IHLT Mandatory Project
Semantic Textual Similarity

Natural Language Research Group

Course 2018/19

1 SemEval 2012

SemEval (Semantic Evaluation Exercises) are a series of workshops which have the main aim of the evaluation and comparison of semantic analysis systems. The data and corpora provided by them have become a 'de facto' set of benchmarks for the NLP comunity.

The SemEval event provide data and evaluation frameworks for several tasks. One of them is Semantic Textual Similarity (STS), the purpose of this project. All information of 2012's edition is available at:

`https://www.cs.york.ac.uk/semeval-2012/`

All information of **6th task (Sematic Textual Similarity)** is available at:

`https://www.cs.york.ac.uk/semeval-2012/task6/index.html`

1.1 Paraphrases

STS is also known as paraphrases detection. A pair of texts is a paraphrase when both texts describe the same meaning with different words.

Real example extracted from trial data set of above task:

- `The bird is bathing in the sink.`
- `Birdie is washing itself in the water basin.`

1.2 Labels

When doing paraphrases detection on a pair of texts, a similarity value should be provided. The following table shows the meaning that this label must have in this task:

label	description
5	They are completely equivalent, as they mean the same thing.
4	They are mostly equivalent, but some unimportant details differ.
3	They are roughly equivalent, but some important information differs/missing.
2	They are not equivalent, but share some details.
1	They are not equivalent, but are on the same topic.
0	They are on different topics.

1.3 Data

All the data involved in the task is available at the data page:

```
https://www.cs.york.ac.uk/semeval-2012/task6/
    index.php%3Fid=data.html
```

It consist of four files:

- *trial*: includes the definition of the scores, a sample of 5 sentence pairs and the input and output formats. It is not needed, but it is useful for prototyping.
- *train*: training data from paraphrasing data sets, input and output formats.
- *test*: test data from paraphrasing data sets.
- *All system submissions*: submissions of the participants.

1.4 Evaluation Measure

In this task, *pearson correlation* is used for comparison purposes. It is available in python through the *scipy* module:

```
from scipy.stats import pearsonr

pearsonr(refs, tsts)[0]
```

1.5 2012 Results

At the following addresses, a summary of the results of the competition could be found:

```
https://www.cs.york.ac.uk/semeval-2012/task6/
    index.php%3Fid=results.html

https://www.cs.york.ac.uk/semeval-2012/task6/
    index.php%3Fid=results-update.html
```

And at the following ones the description of the event and the proceedings of the workshop:

`http://ixa2.si.ehu.es/starsem/proc/pdf/STARSEM-SEMEVAL051.pdf`

`http://ixa2.si.ehu.es/starsem/proc/program.semeval.html`

2 Statement

- Use data set and description of task *Semantic Textual Similarity* in *SemEval 2012*.
- Implement some approaches to detect paraphrase using sentence similarity metrics.
  - Explore some lexical dimensions.
  - Explore the syntactic dimension alone.
  - Explore the combination of both previous.
- Add new components at your choice (**optional**).
- Compare and comment the results achieved by these approaches among them and among the official results.
- Send files to `raco` in *IHLT STS Project* before the oral presentation:
  - Jupyter notebook: sts-[Student1]-[Student2].ipynb
  - Slides: sts-[Student1]-[Student2].pdf

2.1 Project Evaluation

The project evaluation will be:

$$Project\,Grade = 0.2 * Result + 0.2 * Presentation$$

where the *Result* will be constrained by rules in next table:

value	**constraint**
10	if the pearson is over 10th participant (.7562)
0	if the pearson is under the baseline (.311)
proportional	in other case