SemEval-2010 Task 2: Cross-Lingual Lexical Substitution

Rada Mihalcea
University of North Texas
rada@cs.unt.edu

Ravi Sinha
University of North Texas
ravisinha@unt.edu

Diana McCarthy
Lexical Computing Ltd.
diana@dianamccarthy.co.uk

Abstract

In this paper we describe the SemEval-2010 Cross-Lingual Lexical Substitution task, where given an English target word in context, participating systems had to find an alternative substitute word or phrase in Spanish. The task is based on the English Lexical Substitution task run at SemEval-2007. In this paper we provide background and motivation for the task, we describe the data annotation process and the scoring system, and present the results of the participating systems.

1 Introduction

In the Cross-Lingual Lexical Substitution task, annotators and systems had to find an alternative substitute word or phrase in Spanish for an English target word in context. The task is based on the English Lexical Substitution task run at SemEval-2007, where both target words and substitutes were in English.

An automatic system for cross-lingual lexical substitution would be useful for a number of applications. For instance, such a system could be used to assist human translators in their work, by providing a number of correct translations that the human translator can choose from. Similarly, the system could be used to assist language learners, by providing them with the interpretation of the unknown words in a text written in the language they are learning. Last but not least, the output of a cross-lingual lexical substitution system could be used as input to existing systems for cross-language information retrieval or automatic machine translation.

2 Motivation and Related Work

While there has been a lot of discussion on the relevant sense distinctions for monolingual WSD systems, for machine translation applications there is a consensus that the relevant sense distinctions are those that reflect different translations. One early and notable work was the SENSEVAL-2 Japanese Translation task (Kurohashi, 2001) that obtained alternative translation records of typical usages of a test word, also referred to as a *translation memory*. Systems could either select the most appropriate translation memory record for each instance and were scored against a gold-standard set of annotations, or they could provide a translation that was scored by translation experts after the results were submitted. In contrast to this work, in our task we provided actual translations for target instances in advance, rather than predetermine translations using lexicographers or rely on post-hoc evaluation, which does not permit evaluation of new systems after the competition.

Previous standalone WSD tasks based on parallel data have obtained distinct translations for senses as listed in a dictionary (Ng and Chan, 2007). In this way fine-grained senses with the same translations can be lumped together, however this does not fully allow for the fact that some senses for the same words may have some translations in common but also others that are not (Sinha et al., 2009).

In our task, we collected a dataset which allows instances of the same word to have some translations in common, while not necessitating a clustering of translations from a specific resource into senses (in comparison to Lefever and Hoste (2010)). [1] Resnik and Yarowsky (2000) also

[1] Though in that task note that it is possible for a translation to occur in more than one cluster. It will be interesting to

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 9–14,
Uppsala, Sweden, 15-16 July 2010.

conducted experiments using words in context, rather than a predefined sense-inventory however in these experiments the annotators were asked for a single preferred translation. In our case, we allowed annotators to supply as many translations as they felt were equally valid. This allows us to examine more subtle relationships between usages and to allow partial credit to systems that get a close approximation to the annotators' translations. Unlike a full blown machine translation task (Carpuat and Wu, 2007), annotators and systems are not required to translate the whole context but just the target word.

3 Background: The English Lexical Substitution Task

The English Lexical substitution task (hereafter referred to as LEXSUB) was run at SemEval-2007 (McCarthy and Navigli, 2007; McCarthy and Navigli, 2009). LEXSUB was proposed as a task which, while requiring contextual disambiguation, did not presuppose a specific sense inventory. In fact, it is quite possible to use alternative representations of meaning, such as those proposed by Schütze (1998) and Pantel and Lin (2002).

The motivation for a substitution task was that it would reflect capabilities that might be useful for natural language processing tasks such as paraphrasing and textual entailment, while not requiring a complete system that might mask system capabilities at a lexical level and make participation in the task difficult for small research teams.

The task required systems to produce a substitute word for a word in context. The data was collected for 201 words from open class parts-of-speech (PoS) (i.e. nouns, verbs, adjectives and adverbs). Words were selected that have more than one meaning with at least one near synonym. Ten sentences for each word were extracted from the English Internet Corpus (Sharoff, 2006). There were five annotators who annotated each target word as it occurred in the context of a sentence. The annotators were each allowed to provide up to three substitutes, though they could also provide a NIL response if they could not come up with a substitute. They had to indicate if the target word was an integral part of a multiword.

see the extent that this actually occurred in their data and the extent that the translations that our annotators provided might be clustered.

4 The Cross-Lingual Lexical Substitution Task

The Cross-Lingual Lexical Substitution task follows LEXSUB except that the annotations are translations rather than paraphrases. Given a target word in context, the task is to provide several correct translations for that word in a given language. We used English as the source language and Spanish as the target language.

We provided both development and test sets, but no training data. As for LEXSUB, any systems requiring training data had to obtain it from other sources. We included nouns, verbs, adjectives and adverbs in both development and test data. We used the same set of 30 development words as in LEXSUB, and a subset of 100 words from the LEXSUB test set, selected so that they exhibit a wide variety of substitutes. For each word, the same example sentences were used as in LEXSUB.

4.1 Annotation

We used four annotators for the task, all native Spanish speakers from Mexico, with a high level of proficiency in English. As in LEXSUB, the annotators were allowed to use any resources they wanted to, and were required to provide as many substitutes as they could think of.

The inter-tagger agreement (ITA) was calculated as pairwise agreement between sets of substitutes from annotators, as done in LEXSUB. The ITA without mode was determined as 0.2777, which is comparable with the ITA of 0.2775 determined for LEXSUB.

4.2 An Example

One significant outcome of this task is that there are not necessarily clear divisions between usages and senses because we do not use a predefined sense inventory, or restrict the annotations to distinctive translations. This means that there can be usages that overlap to different extents with each other but do not have identical translations. An example is the target adverb *severely*. Four sentences are shown in Figure 1 with the translations provided by one annotator marked in italics and {} braces. Here, all the token occurrences seem related to each other in that they share some translations, but not all. There are sentences like 1 and 2 that appear not to have anything in common. However 1, 3, and 4 seem to be partly related (they share *severamente*), and 2, 3, and 4 are also partly related (they share *seriamente*). When

we look again, sentences 1 and 2, though not directly related, both have translations in common with sentences 3 and 4.

4.3 Scoring

We adopted the **best** and **out-of-ten** precision and recall scores from LEXSUB (oot in the equations below). The systems were allowed to supply as many translations as they feel fit the context. The system translations are then given credit depending on the number of annotators that picked each translation. The credit is divided by the number of annotator responses for the item and since for the **best** score the credit for the system answers for an item is also divided by the number of answers the system provides, this allows more credit to be given to instances where there is less variation. For that reason, a system is better guessing the translation that is most frequent unless it really wants to hedge its bets. Thus if i is an item in the set of instances I, and T_i is the multiset of gold standard translations from the human annotators for i, and a system provides a set of answers S_i for i, then the **best** score for item i is[2]:

$$best\ score(i) = \frac{\sum_{s \in S_i} frequency(s \in T_i)}{|S_i| \cdot |T_i|} \quad (1)$$

Precision is calculated by summing the scores for each item and dividing by the number of items that the system attempted whereas recall divides the sum of scores for each item by $|I|$. Thus:

$$best\ precision = \frac{\sum_i best\ score(i)}{|i \in I : defined(S_i)|} \quad (2)$$

$$best\ recall = \frac{\sum_i best\ score(i)}{|I|} \quad (3)$$

The **out-of-ten** scorer allows up to ten system responses and does not divide the credit attributed to each answer by the number of system responses. This allows a system to be less cautious and for the fact that there is considerable variation on the task and there may be cases where systems select a perfectly good translation that the annotators had not thought of. By allowing up to ten translations in the **out-of-ten** task the systems can hedge their bets to find the translations that the annotators supplied.

[2]NB scores are multiplied by 100, though for **out-of-ten** this is not strictly a percentage.

$$oot\ score(i) = \frac{\sum_{s \in S_i} frequency(s \in T_i)}{|T_i|} \quad (4)$$

$$oot\ precision = \frac{\sum_i oot\ score(i)}{|i \in I : defined(S_i)|} \quad (5)$$

$$oot\ recall = \frac{\sum_i oot\ score(i)}{|I|} \quad (6)$$

We note that there was an issue that the original LEXSUB **out-of-ten** scorer allowed duplicates (McCarthy and Navigli, 2009). The effect of duplicates is that systems can get inflated scores because the credit for each item is not divided by the number of substitutes and because the frequency of each annotator response is used. McCarthy and Navigli (2009) describe this oversight, identify the systems that had included duplicates and explain the implications. For our task, we decided to continue to allow for duplicates, so that systems can boost their scores with duplicates on translations with higher probability.

For both the **best** and **out-of-ten** measures, we also report a *mode* score, which is calculated against the mode from the annotators responses as was done in LEXSUB. Unlike the LEXSUB task, we did not run a separate multi-word subtask and evaluation.

5 Baselines and Upper bound

To place results in perspective, several baselines as well as the upper bound were calculated.

5.1 Baselines

We calculated two baselines, one dictionary-based and one dictionary and corpus-based. The baselines were produced with the help of an online Spanish-English dictionary[3] and the Spanish Wikipedia. For the first baseline, denoted by DICT, for all target words, we collected all the Spanish translations provided by the dictionary, in the order returned on the online query page. The **best** baseline was produced by taking the first translation provided by the online dictionary, while the **out-of-ten** baseline was produced by taking the first 10 translations provided.

The second baseline, DICTCORP, also accounted for the frequency of the translations within a Spanish dictionary. All the translations

[3]www.spanishdict.com

1. Perhaps the effect of West Nile Virus is sufficient to extinguish endemic birds already **severely** stressed by habitat losses. {*fuertemente, severamente, duramente, exageradamente*}

2. She looked as **severely** as she could muster at Draco. {*rigurosamente, seriamente*}

3. A day before he was due to return to the United States Patton was **severely** injured in a road accident. {*seriamente, duramente, severamente*}

4. Use market tools to address environmental issues , such as eliminating subsidies for industries that **severely** harm the environment, like coal. {*peligrosamente, seriamente, severamente*}

5. This picture was **severely** damaged in the flood of 1913 and has rarely been seen until now. {*altamente, seriamente, exageradamente*}

Figure 1: Translations from one annotator for the adverb *severely*

provided by the online dictionary for a given target word were ranked according to their frequencies in the Spanish Wikipedia, producing the DICTCORP baseline.

5.2 Upper bound

The results for the **best** task reflect the inherent variability as less credit is given where annotators express differences. The theoretical upper bound for the **best** recall (and precision if all items are attempted) score is calculated as:

$$best_{ub} = \frac{\sum_{i \in I} \frac{freq_{most\ freq\ substitute_i}}{|T_i|}}{|I|} \times 100 = 40.57 \quad (7)$$

Note of course that this upper bound is theoretical and assumes a human could find the most frequent substitute selected by all annotators. Performance of annotators will undoubtedly be lower than the theoretical upper bound because of human variability on this task. Since we allow for duplicates, the **out-of-ten** upper bound assumes the most frequent word type in T_i is selected for all ten answers. Thus we would obtain ten times the **best** upper bound (equation 7).

$$oot_{ub} = \frac{\sum_{i \in I} \frac{freq_{most\ freq\ substitute_i \times 10}}{|T_i|}}{|I|} \times 100 = 405.78 \quad (8)$$

If we had not allowed duplicates then the **out-of-ten** upper bound would have been just less than 100% (99.97). This is calculated by assuming the top 10 most frequent responses from the annotators are picked in every case. There are only a couple of cases where there are more than 10 translations from the annotators.

6 Systems

Nine teams participated in the task, and several of them entered two systems. The systems used various resources, including bilingual dictionaries, parallel corpora such as Europarl or corpora built from Wikipedia, monolingual corpora such as Web1T or newswire collections, and translation software such as Moses, GIZA or Google. Some systems attempted to select the substitutes on the English side, using a lexical substitution framework or word sense disambiguation, whereas some systems made the selection on the Spanish side using lexical substitution in Spanish.

In the following, we briefly describe each participating system.

CU-SMT relies on a phrase-based statistical machine translation system, trained on the Europarl English-Spanish parallel corpora.

The UvT-v and UvT-g systems make use of k-nearest neighbour classifiers to build one word expert for each target word, and select translations on the basis of a GIZA alignment of the Europarl parallel corpus.

The UBA-T and UBA-W systems both use candidates from Google dictionary, SpanishDict.com and Babylon, which are then confirmed using parallel texts. UBA-T relies on the automatic translation of the source sentence using the Google Translation API, combined with several heuristics. The UBA-W system uses a parallel corpus automatically constructed from DBpedia.

SWAT-E and SWAT-S use a lexical substitution framework applied to either English or Spanish. The SWAT-E system first performs lexical sub-

stitution in English, and then each substitute is translated into Spanish. SWAT-S translates the source sentences into Spanish, identifies the Spanish word corresponding to the target word, and then it performs lexical substitution in Spanish.

TYO uses an English monolingual substitution module, and then it translates the substitution candidates into Spanish using the Freedict and the Google English-Spanish dictionary.

FCC-LS uses the probability of a word to be translated into a candidate based on estimates obtained from the GIZA alignment of the Europarl corpus. These translations are subsequently filtered to include only those that appear in a translation of the target word using Google translate.

WLVUSP determines candidates using the best N translations of the test sentences obtained with the Moses system, which are further filtered using an English-Spanish dictionary. USPWLV uses candidates from an alignment of Europarl, which are then selected using various features and a classifier tuned on the development data.

IRST-1 generates the **best** substitute using a PoS constrained alignment of Moses translations of the source sentences, with a back-off to a bilingual dictionary. For **out-of-ten**, dictionary translations are filtered using the LSA similarity between candidates and the sentence translation into Spanish. IRSTbs is intended as a baseline, and it uses only the PoS constrained Moses translation for **best**, and the dictionary translations for **out-of-ten**.

ColEur and ColSlm use a supervised word sense disambiguation algorithm to distinguish between senses in the English source sentences. Translations are then assigned by using GIZA alignments from a parallel corpus, collected for the word senses of interest.

7 Results

Tables 1 and 2 show the precision P and recall R for the **best** and **out-of-ten** tasks respectively, for normal and mode. The rows are ordered by R. The **out-of-ten** systems were allowed to provide up to 10 substitutes and did not have any advantage by providing less. Since duplicates were allowed so that a system can put more emphasis on items it is more confident of, this means that **out-of-ten** R and P scores might exceed 100% because the credit for each of the human answers is used for each of the duplicates (McCarthy and Navigli, 2009). Duplicates will not help the mode scores, and can be detrimental as valuable guesses which would not be penalised are taken up with

Systems	R	P	$Mode\ R$	$Mode\ P$
UBA-T	27.15	27.15	57.20	57.20
USPWLV	26.81	26.81	58.85	58.85
ColSlm	25.99	27.59	56.24	59.16
WLVUSP	25.27	25.27	52.81	52.81
SWAT-E	21.46	21.46	43.21	43.21
UvT-v	21.09	21.09	43.76	43.76
CU-SMT	20.56	21.62	44.58	45.01
UBA-W	19.68	19.68	39.09	39.09
UvT-g	19.59	19.59	41.02	41.02
SWAT-S	18.87	18.87	36.63	36.63
ColEur	18.15	19.47	37.72	40.03
IRST-1	15.38	22.16	33.47	45.95
IRSTbs	13.21	22.51	28.26	45.27
TYO	8.39	8.62	14.95	15.31
DICT	24.34	24.34	50.34	50.34
DICTCORP	15.09	15.09	29.22	29.22

Table 1: **best** results

duplicates. In table 2, in the column marked dups, we display the number of test items for which at least one duplicate answer was provided. [4] Although systems were perfectly free to use duplicates, some may not have realised this. [5] Duplicates help when a system is fairly confident of a subset of its 10 answers.

We had anticipated a practical issue to come up with all participants, which is the issue of different character encodings, especially when using bilingual dictionaries from the Web. While we were counting on the participants to clean their files and provide us with clean characters only, we ended up with result files following different encodings (e.g, UTF-8, ANSI), some of them including diacritics, and some of them containing malformed characters. We were able to perform a basic cleaning of the files, and transform the diacritics into their diacriticless counterparts, however it was not possible to clean all the malformed characters without a significant manual effort that was not possible due to time constraints. As a result, a few of the participants ended up losing a few points because their translations, while being correct, contained an invalid, malformed character that was not recognized as correct by the scorer.

There is some variation in rank order of the systems depending on which measures are used. [6]

[4] Please note that any residual character encoding issues were not considered by the scorer and so the number of duplicates may be slightly higher than if diacritics/different encodings had been considered.

[5] Also, note that some systems did not supply 10 translations. Their scores would possibly have improved if they had done so.

[6] There is not a big difference between P and R because

Systems	R	P	$Mode\ R$	$Mode\ P$	dups
SWAT-E	174.59	174.59	66.94	66.94	968
SWAT-S	97.98	97.98	79.01	79.01	872
UvT-v	58.91	58.91	62.96	62.96	345
UvT-g	55.29	55.29	73.94	73.94	146
UBA-W	52.75	52.75	83.54	83.54	-
WLVUSP	48.48	48.48	77.91	77.91	64
UBA-T	47.99	47.99	81.07	81.07	-
USPWLV	47.60	47.60	79.84	79.84	30
Colslm	43.91	46.61	65.98	69.41	509
ColEur	41.72	44.77	67.35	71.47	125
TYO	34.54	35.46	58.02	59.16	-
IRST-1	31.48	33.14	55.42	58.30	-
FCC-LS	23.90	23.90	31.96	31.96	308
IRSTbs	8.33	29.74	19.89	64.44	-
DICT	44.04	44.04	73.53	73.53	30
DICTCORP	42.65	42.65	71.60	71.60	-

Table 2: **out-of-ten** results

UBA-T has the highest ranking on R for **best**. USPWLV is best at finding the mode, for **best** however the UBA-W and UBA-T systems (particularly the former) both have exceptional performance for finding the mode in the **out-of-ten** task, though note that SWAT-S performs competitively given that its duplicate responses will reduce its chances on this metric. SWAT-E is the best system for **out-of-ten**, as several of the items that were emphasized through duplication were also correct.

The results are much higher than for LEXSUB (McCarthy and Navigli, 2007). There are several possible causes for this. It is perhaps easier for humans, and machines to come up with translations compared to paraphrases. Though the ITA figures are comparable on both tasks, our task contained only a subset of the data in LEXSUB and we specifically avoided data where the LEXSUB annotators had not been able to come up with a substitute or had labelled the instance as a name e.g. measurements such as *pound*, *yard* or terms such as *mad* in *mad cow disease*. Another reason for this difference may be that there are many parallel corpora available for training a system for this task whereas that was not the case for LEXSUB.

8 Conclusions

In this paper we described the SemEval-2010 cross-lingual lexical substitution task, including the motivation behind the task, the annotation process and the scoring system, as well as the participating systems. Nine different teams with a total of 15 different systems participated in the task, using a variety of resources and approaches. Comparative evaluations using different metrics helped determine what works well for the selection of cross-lingual lexical substitutes.

systems typically supplied answers for most items. However, IRST-1 and IRSTbs did considerably better on precision compared to recall since they did not cover all test items.

9 Acknowledgements

The work of the first and second authors has been partially supported by a National Science Foundation CAREER award #0747340. The work of the third author has been supported by a Royal Society UK Dorothy Hodgkin Fellowship. The authors are grateful to Samer Hassan for his help with the annotation interface.

References

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, pages 61–72, Prague, Czech Republic, June. Association for Computational Linguistics.

Sadao Kurohashi. 2001. SENSEVAL-2 japanese translation task. In *Proceedings of the* SENSEVAL-2 *workshop*, pages 37–44.

Els Lefever and Veronique Hoste. 2010. SemEval-2007 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, Uppsala, Sweden.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic.

Diana McCarthy and Roberto Navigli. 2009. The English lexical substitution task. *Language Resources and Evaluation Special Issue on Computational Semantic Analysis of Language: SemEval-2007 and Beyond*, 43(2):139–159.

Hwee Tou Ng and Yee Seng Chan. 2007. SemEval-2007 task 11: English lexical sample task via English-Chinese parallel text. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 54–58, Prague, Czech Republic.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Canada.

Philip Resnik and David Yarowsky. 2000. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(3):113–133.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Serge Sharoff. 2006. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.

Ravi Sinha, Diana McCarthy, and Rada Mihalcea. 2009. Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the NAACL-HLT Workshop SEW-2009 - Semantic Evaluations: Recent Achievements and Future Directions*, Boulder, Colorado, USA.