SemEval-2010 Task 2: Cross-Lingual Lexical Substitution

Ravi Sinha
University of North Texas
ravisinha@unt.edu

Diana McCarthy
University of Sussex
dianam@sussex.ac.uk

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

Abstract

In this paper we describe the SemEval-2010 Cross-Lingual Lexical Substitution task, which is based on the English Lexical Substitution task run at SemEval-2007. In the English version of the task, annotators and systems had to find an alternative substitute word or phrase for a target word in context. In this paper we propose a task where the target word and contexts will be in English, but the substitutes will be in Spanish. In this paper we provide background and motivation for the task and describe how the dataset will differ from a machine translation task and previous word sense disambiguation tasks based on parallel data. We describe the annotation process and how we anticipate scoring the system output. We finish with some ideas for participating systems.

1 Introduction

The Cross-Lingual Lexical Substitution task is based on the English Lexical Substitution task run at SemEval-2007. In the 2007 English Lexical Substitution Task, annotators and systems had to find an alternative substitute word or phrase for a target word in context. In this cross-lingual task the target word and contexts will be in English, but the substitutes will be in Spanish.

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 Background: The English Lexical Substitution Task

The English Lexical substitution task (hereafter referred to as LEXSUB) was run at SemEval-2007 following earlier ideas on a method of testing WSD systems without predetermining the inventory (McCarthy, 2002). The issue of which inventory is appropriate for the task has been a long standing issue for debate, and while there is hope that coarse-grained inventories will allow for increased system performance (Ide and Wilks, 2006) we do not yet know if these will make the distinctions that will most benefit practical systems (Stokoe, 2005) or reflect cognitive processes (Kilgarriff, 2006). 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 (Schütze, 1998; 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 only focusing on one aspect of the problem and therefore not requiring a complete system that might mask system capabilities at a lexical level and at the same time make

Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pages 76–81,
Boulder, Colorado, June 2009.

participation in the task difficult for small research teams.

The task required systems to produce a substitute word for a word in context. For example a substitute of *tournament* might be given for the second occurrence of *match* (shown in bold) in the following sentence:

The ideal preparation would be a light meal about 2-2 1/2 hours pre-match, followed by a warm-up hit and perhaps a top-up with extra fluid before the ***match****.*

In LEXSUB, 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.

A development and test dataset were provided, but no training data. Any system that relied on training data, such as sense annotated corpora, had to use resources available from other sources. The task had eight participating teams. Teams were allowed to submit up to two systems and there were a total of ten different systems. The scoring was conducted using recall and precision measures using:

- the frequency distribution of responses from the annotators and
- the mode of the annotators (the most frequent response).

The systems were scored using their **best** guess as well as an **out-of-ten** score which allowed up to 10 attempts. [1] The results are reported in McCarthy and Navigli (2007) and in more detail in McCarthy and Navigli (in press).

[1]The details are available at http://nlp.cs.swarthmore.edu/semeval/tasks/task10/task10documentation.pdf.

3 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, we propose to provide 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. An example from Resnik and Yarowsky (2000) (table 4 in that paper) is the first two senses from WordNet for the noun *interest*:

WordNet sense	Spanish Translation
monetary e.g. on loan	*interés, rédito*
stake/share	*interés,participación*

For WSD tasks, a decision can be made to lump senses with such overlap, or split them using the distinctive translation and then use the distinctive translations as a sense inventory. This sense inventory is then used to collect training from parallel data (Ng and Chan, 2007). We propose that it would be interesting to collect a dataset where the overlap in translations for an instance can remain and that this will depend on the token instance rather than mapping to a pre-defined sense inventory. Resnik and Yarowsky (2000) also conducted their experiments using words in context, rather than a predefined

sense-inventory as in (Ng and Chan, 2007; Chan and Ng, 2005), however in these experiments the annotators were asked for a single preferred translation. We intend to allow annotators to supply as many translations as they feel are equally valid. This will allow us to examine more subtle relationships between usages and to allow partial credit to systems which get a close approximation to the annotators' translations. Unlike a full blown machine translation task (Carpuat and Wu, 2007), annotators and systems will not be required to translate the whole context but just the target word.

4 The Cross-Lingual Lexical Substitution Task

Here we discuss our proposal for a Cross-Lingual Lexical Substitution task. The task will follow LEXSUB except that the annotations will be 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 will use English as the source language and Spanish as the target language. Multiwords are 'part and parcel' of natural language. For this reason, rather than try and filter multiwords, which is very hard to do without assuming a fixed inventory, [2] we will ask annotators to indicate where the target word is part of a multiword and what that multiword is. This way, we know what the substitute translation is replacing.

We will provide both development and test sets, but no training data. As for LEXSUB, any systems requiring data will need to obtain it from other sources. We will include nouns, verbs, adjectives and adverbs in both development and test data. Unlike LEXSUB, the annotators will be told the PoS of the current target word.

4.1 Annotation

We are going to use four annotators for our task, all native Spanish speakers from Mexico, with a high level of proficiency in English. The annotation interface is shown in figure 1. We will calculate inter-tagger agreement as pairwise agreement between sets of substitutes from annotators, as was done in LEXSUB.

4.2 An Example

One significant outcome of this task is that there will not necessarily be clear divisions between usages and senses because we do not use a predefined sense inventory, or restrict the annotations to distinctive translations. This will mean that there can be usages that overlap to different extents with each other but do not have identical translations. An example from our preliminary annotation trials is the target adverb *severely*. Four sentences are shown in figure 2 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 will adopt the **best** and **out-of-ten** precision and recall scores from LEXSUB. The systems can supply as many translations as they feel fit the context. The system translations will be given credit depending on the number of annotators that picked each translation. The credit will be 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 will be:

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

[2] The multiword inventories that do exist are far from complete.

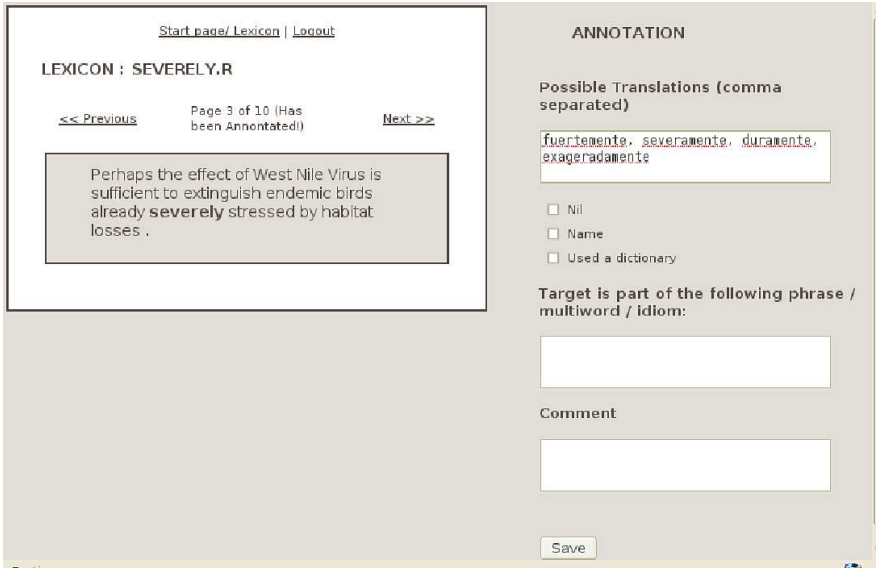

Figure 1: The Cross-Lingual Lexical Substitution Interface

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 2: Translations from one annotator for the adverb *severely*

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 will allow up to ten system responses and will not divide the credit attributed to each answer by the number of system responses. This allows the 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.

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

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

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

We will refine the scores before June 2009 when we will release the development data for this cross-lingual task. We note that there was an issue that the original LEXSUB **out-of-ten** scorer allowed duplicates (McCarthy and Navigli, in press). 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 (in press) describe this oversight, identify the systems that had included duplicates and explain the implications. For our task there is an option for the **out-of-ten** score. Either:

1. we remove duplicates before scoring or,
2. we allow duplicates so that systems can boost their scores with duplicates on translations with higher probability

We will probably allow duplicates but make this clear to participants.

We may calculate additional **best** and **out-of-ten** scores against the mode from the annotators responses as was done in LEXSUB, but we have not decided on this yet. We will not run a multiword task, but we will use the items identified as multiwords as an optional filter to the scoring i.e. to see how systems did without these items.

We will provide baselines and upper-bounds.

5 Systems

In the cross-lingual LEXSUB task, the systems will have to deal with two parts of the problem, namely:

1. candidate collection
2. candidate selection

The first sub-task, *candidate collection*, refers to consulting several resources and coming up with a list of potential translation candidates for each target word and part of speech. We do not provide any inventories, as with the original LEXSUB task, and thus leave this task of coming up with the most suitable translation list (in contrast to the synonym list required for LEXSUB) to the participants. As was observed with LEXSUB, it is our intuition that the quality of this translation list that the systems come up with will determine to a large extent how well the final performance of the system will be. Participants are free to use any ideas. However, a few possibilities might be to use parallel corpora, bilingual dictionaries, a translation engine that only translates the target word, or a machine translation system that translates the entire sentences. Several of the bilingual dictionaries or even other resources might be combined together to come up with a comprehensive translation candidate list, if that seems to improve performance.

The second phase, *candidate selection*, concerns fitting the translation candidates in context, and thus coming up with a ranking as to which translations are the most suitable for each instance. The highest ranking candidate will be the output for **best**, and the list of the top 10 ranking candidates will be the output for **out-of-ten**. Again, participants are free to use their creativity in this, while a range of possible algorithms might include using a machine translation system, using language models, word sense disambiguation models, semantic similarity-based techniques, graph-based models etc. Again, combinations of these might be used if they are feasible as far as time and space are concerned.

We anticipate a minor practical issue to come up with all participants, and that is the issue of different character encodings, especially when using bilingual dictionaries from the Web. This is directly related to the issue of dealing with characters with diacritics, and in our experience not all available software packages and programs are able to handle diacritics and different character encodings in the same way. This issue is inherent in all cross-lingual tasks, and we leave it up to the discretion of the participants to effectively deal with it.

6 Post Hoc Issues

In LEXSUB a post hoc evaluation was conducted using fresh annotators to ensure that the substitutes the systems came up with were not typically better than those of the original annotators. This was done as a sanity check because there was no fixed inventory for the task and there will be a lot of varia-

tion in the task and sometimes the systems might do better than the annotators. The post hoc evaluation demonstrated that the post hoc annotators typically preferred the substitutes provided by humans.

We have not yet determined whether we will run a post hoc evaluation because of the costs of doing this and the time constraints. Another option is to reannotate a portion of our data using a new set of annotators but restricting them to the translations supplied by the initial set of annotations and other translations from available resources. This would be worthwhile but it could be done at any stage when funds permit because we do not intend to supply a set of candidate translations to the annotators since we wish to evaluate candidate collection as well as candidate selection.

7 Conclusions

In this paper we have outlined the cross-lingual lexical substitution task to be run under the auspices of SemEval-2010. The task will require annotators and systems to find translations for a target word in context. Unlike machine translation tasks, the whole text is not translated and annotators are encouraged to supply as many translations as fit the context. Unlike previous WSD tasks based on parallel data, because we allow multiple translations and because we do not restrict translations to those that provide clear cut sense distinctions, we will be able to use the dataset collected to investigate more subtle representations of meaning.

8 Acknowledgements

The work of the first and third authors has been partially supported by a National Science Foundation CAREER award #0747340. The work of the second author has been supported by a Royal Society UK Dorothy Hodgkin Fellowship.

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.

Yee Seng Chan and Hwee Tou Ng. 2005. Word sense disambiguation with distribution estimation. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005)*, pages 1010–1015, Edinburgh, Scotland.

Nancy Ide and Yorick Wilks. 2006. Making sense about sense. In Eneko Agirre and Phil Edmonds, editors, *Word Sense Disambiguation, Algorithms and Applications*, pages 47–73. Springer.

Adam Kilgarriff. 2006. Word senses. In Eneko Agirre and Phil Edmonds, editors, *Word Sense Disambiguation, Algorithms and Applications*, pages 29–46. Springer.

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

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. in press. The english lexical substitution task. *Language Resources and Evaluation Special Issue on Computational Semantic Analysis of Language: SemEval-2007 and Beyond*.

Diana McCarthy. 2002. Lexical substitution as a task for wsd evaluation. In *Proceedings of the ACL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 109–115, Philadelphia, USA.

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.

Christopher Stokoe. 2005. Differentiating homonymy and polysemy in information retrieval. In *Proceedings of the joint conference on Human Language Technology and Empirical methods in Natural Language Processing*, pages 403–410, Vancouver, B.C., Canada.