Cross Lingual Lexical Substitution

Diana McCarthy, Ravi Sinha and Rada Mihalcea

September 8, 2009

1 Introduction

This document contains the information about scoring for the cross lingual lexical substitution (CLLS) task to be run at SEMEVAL-2010. It includes information on the format of the input files, the gold standard files and the format required for the system output files. It also contains the details needed for running the scorer and the measures used for evaluation.

There are two types of scoring. Systems can be evaluated on either or both of these scoring types:

best Scoring the best substitutes for a given item

oot Scoring for the best 10 substitutes for a given item. 10 responses are anticipated and systems will not benefit from providing less responses

The details of scoring for these types are described below in section 4. First we will describe the format of the input files, the gold standard files and the system output files.

2 Format

Please note that in this section we are using { } brackets to indicate variables in our textual description so that we can distinguish variables from xml tags. The variables used in our equations in section 4 will be indicated with symbols introduced in the text of section 4.

2.1 Input Format for Trial and Test Set: see trial dataset clls.trial.xml

The file input to systems for evaluation will adhere to the following format:

```
<corpus lang="english">
   <lexelt item="{lemma}.{pos}">
      <instance id="{id}">
         <context>...<head>...</head>...</context>
      </instance>

      ⋮

     <instance id="{id}">
         <context>...<head>...</head>...</context>
      </instance>
   </lexelt>

   ⋮

   <lexelt item="{lemma}.{pos}">
      <instance id="{id}">
         <context>...<head>...</head>...</context>
      </instance>

      ⋮

      <instance id="{id}">
         <context>...<head>...</head>...</context>
      </instance>
   </lexelt>
</corpus>
```

where each <lexelt> tag focuses on a specific lemma and part of speech, as specified in the attribute item. {pos} can assume one of the following four values: a, v, n, r (for adjective, verb, noun, adverb, respectively).

Each sentence is represented with an <instance> tag (each specifying a unique numeric id attribute). Each <instance> tag contains a <context> tag which includes the sentence in which an instance of the lemma co-occurs. The word instance is in turn enclosed in a <head> tag. For instance:

```
<corpus lang="english">

   ⋮

   <lexelt item="bright.a">
      <instance id="3">
         <context>The roses have grown out of control ,
```

```
        wild and carefree , their <head>bright</head>
        blooming faces turned to bathe in the early
        autumn sun .</context>
    </instance>
  </lexelt>

  ⋮

</corpus>
```

2.2 Gold Standard and System Output Formats

The gold-standard format will be the same for the **best** and **oot** evaluation. The system output files will differ for these two scoring methods.

Please note that all human responses are semi-automatically lemmatised so systems should ensure that all their answers are provided in lemmatised form. We will provide masculine forms where the appropriate inflection depends on the context. The reason for this is that we want scoring to focus on the correct choice of the lemma and not the surface form.

We are removing all diacritics from the gold standard as this will make it easier for systems and will account for the fact that annotators do not always provide them.

2.2.1 gold standard format for best and oot: see example file trial.gold

This file is provided by the task organisers. The format is

`{lexelt}\s{id}\s::\s{list of substitues with frequency}`

where the `\s` represents a single space and `lexelt` is the `lemma.pos` described in the section 2.1 above. Each item of the list or substitutes is separated by ; and consists of the lemmatised word of phrase and a frequency count indicating the number of annotators that provided this substitute.
Example:
bright.a 1 :: inteligente 3;brillante 2;listo 1;
bright.a 2 :: brillante 2;luminoso 2;claro 1;

2.2.2 system format for best: see example file trial.nowiki.best

The file output by systems for evaluation should confirm to the format:

`{lexelt}\s{id}\s::{list of substitutes}`

where the `\s` represents a single space and `lexelt` is the `lemma.pos` (or `lemma.ori_pos.pos`) described in the section 2.1 above. Each item of the list of substitutes is separated by ; and consists of the lemmatised word or phrase.

The best guess should appear first in the list. Example:
severely.r 127 :: severamente
tight.r 32 :: encoger
wild.n 160 :: habitat natural;libertad

The order of items in the file does not matter. In case the results file contains two or more lines for the same reference number for the same reference id, the first such line will be counted as the system's answer and the subsequent lines will be disregarded.

2.2.3 system format for oot: see example file trial.nowiki.oot

The file output by systems for evaluation should confirm to the format:

```
{lexelt}\s{id}\s:::\s{list of substitutes}
```

(N.B. three colons to differentiate from the **best** output file!)
where the `\s` represents a single space and `lexelt` is the `lemma.pos` described in the section 2.1 above. Each item of the list of substitutes is separated by ; and consists of the lemmatised word or phrase. Systems can provide up to 10 substitutes and will not have any advantage by providing less. Duplicates are allowed so a system may put more emphasis on items it is more confident of. NB This can give scores that might exceed 100% because the credit for each of the human answers ($freq_{res}$) is used for each of the duplicates (McCarthy and Navigli, 2009).

Example:
wild.a 151 ::: interesante;bravio;salvaje;extravagante;loco;fantastica;indomito;extrano;brutal;fiero;
manage.v 95 ::: manejar;dirigir;lograr;condicir;gobernar;guiar;administrar;mandar;coordinar;operar;

The order of items in the file does not matter. In case the results file contains two or more lines for the same reference number for the same reference id, the first such line will be counted as the system's answer and the subsequent lines will be disregarded.

3 Running the Scorer: score.pl

The scorer is a perl program. To run this do:

```
perl score.pl system_file gold_file  [-t best|oot] [-v]
```

where

required parameters	
system_file	is the file of formatted answers output by a system.
gold_file	is the file with the gold standard provided by human annotators.

optional parameters	
-t	specifies the type of scoring: best, oot (out of ten) with best as the default.
-v	causes line-by-line scoring calculations to be printed.

For example:

```
perl score.pl trial.nowiki.best clls.trial.gold
perl score.pl trial.nowiki.best clls.trial.gold -v
perl score.pl trial.nowiki.oot  clls.trial.gold -t oot
perl score.pl trial.nowiki.oot  clls.trial.gold -t oot -v
```

4 Details of the Evaluation Measures

We have 2 separate scoring functions to allow scoring on

1. any number of best guesses, with the very best guess (bg) first
2. up to 10 guesses (no penalising for multiple guesses to cope with fact that we only have 5 annotators and systems may come up with a larger, but equally valid, set of substitutes)

Let H be the set of annotators, T be the set of test items with 2 or more responses (non NIL or proper name) from the annotators and h_i be the set of responses for an item $i \in T$ for annotator $h \in H$.

For each $i \in T$ we will calculate the mode (m_i) which is the most frequent response, provided that there is a response more frequent than the others. The set of items where there is such a mode is referred to as T_m. Let A (and A_m) be the set of items from T (or T_m) where the system provides at least one substitute. Let $a_i : i \in A$ (or $a_i : i \in A_m$) be the set of guesses from the system for item i. For each i we calculate the multiset union (H_i) for all h_i for all $h \in H$ and for each unique type (res) in H_i will have an associated frequency ($freq_{res}$) for the number of times it appears in H_i. For example: Given an item (id 99) for ***happy.a*** supposing the annotators had supplied answers as follows:

annotator	responses
1	feliz extatico
2	feliz
3	contento feliz
4	extatico
5	alegre

and the system's responses for this item was *feliz; contento* then H_i would be { *feliz feliz feliz extatico extatico contento alegre* }. The res with associated frequencies would be *feliz 3 extatico 2 contento 1* and *alegre 1*.

4.1 Measures for best

This requires the **best** file produced by the system which gives as many guesses as the system believes are fitting, but where the credit for each correct guess will be divided by the number of guesses. The first guess in the list will be taken as the best (bg).

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

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

$$Mode\ precision = \frac{\sum_{bg_i \in A_m} 1\ if\ bg_i = m_i}{|A_m|} \tag{3}$$

$$Mode\ recall = \frac{\sum_{bg_i \in T_m} 1\ if\ bg_i = m_i}{|T_m|} \tag{4}$$

Using the example for *happy.a* id 99 in section 4, the credit for a_{99} in the numerator of precision and recall would be $\frac{3+1}{2 \cdot 7} = .286$

4.2 Measures for oot

This allows a system to make up to 10 guesses. The credit for each correct guess will not be divided by the number of guesses. There is no ordering of the guesses

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

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

$$Mode\ precision = \frac{\sum_{a_i:i \in A_m} 1\ if\ m_i \in a_i}{|A_m|} \quad (7)$$

$$Mode\ recall = \frac{\sum_{a_i:i \in T_m} 1\ if\ m_i \in a_i}{|T_m|} \quad (8)$$

5 Baselines

We have two baselines for both the **best** and **oot** tasks.

	files
best	i) trial.nowiki.best, ii) trial.wiki.best
oot	i) trial.nowiki.oot, ii) trial.wiki.oot

The baselines were produced with help from an online Spanish-English dictionary (www.spanishdict.com) and the Spanish Wikipedia. For all target English words, we collected all the Spanish translations provided by the dictionary. Especial care was taken that the part of speech of the translations collected stays the same as the part of speech of the target word. This online dictionary provides translations from two different resources, and we collected the translations in the order returned on the online query page. The file trial.nowiki.best was produced by taking the first translation provided by the Website, while the file trial.nowiki.oot was produced by taking the first 10 translations provided. In addition to this 'most-frequent-translation' heuristic, we also implemented a frequency-based most-frequent-translation heuristic based on the Spanish Wikipedia. All the translations provided by the online dictionary for a given target word were ranked according to their frequencies in the Spanish Wikipedia, and the files trial.wiki.best and trial.wiki.oot were produced analogously to the first two baseline files, although with different rankings than before. We did the same preprocessing on all the translations provided by the dictionary and also all the words in the Spanish Wikipedia - namely the diacritics were removed (accented characters were converted to regular characters), only lemmatized words (masculine gender, for example) were considered in counting frequencies, etc.

6 Measuring Human Agreement

We will measure pairwise agreement using the multisets. For each paired set of responses (h_i) from $i \in T$ for 2 annotators ($h \in H$) where both have provided a response, we calculate agreement as the multiset intersection divided by the multiset union. The sum of these pairwise scores is divided by the sum of all non nil paired annotator sets (h_i for all $h \in H$ and $i \in T$).

We will also calculate pairwise annotator agreement with the mode for each item. The details of these calculations are as reported for the original LEXSUB task in McCarthy and Navigli (2009).

7 Acknowledgements

This document is based on the SemEval 2007 English Lexical Substitution Task scoring documentation (*task10documentation.pdf* available at http://nlp.cs.swarthmore.edu/semeval/tasks/task10/task10documentation.pdf) which was written by Diana McCarthy and Roberto Navigli.

References

Diana McCarthy and Roberto Navigli. 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, 2009.