EVALITA 2009
Lexical Substitution Task
Guidelines for Participants

Antonio Toral
Istituto di Linguistica Computazionale
Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1 - 56124 Pisa, Italy
antonio.toral@ilc.cnr.it

1 Introduction

This document contains the guidelines for the participants of the Lexical Substitution task[1], part of the EVALITA 2009 evaluation campaign. In this task participants will be 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. This document covers the different aspects that are relevant to participants, i.e. (i) details about the different formats (input, gold standard, answer); (ii) the evaluation metrics that will be used and the scoring types; (iii) details about the scorer script.

2 Formats

2.1 Input

The input files that participants will have access to will adhere to the format of the DTD file "lexsub_input.dtd". The structure of such files is the following (we use brackets to indicate variables):

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

[1] http://evalita.fbk.eu/lexical.html

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

Each lexelt element corresponds to a lemma (with a specific part of speech) of the evaluation set. The variable pos can contain one of the following values: n, v, a, r (corresponding, respectively, to noun, verb, adjective and adverb). Each lexelt element contains a set of instance elements; each instance is identified by a unique id and contains one context element. A context element consists of a sentence in which a word corresponding to the lexelt lemma appears in an inner tag named head. Let's take a look at an example:

```
<corpus lang="it">
  [...]
  <lexelt item="fulmineo.a">
    <instance id="7">
      <context>E' successo due volte nel corso della <head>fulminea
      </head> offensiva</context>
    </instance>
    [...]
  </lexelt>
  [...]
</corpus>
```

2.2 Gold Standard

The gold standard files, hand tagged by annotators, follow this format:
`{lexelt}${id}$::${list of synonyms with their frecuency}`

Example:
`fulmineo.a$7$::$veloce 3;rapido 2; istantaneo 1`

This means that for the word "fulmineo" three annotators picked the synonym "veloce", two "rapido" and one "istantaneo".

2.3 Answer

The answer format is the format to which the output from the participant systems have to adhere.

```
{lexelt}${id}$::${list of guessed synonyms}
```

Example:

```
fulmineo.a$7$::$veloce;rapido
```

3 Evaluation Metrics

Systems will be evaluated on two scoring types:

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

Participants can submit results for any of the two scoring types or for both types. The evaluation measures that will be used for both scoring types are precision, recall, 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).

Taking the example data from 2.2 and 2.3 we would have $H_i =$ *veloce veloce veloce rapido rapido istantaneo* and res (with associated frequencies) = *veloce 3 rapido 2 istantaneo 1.*

The equations for the scoring type best are:

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

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

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

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

Taking the example data, the numerator for the precision equation would be $\frac{\frac{3+2}{2}}{6} = .4166$

The equations for the scoring type oot are:

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

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

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

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

4 Scorer

The scorer that will be used to evaluate the systems was originally developed for the Lexical Substitution task at Semeval 2007 and has been modified to meet the requirements of the current task. You can run it by typing:

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

where

- system_file is the output file from a system
- gold_file is the gold standard file tagged by the annotators
- -t specifies the type of scoring type that will be performed (best or oot)
- -v is optional and activates the verbose mode (causes line-by-line scoring calculations to be printed)