SemEval2017 - Task 2

Task overview

Given a pair of words, the task is to automatically measure their semantic similarity. The task comprises two subtasks: 1. Monolingual word similarity and 2. Cross-lingual word similarity. Unlike most existing word similarity datasets, the datasets include: Multi-word expressions; Domain-specific terms; Named entities.

For subtask 1. both words in a pair are in the same language. The subtask provides five monolingual word similarity datasets in English, German, Italian, Spanish and Farsi.

For subtask 2, the words in a pair are in different languages. This subtask is composed of ten cross-lingual word similarity datasets: EN-DE, EN-ES, EN-FA, EN-IT, DE-ES, DE-FA, DE-IT, ES-FA, ES-IT, and FA-IT.[1]

A system need not compete on all datasets.[2]

Input

The input file contains word pairs, one word pair per line. The words in a pair are separated by a tab.[3] Probably this will accomodate multi-word expressions - the words in a multi-word expression could be separated by whitespace characters other that tab and newline, perhaps by spaces.[4]

The language ISO codes for the five languages are DE-German, EN-English, ES-Spanish, FA-Farsi (Persian), IT-Italian.[5]

The monolingual datasets have a fixed length of 500 word pairs each; the cross-lingual datasets may vary in size, but is estimated at 600-1000 word pairs each.[6]

Output

There is one output file for each input dataset file. On each line of the

[1] http://alt.qcri.org/semeval2017/task2/index.php?id=task-details
[2] http://alt.qcri.org/semeval2017/task2/index.php?id=evaluation
[3] http://alt.qcri.org/semeval2017/task2/index.php?id=data-and-tools
[4] The sample input files which have been released do not contain any muti-word expressions.
[5] See the README.txt file inside the archive containing the trial data, available at http://alt.qcri.org/semeval2017/task2/data/uploads/semeval2017-task2.zip
[6] http://alt.qcri.org/semeval2017/task2/index.php?id=data-and-tools

output file, there must be the corresponding score of the word pair.

There should be no empty lines in the output file. The organizers recommend to set a score corresponding to the middle point of the scale, if the system does not cover a certain word in a pair.[7]

As regards the similarity scale, in one place is says that any consistent similarity scale can be used, for example [0-4], [-1-1], [0-1].[8] But in another place, it states that a [0-4] scale is used.[9] So it is probably a good idea to use the [0-4] scale, where 4 is “very similar” and 0 is “unrelated”.[10]

The floating point numbers in the output file seem to consist of a whole and a fractional part in base 10, separated by a period, for example 1.992[11]

Each team will only be allowed to submit a maximum of two systems / runs.[12]

Scoring and rankings

For each individual dataset, monolingual or cross-lingual, there will be a separate ranking. There will also be a global ranking for each subtask, as well as an additional ranking for corpus-based approaches.

For each individual ranking - in a certain dataset, the score is computed as the harmonic mean of Pearson and Spearman correlations on the corresponding dataset.

To participate in the global ranking for subtask 1, a system - multilingual or language-independent - must compete in at least 4 of the 5 datasets. The final score for the global ranking is computed as the mean of the 4 languages on which the system performs best.

For the global ranking for subtask 2, a system must compete in at least 6 of the 10 datasets, and the score for the global ranking is the mean of the 6 cross-lingual datasets on which the system performs best.[13]

For corpus-based approaches, there is an additional separate ranking. For systems that will not compete in this additional ranking also, there is no restriction on the training corpora that can be used. But to also participate in this additional ranking, corpus-based models must only use the following corpora for training:

- for subtask 1, the Wikipedia corpus corresponding to the given language (tokenized dumps available at https://sites.google.com/site/rmyeid/projects/polyglot);

- for subtask 2, the Europarl parallel corpus for any two languages that do

[7] http://alt.qcri.org/semeval2017/task2/index.php?id=evaluation
[8] http://alt.qcri.org/semeval2017/task2/index.php?id=evaluation
[9] http://alt.qcri.org/semeval2017/task2/index.php?id=data-and-tools
[10] For an overview of the rating scale, see http://alt.qcri.org/semeval2017/task2/index.php?id=data-and-tools
[11] README.txt
[12] http://alt.qcri.org/semeval2017/task2/index.php?id=evaluation
[13] http://alt.qcri.org/semeval2017/task2/index.php?id=evaluation

not include Farsi (http://opus.lingfil.uu.se/Europarl.php), and the OpenSubtitles2016 parallel corpora for any two languages one of which is Farsi (http://opus.lingfil.uu.se/OpenSubtitles2016.php).[14]

Calendar

Mon 01 Aug 2016: Trial data ready
Mon 09 Jan 2017: Evaluation start
Mon 30 Jan 2017: Evaluation end[15]
Mon 06 Feb 2017: Results posted
Mon 27 Feb 2017: Paper submissions due
Mon 03 Apr 2017: Author notifications
Mon 17 Apr 2017: Camera ready submissions due

[14] http://alt.qcri.org/semeval2017/task2/index.php?id=data-and-tools

[15] As regards the tuning of the system after the test data will have been released, on the one hand, there is a limit of two runs per team that can be submitted - this is to make it more difficult to tune the system to the test dataset. On the other hand, the evaluation period is three weeks long. This means that after an early first run, there is plenty of time to tune the system. But having received only the results from one run will give only limited insight, which will make it more difficult to tune the system.