Introduction to Human Language Technologies

IHLT Mandatory Project

Semantic Textual Similarity

Gerard Escudero & Jordi Turmo

Natural Language Research Group

Master on Artificial Intelligence

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

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 bench- marks for the NLP comunity.

The *SemEval* event provides data and evaluation frameworks for several tasks. Task 6 is *Semantic Textual Similarity* (STS), the purpose of this project.

- The description of the event is available at:
  - http://ixa2.si.ehu.es/starsem/proc/pdf/STARSEM-SEMEVAL051.pdf
- and the proceedings of the workshop at:
  - http://ixa2.si.ehu.es/starsem/proc/program.semeval.html

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.`

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.

Data

- All the data consists 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.
- No other source data is allowed.

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]
```

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).
- Already generated word or sentence embeddings models are not allowed, such as BERT.
- 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`

Project Evaluation

The project evaluation will be:

```
ProjectGrade = 0.1  * Code Effectiveness +
               0.05 * Code Readability and Efficiency +
               0.05 * Use of NLP Libraries and Resources +
               0.4  * Analysis and Representation of Results +
               0.2  * Results +
               0.2  * Oral 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

Papers

- Task 6 results

Participants (>0.65 pearson):

- 1-ukp
- 2-takelab
- 5-unt
- 6-ets
- 10-sriubc
- 12-unitor
- 15-soft-cardinality
- 17-sheffield
- 18-umcc
- 20-weiwei
- 22-limsi
- 23-sbdlrhmn
- 24-sranjans
- 25-buap
- 27-penn
- 31-polyucomp
- 32-fbk