Extending the gold standard for a lexical substitution task: is it worth it?

Ludovic Tanguy (1), Cécile Fabre (1) , Laura Rivière (2)
(1) CLLE-ERSS: CNRS & University of Toulouse, France
(2) University of Toulouse, France
ludovic.tanguy@univ-tlse2.fr, cecile.fabre@univ-tlse2.fr, laurariviere12@gmail.com

Abstract

We present a new evaluation scheme for the lexical substitution task. Following (McCarthy and Navigli, 2007) we conducted an annotation task that mixes two datasets: in the first one, 300 sentences containing a target word (among 30 different) were submitted to human subjects who were asked to provide substitutes. The second one contains the propositions of the systems that participated to the lexical substitution task based on the same data. The idea is first, to assess the capacity of the systems to provide good substitutes that would not have been proposed by the subjects and second, to measure the impact on the task evaluation of a new gold standard that incorporates these additional data. While (McCarthy and Navigli, 2009) have conducted a similar post-hoc analysis, re-evaluation of the systems' performances has not been carried out to our knowledge. This experiment shows interesting differences between the two resulting datasets and gives insight on how automatically retrieved substitutes can provide complementary data to a lexical production task, without however a major impact on the evaluation of the systems.

1. Introduction

The lexical substitution task consists in providing the best substitute for a given word in a given context, usually the sentence. This is a crucial task in Natural Language Processing that requires systems to identify words that are semantically close to the target, and to select among candidates those that best fit the context. The task has been introduced in *SemEval-2007* by (McCarthy and Navigli, 2007) and involves annotators that are asked to provide substitutes for a single target word in context. This task has been reproduced with variations on the language (Cholakov et al., 2014; Fabre et al., 2014) or the size of the dataset (Kremer et al., 2014).

In this paper, we estimate the reliability of this setup by performing an additional annotation task based on the results of the systems. We follow and extend (McCarthy and Navigli, 2009) who conducted a post hoc analysis to evaluate the reliability of the gold standard. The idea is to assess the capacity of the systems to provide good substitutes that would not have been proposed by the annotators. This question is particularly relevant for the lexical substitution task for two main reasons: first, finding a lexical substitute for a target word in the context of a sentence is a fairly difficult production task; second, systems that rely on huge corpora and lexical resources to detect semantic equivalents are likely to provide supplementary candidates that may be mistakenly ruled out by the evaluation procedure. Following (McCarthy and Navigli, 2009) we created a new annotation task mixing man-made and automatically-retrieved substitutes. But whereas (McCarthy and Navigli, 2009) simply evaluated the discrepancy between the two datasets, we go a step further and use this new source of data to evaluate the performance of the systems. This experiment shows interesting differences between the two resulting datasets. Yet it enables us to conclude that these differences have little impact on the evaluation of the systems.

2. The SemDis campaign: lexical substitution task in French

The SemDis 2014 evaluation campaign (Fabre et al., 2014), organized jointly by CLLE and IRIT research laboratories (University of Toulouse, France) was dedicated to a lexical substitution task in French, adapting the procedure defined for English by (McCarthy and Navigli, 2007). A dataset has been designed to evaluate and rank the systems that participated in the task, consisting of:

- 30 target words: 10 adjectives, 10 nouns and 10 verbs, whose frequency and degree of polysemy have been controlled;
- 300 sentences: 10 for each target word, selected from FRWaC (Baroni et al., 2009) and exemplifying different senses of the words;
- a gold standard made up of the substitutes that have been proposed by 7 judges for each sentence. Each judge could provide up to 3 answers per sentence.

For example, one of the target words is the noun *espace* (space), and one of the 10 target sentences for this word is: *No 208: Les sièges sont plus étroits, il y a moins d'espace entre les rangées.* (The seats are narrower, there is less space between the rows.)
Substitutes proposed by the judges are: *distance* (distance), *place* (room), *écart* (gap), *espacement* (spacing), *volume* (volume).
Each system competing for this task could propose up to 10 substitutes per sentence. The first substitute in each list was considered the best candidate, but those that come next were not considered in any particular order (see below for the details on evaluation measures).
This gold standard was used to evaluate the 9 participating systems and a simple baseline (which proposed the synonyms of the target words found in a dictionary, ordered by decreasing frequency in a reference corpus). It is freely available for further use by the community[1].

[1] http://redac.univ-tlse2.fr/datasets/semdis-gold/lexicalsubstitution/

3. Two stages of annotation

3.1. First version: annotation prior to runs

A total of 1,771 substitutes[2] were proposed by at least one judge (see (Fabre et al., 2014) for the details on filtering and normalizing the substitutes). Inter-annotator agreement was measured with two techniques:

- *pairwise inter-annotator agreement*: 25.8%. This is the average rate of similar answers (proposed/not proposed) over every substitute in the dataset and for each pair of annotators;
- *mode inter-annotator agreement*: 73%. This is the rate of propositions which are the most common substitute (calculated only for the 77% of items for which a mode exists).

The base score for each substitute for this dataset was the number of different judges who proposed it ($1 \leq score_1 \leq 7$). For example, the values for the substitutes of sentence 208 shown above are: *distance* (4), *place* (4), *écart* (2), *espacement* (2), *volume* (1).

3.2. Second version: post-hoc annotation

It is common practice in Information Retrieval to use the results of the systems to build the set of documents that will be submitted to the annotators. The pooling method (Teufel, 2007) is a solution to the problem of non exhaustive relevance judgments. Following (McCarthy and Navigli, 2009), we adapted this principle to the task, by complementing the initial annotation dataset with all the substitutes provided by the systems: for each target sentence, every substitute either proposed by one of the participant systems, the baseline, or found in the first gold standard dataset has been evaluated. This gave a total of 13,089 candidate substitutes, among which 983 (7.5%) were in the gold standard and had been proposed by at least one system; 788 (6.0%) were in the gold standard only. The bulk (86.4%) of the candidate substitutes were proposed only by the systems, and thus were not taken into account in the evaluation of the systems.

The 13,089 candidate substitutes were evaluated by 3 to 7 different anonymous judges (with an average of 4.2 judges per item). This annotation process was spanned over several months using an online survey platform. The judges were contacted by word of mouth and had the sole constraint of being native French speakers; we did not control other variables (such as age and education). Each judge could participate by annotating any number of the 90 subsets in which these substitutes were dispatched in order to reduce the time spent on a single session. Dispatching the substitutes was pseudo-random, as we made sure that each subset for a given sentence contained at least one item from the initial gold standard, to prevent cases where the judge was presented with only unsuitable substitutes. Evaluation itself required the judge to choose a value on a scale ranging from 0 (inappropriate substitute) to 3 (perfect substitute). Annotators were instructed to focus on the meaning, and to be permissive of slight agrammaticalities induced by the substitution (inflection, elision, choice of preposition, etc.). Intermediate values were to be used for substitutes that were acceptable but induced a slight modification in the meaning of the sentence. Overall, 55,000 individual scoring decisions were made that could be exploited after filtering out inconsistent and incomplete answers.

The base score for each substitute in this dataset is the average score over the 3 to 7 judges who rated it ($0 \leq score_2 \leq 3$). In the end, a total of 6.034 substitutes received a positive score, the average score being 0.51. Median score is 0, as 7,055 substitutes received an (unanimous) score of 0.

If we go back to sentence no 208, in addition to the 5 substitutes from the gold standard, 47 other candidates were found in the systems' propositions. In the end, 22 words received a positive score for this sentence. Here is a subset (words absent from the first gold standard are in italics): distance (3), place (3), espacement (3), écart (2.75), *écartement* (2.5), *intervalle* (2.5), *éloignement* (2), *interstice* (2), *marge* (1), *surface* (1), volume (0.75), *étendue* (0.75), *ouverture* (0.75), *air* (0.5), [...], *zone* (0.25), *an* (0), *atelier* (0), *attribut* (0), [...], *blanc* (0), *centre* (0), [...], *interligne* (0), *jardin* (0), [...] *univers* (0), *visa* (0)

Inter-annotator agreement was measured differently from the first dataset, because the annotated value is now a scale rating. We used:

- *rate of unanimous decision*: 0.56. 7,289 substitutes (out of 13,089) were given the exact same score by all judges.
- *average standard deviation*: 0.38 ($\pm$ 0.01, 95% CI). Standard deviation is calculated independently on the judges' rate for each substitute.

Although the two tasks clearly differ in their scope and nature (lexical production *vs* acceptability rating), it seems that the relevance of the collected data is on a par with what is expected when annotating lexical semantic phenomenons. In the next section we have a closer look at the differences between the two datasets.

This second dataset is also freely available[1].

4. Comparison of the two datasets

4.1. Quantitative differences

The two datasets agree on most items, as indicated by a high Pearson correlation coefficient measured between the two base scores. On the smaller dataset (first version, N=1,771), $\rho = 0.42$. On the larger dataset (N=13,089), $\rho = 0.61$ (considering a $score_1$ of 0 for substitutes absent from the first gold standard). More detail of this high correlation is indicated in the boxplots in Figure 1, where it appears clearly that substitutes initially proposed by 2 or more judges systematically receive a very high score in the post-hoc annotation.

Yet we notice contrast: as expected, the second dataset is much larger than the first one. More precisely, it contains a total of 4,336 new substitutes, which are words with a positive score that did not appear in the first stage. Conversely,

[2]The initial dataset contained 2,152 substitutes. However, we removed the multi-word substitutes as all the systems submitted single-word candidates. We also removed the data related to the adjective *compris* (understood/included), as most target sentences contained occurrences of the past participle instead of the adjective.

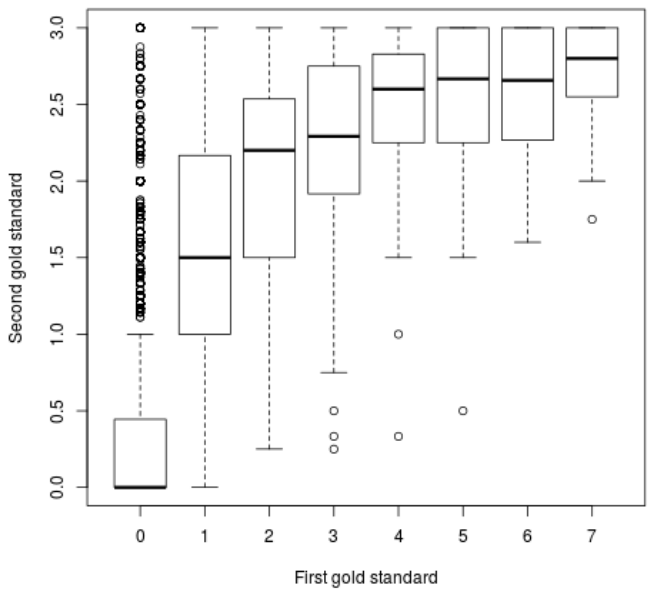

Figure 1: Correlation between the two annotations

73 substitutes from the initial gold standard (out of 1,117) received a null score with the post hoc evaluation (i.e. no judge considered they could be a possible substitute). In every such case, the substitute had an initial score of 1 (i.e. was proposed by only one judge in the first dataset).

In Table 1 is shown the breakdown of the candidate substitutes based on their origin, i.e. whether they were proposed by the judges, submitted by the competing systems (or baseline), or both. It appears clearly that the substitutes which were proposed both by the judges and the systems get the highest scores (more than 78% with a score of 2 or more), even higher than the substitutes which were proposed by the judges only.

$\mathbf{Score_2}$	Judges only (788)	Systems only (11,318)	Judges & systems (983)
$[0, 1[$	16%	80%	7%
$[1, 2[$	28%	14%	15%
$[2, 3[$	43%	6%	50%
$[3]$	13%	1%	28%

Table 1: Breakdown of substitutes across annotations

We also measured the approximate frequency for each substitute, based on the FrWac corpus, and found that, if in the first dataset the scores are positively correlated to the logarithm of the frequency ($\rho = 0.12$, $p < 0.01$) this is not true for the second ($\rho = 0.01$, $p > 0.05$). In addition, it appears that the substitutes which were added in the second dataset have a significantly lower frequency than those present in the first one (Wilcoxon test, $p < 0.01$).

4.2. Qualitative differences

Sentence 208 presented in Section 2 shows a typical example of the results obtained with the second annotation: as expected, only a few substitutes receive a very high score, most of which appearing in the first gold standard, while there is a long tail of words that were unanimously rejected by the annotators (corresponding to the very low box on the left of Figure 1). Some of these rejected candidates share with the word target *space* the very broad idea of location (*jardin* (garden), *univers* (universe)), others are related to a different sense than the one used in the sentence (such as *interligne* (line space)), while some are totally unrelated (*visa* (visa)). This illustrates the difficulty of the systems to capture the right level of semantic proximity or to perform disambiguation.

We had a closer look at words which were never proposed in the first annotation but got a high score in the second one (i.e. the outliers of the 0 value box on the left of figure 1). We identified several patterns. First, in some sentences, good substitutes are too numerous to be all proposed by annotators who focus on the most obvious synonyms within the limit of 3 answers. As the systems provide additional solutions the second task enables subjects to rate a larger set of substitutes. For example, considering again sentence 208, the word *écartement*, which is a morphological variant of the substitute *écart*, appears as a new valid substitute. In contrast, the judges may find no answer for a given sentence because there is only one good substitute corresponding to a rare word (e.g. *diction* (diction) for *débit* (speech delivery)). The subjects are not able to produce this word, but they rate it highly. More generally, new substitutes often depart from common vocabulary: colloquial or formal words do not come to mind but they are considered acceptable in the rating task. This could be an explanation to the difference in frequencies measured across the two sets. Lastly, some new substitutes exhibit a looser semantic relation with the target word (e.g. hypernyms such as *progression* (progression) as a substitute for *montée* (climbing)).

These new substitutes are the most interesting contribution to this second stage: they illustrate the differences between the two tasks (producing substitutes or annotating candidate substitutes) and they are likely to have an impact on the evaluation of the systems.

5. Impact on the system rankings

In this last section we compare the overall ratings of each participant system on the two benchmarks.

5.1. Evaluation measures

Two scoring measures were initially used to assess the participants. In the formulae below, for sentence number i, G_i is the set of substitutes in the gold standard, P_i is the set of candidate substitutes proposed among which $best_i$ is the first one, and $score_i(c)$ is the score given to candidate c in the reference.

- BEST: considers only the first proposal. The score is based on the raw gold standard score of the proposed substitute.
  $best(i) = \frac{score_i(best_i)}{\sum_{a \in G_i} score_i(a)}$
- OOT (Out Of Ten): considers the 10 proposals without taking the order into consideration.
  $oot(i) = \frac{\sum_{a \in P_i} score_i(a)}{\sum_{a \in G_i} score_i(a)}$

As can be seen, for both measures the scores are scaled by the sum scores of all substitutes for the target word. However, there are important differences from one target sen-

System	BEST (Gold1)	Rank	BEST (Gold2)	Rank	Rank difference
Proxteam_JDM_Syn	0.29	1	0.48	1	0
Proxteam_AxeParaProx_JDM_Syn	0.20	3	0.37	2	-1
CEA_LIST-word_cos_sent	0.23	2	0.33	3	+1
Alpage_WoDiS	0.17	4	0.29	4	0
CEA_LIST-fredist_cos_sent	0.12	7	0.25	5	-2
CEA_LIST-isc_cos_w2	0.12	8	0.22	6	-2
Proxteam_LM	0.15	5	0.18	7	+2
CEA_LIST-isc_cos_sent	0.11	9	0.18	8	-1
Baseline	0.13	6	0.17	9	+3
CEA_LIST-isc_l2_sent	0.03	10	0.09	10	0

Table 2: Normalized BEST scores and ranks for all systems as evaluated on both versions of the gold standard

System	OOT (Gold1)	Rank	OOT (Gold2)	Rank	Rank difference
Proxteam_JDM_Syn	0.41	1	0.38	1	0
Proxteam_AxeParaProx_JDM_Syn	0.37	2	0.35	2	0
CEA_LIST-isc_cos_sent	0.29	4	0.33	3	-1
CEA_LIST-isc_cos_w2	0.29	5	0.33	4	-1
Baseline	0.33	3	0.28	5	+2
CEA_LIST-fredist_cos_sent	0.24	6	0.23	6	0
CEA_LIST-isc_l2_sent	0.23	8	0.23	7	-1
Proxteam_LM	0.23	9	0.22	8	-1
CEA_LIST-word_cos_sent	0.24	7	0.19	9	+2
Alpage_WoDiS	0.22	10	0.19	10	0

Table 3: Normalized OOT scores and ranks for all systems as evaluated on both versions of the gold standard

tence to another, the number of substitutes with a positive score in the second gold standard varying from 5 to 56. It is thus impossible for a system to get a high score for a sentence with a large (> 10) number of suitable substitutes. For further evaluation, we use a normalized version of both evaluation measures, by dividing the measure value with the maximum expected value (i.e. the score obtained by a perfect system that proposes the 10 best substitutes in decreasing order of their gold standard score). We thus get a score that can reach a value of 1, meaning that it either proposed the highest rated substitute (BEST) or the 10 highest rated substitutes (OOT).

5.2. Comparison

We computed the scores for each initial participant submission and the baseline using both benchmarks, in order to measure the impact of the second annotation on the results of the substitution task. For details on the competing systems, please refer to (Fabre et al., 2014), (Desalle et al., 2014), (Ferret, 2014) and (Gábor, 2014).

Table 2 (resp. 3) gives the normalized BEST (resp. OOT) scores for all submitted participants with their scores according to both versions of the gold standards. As can be seen, for both evaluation measures there are little changes in terms of system ranking. Systems at both ends of the score range remain the same. The cases of bigger variations (up to three ranks) occur when the initial scores were very close to each others (e.g. for BEST there were 5 systems in the 0.11-0.15 range), this explains the relative changes.

6. Conclusion

The work presented here originated in a legitimate interrogation from the participants to the SemDis 2014 lexical substitution task. There was a possibility that the candidate systems proposed better (more accurate or more varied) substitutes than the ones proposed by the judges. To answer this question, we had to evaluate all the proposed substitutes. It required a substantial annotation effort, given the sheer number of items and the need for a cross-evaluation as we wanted to get reliable data for such a difficult task. We finally opted for a pseudo-crowd-sourcing process in which all the judges were contacted individually.

The main result of this work is of course the second gold standard dataset itself. Much more extended than the first, we measured that it is at least as reliable. It is an important added value to the test set itself, and we hope that it can still be of use for further experiments on lexical semantic processing techniques for French.

The answer to the title question is twofold. As far as ranking the participants is concerned, the changes exist but are quite marginal: the best systems remain on top, and for the others the relative changes are mostly irrelevant. So the short answer is that this second annotation was not worth the effort, and we hope that this can be of use for future work on the development of such evaluation data.

However, the comparison of the two data sets gives us useful insights on the task itself, and helps us understand the gaps between the systems. By beginning to identify the main differences in terms of substitutes proposed by humans and NLP systems, we can complete the initial analysis proposed by (Tanguy et al., 2016) who found that there are also important differences in the difficulty encountered for specific target sentences.

7. Acknowledgements

We would like to thank the organizers of the SemDis 2014 task (N. Hathout, M. Ho-Dac, F. Morlane-Hondère, P. Muller, F. Sajous and T. Van de Cruys); the participants (Y. Chudy, Y. Desalle, O. Ferret, K. Gábor, B. Gaume, P. Magistry, E. Navarro) and of course all the annotators.

8. Bibliographical References

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The Wacky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Cholakov, K., Biemann, C., Eckle-Kohler, J., and Gurevych, I. (2014). Lexical substitution dataset for german. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 1406–1411, Reykjavik, Iceland.

Desalle, Y., Navarro, E., Chudy, Y., Magistry, P., and Gaume, B. (2014). BACANAL : Balades Aléatoires Courtes pour ANAlyses Lexicales : Application à la substitution lexicale. In *Proceedings of the TALN Conference (Traitement Automatique du Langage Naturel), SemDis workshop*, Marseille, France.

Fabre, C., Hathout, N., Ho-Dac, L.-M., Morlane-Hondère, F., Muller, P., Sajous, F., Tanguy, L., and Van de Cruys, T. (2014). Présentation de l'atelier SemDis 2014 : sémantique distributionnelle pour la substitution lexicale et l'exploration de corpus spécialisés. In *Proceedings of the TALN Conference (Traitement Automatique du Langage Naturel), SemDis workshop*, Marseille, France.

Ferret, O. (2014). Utiliser un modèle neuronal générique pour la substitution lexicale. In *Proceedings of the TALN Conference (Traitement Automatique du Langage Naturel), SemDis workshop*, Marseille, France.

Gábor, K. (2014). Le système WoDiS - WOlf & DIStributions pour la substitution lexicale. In *Proceedings of the TALN Conference (Traitement Automatique du Langage Naturel), SemDis workshop*, Marseille, France.

Kremer, G., Erk, K., Padó, S., and Thater, S. (2014). What Substitutes tell us - analysis of an" all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics , EACL*, pages 540–549, Gothenburg, Sweden.

McCarthy, D. and Navigli, R. (2007). Semeval-2007 task 10: English Lexical Substitution Task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. Association for Computational Linguistics.

McCarthy, D. and Navigli, R. (2009). The English Lexical Substitution Task. *Language Resources and Evaluation*, 43(2):139–159.

Tanguy, L., Fabre, C., and Mercier, C. (2016). Analyse d'une tâche de substitution lexicale : quelles sont les sources de difficulté ? In *Proceedings of the TALN Conference (Traitement Automatique du Langage Naturel)*, Paris, France.

Teufel, S. (2007). An Overview of Evaluation Methods in TREC ad hoc Information Retrieval and TREC Question Answering. In L. Dybkjaer, et al., editors, *Evaluation of text and speech systems*, pages 163–186. Springer.