correct sense of words according to the senses present in computational lexicons (especially WordNet [2]). The main problem that arose is that the granularity of such resources is too detailed; while these fine distinctions might be useful for human users, they are not necessary for many computer applications [3].

An alternative way to evaluate WSD consists of performing Lexical Substitution [4]. In this case, given a word in a specific context, the participant is asked to provide the synonyms which best fit in that context. An important aspect of Lexical Substitution is the absence of a predefined sense inventory, thus allowing the participation of unsupervised approaches. For the present edition of EVALITA, a Lexical Substitution task has been organised. This is the first time that this type of task is evaluated for Italian and inspires in the task carried out for English at Semeval-2007[2] [5].

[1] http://www.senseval.org/
[2] http://nlp.cs.swarthmore.edu/semeval/tasks/task10/summary.shtml

2 Definition of the task

In this task participants are provided with a set of words, each of them appearing in different contexts, and are asked to return for each word synonyms that fit for each of the contexts in which the word appears. For example consider the following two contexts where the word "giudizio" appears:

- Alle nove, nel tribunale di Pescara, riprendeva il **processo** contro la banda Battestini.
  At nine o'clock, in the court of Pescara, the ***trial*** *against the Battestini gang resumed.*
- Ma la crisi dell'industria automobilistica francese deve avere convinto i vertici di Clermont-Ferrand della necessit di accelerare il **processo** di ristrutturazione.
  But the crisis of the French automobile industry should have convinced the top brass of Clermont-Ferrand of the need to accelerate the restructuring ***process****.*

In the first context, "processo" could be substituted by "giudizio" (trial) while in the second synonyms include "sviluppo" (development) and "procedura" (procedure).

3 Dataset

The words to be substituted were selected automatically from the Language Resources (LRs) ItalWordNet (IWN) [6] and PAROLE-SIMPLE-CLIPS (PSC) [7] by following several criteria that would guarantee, (i) that they have a high level of polysemy and (ii) that they are representative in the language. These criteria are:

- Words that belong to more than n semantic types (2 for nouns and verbs, 1 for adjectives) in the ontology of PSC.
- Words that have more than n hyponyms (5 for nouns, 2 for verbs) in PSC.
- Words that have more than 1 synonym in PSC.
- Base Concepts that represent at least 10 synsets automatically extracted from IWN [8] (only for nouns).

From the words selected we filtered out those that occur in less than ten sentences of the Italian Syntactic Semantic Treebank (ISST) [9]. As none of the criteria produced any adverb, we considered the adverbs that have more than 10 occurrences in the ISST. These criteria led to 231 words: 75 nouns, 58 adjectives, 63 verbs and 36 adverbs. We divided them into two groups: (i) 80 "manual", (ii) 151 "random".

For the first set we manually selected 10 sentences for each of the words from the ISST, whilst for the second 10 sentences for each word were automatically extracted. The resulting dataset is made up of 2,310 contexts. From these, 2,010 were annotated by three different annotators. The annotation guidelines

established by [5] were used; it was preferred to choose the synonyms without consulting any resource, but if the annotator could not find a synonym, she/he could use a dictionary, provided that it was not IWN or PSC as these are used for the baseline system.

The annotated contexts were split in two sets in order to carry out the evaluation: 300 were used as the trial set and the remaining 1,710 as the test set.

Let us consider two examples of contexts for one word and their annotations. The format of the corresponding files is described in the detailed guidelines of the task—footnotehttp://evalita.itc.it/doc/Guidelines_evalita09_lexical_substitution.pdf. This is the input file:

```
<lexelt item="processo.n">
...
<instance id="1925">
  <context>Ma la crisi dell' industria automobilistica francese e in
  particolare le difficolta' della Renault , accompagnata ai problemi
  internazionali di tutto il settore dell' auto , devono avere convinto
  i vertici di Clermont-Ferrand della necessita' di accelerare il <head>
  processo</head> di ristrutturazione . </context>
</instance>
<instance id="1926">
  <context>Alle nove , nel tribunale di Pescara , a poche centinaia di
  metri dal carcere , riprendeva il <head>processo</head> contro la banda
  Battestini e tutte le forze dell' ordine della citta' erano mobilitate
   attorno al palazzo di giustizia per garantirne la sicurezza . </context>
</instance>
...
</lexelt>
```

and these are the corresponding lines in the gold standard derived from the annotations:

```
processo.n 1925 :: sviluppo 1;sistema 1;procedura 1;
processo.n 1926 :: giudizio 1;
```

3.1 Baseline

Apart from the dataset, the organisation of the task has provided a baseline system. This system exploits the semantic relations present in IWN and PSC but does not perform any distinction regarding the context. Given a word, its substitutions are selected according to the following criteria:

- The synonyms and hyperonyms from all the senses of the LRs that correspond to the word are extracted and proposed as substitutions. Synonyms are given a weight value 3 and hyperonyms 1.
- If a word has been extracted more than once (e.g. from different senses and/or from both resources), its different weights are summed up.
- The list of substitutions is output in order according to the weights.

We provide three runs for each scoring type: one using IWN, one using PSC and finally, one using both LRs.

4 Evaluation measures

Participants' systems and the baselines have been evaluated according to two scoring types:

- Best. Scores the best guessed synonym.
- Out-of-ten (oot). Scores the best 10 guessed synonyms.

The evaluation measures used for both scoring types are precision, recall, F-measure, mode precision and mode recall. Mode precision and mode recall calculate precision and recall, respectively, against the synonym chosen by the majority of annotators (if there is a majority).

Prior to present the equations of the different evaluation measures consider the following variables:

- H, the set of annotators.
- T, the set of items with at least one answer from the annotators.
- h_i, the set of answers for an item $i \in T$ for an annotator $h \in H$
- m_i, the mode for an item i, i.e. the most frequent answer (if there is an answer more frequent than the others)
- TM the set of items for which there is an answer more frequent than the others.
- A (and AM), the set of items from T (or TM) where a system provides at least one synonym.
- $a_i : i \in A$ (or $a_i : i \in AM$), the set of guesses from a system for an item i.
- H_i, the multiset union for an item i for all $h \in H$.
- res, the unique types in H_i.
- $freq_{res}$ the associated frequency for each type in res (according to the number of types it appears in H_i).

The equations for the scoring type best are:

$$precision = \frac{\sum_{a_i : i \in A} \frac{\sum_{res \in a_i} \frac{freq_{res}}{|a_i|}}{|H_i|}}{|A|} \tag{1}$$

$$recall = \frac{\sum_{a_i : i \in T} \frac{\sum_{res \in a_i} \frac{freq_{res}}{|a_i|}}{|H_i|}}{|T|} \tag{2}$$

$$modeP = \frac{\sum_{bestguess_i \in AM} 1if bestguess = m_i}{|AM|} \tag{3}$$

$$modeR = \frac{\sum_{bestguess_i \in TM} 1if bestguess = m_i}{|TM|} \tag{4}$$

The equations for the scoring type oot are:

$$precision = \frac{\sum_{a_i:i \in A} \frac{\sum_{res \in a_i} freq_{res}}{|H_i|}}{|A|} \tag{5}$$

$$recall = \frac{\sum_{a_i:i \in T} \frac{\sum_{res \in a_i} freq_{res}}{|H_i|}}{|T|} \tag{6}$$

$$modeP = \frac{\sum_{a_i:i \in AM} 1if anyguess \in a_i = m_i}{|AM|} \tag{7}$$

$$modeR = \frac{\sum_{a_i:i \in TM} 1if anyguess \in a_i = m_i}{|TM|} \tag{8}$$

Finally, for both scoring types we calculate the F-measure as:

$$F = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{9}$$

5 Participation results

This section presents the results obtained by the participants' systems and the baselines provided for the two scoring types considered. Table 1 shows the results for the scoring type best and table 2 presents the scores for oot. Both tables are ordered in descending order according to the value of the F measure.

Table 1. Results obtained using the scoring type best

Run	Precision	Recall	F	mode P	mode R
uniba2	**8.16**	**7.18**	**7.64**	10.58	10.58
baroniCutugnoLenciPucci	6.26	6.01	6.13	**11.28**	**10.84**
uniba1	6.80	5.53	6.10	8.90	8.90
uniba3	6.28	5.46	5.84	8.13	8.13
decao3	3.95	3.21	3.54	6.58	6.58
decao2	3.90	3.17	3.50	6.71	6.71
decao1	3.16	3.16	3.16	6.97	6.97
decao4	3.52	2.80	3.12	5.03	5.03
baseline_psc	**10.86**	**9.06**	**9.88**	**13.94**	**13.94**
baseline_iwn_psc	9.71	8.19	8.89	13.16	13.16
baseline_iwn	2.72	1.78	2.15	2.19	2.19

Table 2. Results obtained using the scoring type oot

Run	Precision	Recall	F	mode P	mode R
uniba2	**41.46**	**36.50**	**38.82**	**47.23**	**47.23**
uniba1	37.74	30.69	33.85	34.84	34.84
uniba3	28.54	24.79	26.53	34.58	34.58
decao3	23.48	19.11	21.07	26.58	26.58
decao2	23.00	18.72	20.64	26.32	26.32
decao1	20.09	20.09	20.09	27.74	27.74
decao4	18.62	14.78	16.48	20.52	20.52
baroniCutugnoLenciPucci	16.65	16.00	16.32	24.97	24.00
baseline_iwn_psc	**27.52**	**23.23**	**25.19**	**37.24**	**32.39**
baseline_psc	23.00	19.20	20.93	26.97	26.97
baseline_iwn	14.51	9.51	11.49	12.77	12.77

6 Discussion

The task has been tackled by the participants using different methodologies:

- Basile used two different approaches: Basile_a (used for the run uniba1) uses WSD and extracts synonyms from the selected senses in the lexicon used as sense inventory; Basile_b (used for runs uniba2 and uniba3) builds n-grams of the contexts where the target word is substituted with synonyms from a lexicon, then he searches these contexts in a corpus and the substitutes are weighted according, among other factors, to the number of documents retrieved.
- Baroni et al propose to exploit co-occurrence statistics from a PoS-tagged corpus. They exploit Word Space Semantic Models (WSM) and represent the target word by a composite vector that takes into account both the overall distribution of the word in the corpus and its local context.
- De Cao et al. select a set of candidates from a lexicon and weight them according to a combination of similarity metrics in an extended Latent Semantic Analysis (LSA) space.

Table 3 shows the different techniques, tools, lexicons and corpora used by each of the systems that participated in the evaluation task.

The best F value obtained by a participant corresponds, for both the best (7.64) and oot (38.82) scoring types, to the uniba2 run. It is worth mentioning that the only system that does not use any LR to extract a set of candidate synonyms, Baroni et al., obtains the second best F measure (6.13) for the best scoring type. However, it ranks last in the oot type (F 16.32). Basile b and De Cao follow a similar two phase approach: in the first step a LR is used to extract a set of substitution candidates. In the second some method is applied to rank the candidates; Basile uses n-grams and IR while De Cao relies on LSA. The first approach obtains better results.

Table 3. Features of participants' systems

System	Techniques, tools	Lexicons	Corpora
Basile_a	WSD	ItalWordNet	
Basile_b	n-grams	ItalWordNet	ItWaC
	IR	De Mauro	
De Cao et al.	LSA	ItalWordNet	Repubblica
	SVD	MultiWordNet	
Baroni et al.	WSM	None	Repubblica
	PoS-tagger		ItWac
	Lemmatiser		Italian Wikipedia

Regarding the baselines, PSC obtained scores significantly higher than IWN. In fact, for the best score, this baseline obtains the best F score (9.88). As the baseline does not carry out any distinction of the different contexts, the fact that it beats all the systems for the best score type seems to indicate that there is a lot of room for the improvement of approaches that tackle Lexical Substitution in Italian. Given that the systems that used IWN obtained better scores than the baseline based on this LR and that the baseline using PSC beat these systems (at least for the best score type), it is hypothesised that these systems would obtain significantly better results by replacing IWN by PSC.

Finally, we compare the best results obtained for the current task with the best results obtained for the English task at Semeval 2007. Results are shown in table 4. The last column, *difference*, indicates the improvement percentage of the English score with respect to the Italian one.

Table 4. Comparison of best results in Italian and English

Scoring type	Measure	Italian	English	Difference
best	**P**	10.86	12.90	18.78
	R	9.06	12.90	42.38
	mode P	13.94	20.73	48.70
	mode R	13.94	20.73	48.70
oot	**P**	41.46	69.03	66.50
	R	36.50	68.90	88.80
	mode P	47.23	66.26	40.30
	mode R	47.23	66.26	40.30

As it can be seen, the English best score is higher for all the measures and scoring types. These results corroborate the fact that there is room for the improvement of Lexical Substitution in Italian.

Acknowledgements

I would like to thank the annotators: Stefania Bracale, Adriana Roventini and Giulia Sarti; the organisation of this task would not have been possible without their work. Thanks also to Monica Monachini for her help regarding the linguistic aspects of the task. Finally, thanks to Diana McCarthy and Roberto Navigli, the organisers of the Lexical Substitution task at Semeval 2007, for their help and advice regarding the set up of the current task.

References

1. Magnini, B., Cappelli, A., Tamburini, F., Bosco, C., Mazzei, A., Lombardo, V., Bertagna, F., Calzolari, N., Toral, A., Lenzi, V.B., Sprugnoli, R., Speranza, M.: Evaluation of natural language tools for italian: Evalita 2007. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Tapias, D. (eds.) Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), http://www.lrec-conf.org/proceedings/lrec2008. Marrakech, Morocco (2008)
2. Fellbaum, C.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)
3. Ide, N., Wilks, Y.: Making Sense About Sense. Springer (2006)
4. McCarthy, D.: Lexical substitution as a task for wsd evaluation. In: Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions, pp. 109–115. Association for Computational Linguistics, Philadelphia (2002)
5. McCarthy, D., Navigli, R.: The english lexical substitution task. Language Resources and Evaluation, vol. 43, issue 2 (2009)
6. Alonge, A., Bertagna, F., Calzolari, N., Roventini, A.: The Italian Wordnet, EuroWordNet Deliverable D032D033 part B5. Technical report (1999)
7. Ruimy, N., Monachini, M., Distante, R., Guazzini, E., Molino, S., Ulivieri, M., Calzolari, N., Zampolli, A.: Clips, a multi-level italian computational lexicon: A glimpse to data. In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02). Las Palmas de Gran Canaria, Spain (2002)
8. Izquierdo, R., Surez, A., Rigau, G.: Exploring the automatic selection of basic level concepts. In: Proceedings of RANLP 2007. Borovets, Bulgaria (2007)
9. Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M.T., Saracino, D., Zanzotto, F.M., Mana, N., Pianesi, F., Delmonte, R. Building the Italian syntactic-semantic treebank. Anne Abeill (ed.) Building and using Parsed Corpora, Language and Speech series, pp. 189–210. Kluwer, Dordrecht (2000)