INESC-ID at ASSIN: measuring semantic similarity and recognizing textual entailment

Pedro Fialho
Universidade de Évora, INESC-ID
pedro.fialho@l2f.inesc-id.pt

Ricardo Marques
IST/UTL
ricardo.sa.marques@tecnico.ulisboa.pt

Bruno Martins
IST/UTL, INESC-ID
bruno.g.martins@tecnico.ulisboa.pt

Luísa Coheur
IST/UTL, INESC-ID
luisa.coheur@l2f.inesc-id.pt

Paulo Quaresma
Universidade de Évora, INESC-ID
pq@di.uevora.pt

Objective

- **What**: Detect textual entailment type (entailment, paraphrase or no relation) and measure semantic relatedness on a continuous scale from 1 to 5

- **How**: employ supervised machine learning on 96 language independent lexical features denoting relatedness among two sentences

- **Why**: multiple applications; whenever measuring the similarity between 2 texts is useful (dialogue systems, machine translation, etc)

Related work

- Popular in English, with datasets/rankings such as Microsoft Research Paraphrase Corpus or Stanford's SNLI

- Multiple approaches, both on Recognizing Textual Entailment (RTE) and paraphrase identification, using ML on lexical, syntactic and/or semantic features:
  - Madnani, Tetreault, and Chodorow (2012): paraphrase identification using string alignment metrics from Machine Translation (MT). Top 5 result on the Microsoft Research Paraphrase Corpus
  - Ul-Qayyum and Wasif (2012): lexical and semantic features for paraphrase identification
  - RTE based on lexical and:
    - syntactic (Pakray, Bandyopadhyay, and Gelbukh 2011)
    - sentence structure (Tsuchida & Ishikawa 2011)

INESC-ID@ASSIN

- Machine Learning (scikit-learn):
  - Semantic relatedness: **Kernel Ridge Regression**
  - RTE: **Support Vector Machines**

- 96 features, language independent (except for stopword removal and stemming), inspired on:
  - **Strings/sets**: Cosine, Jaccard, Edit distance, etc
  - **RTE**: Jaccard on selected words (named entities, negative words, etc)
  - **Paraphrases**: from text similarity (ROUGE, etc) and MT (Bleu, etc)
  - **Numeric**: Combine 2 Jaccard computations: on numbers and on their surrounding words
  - Each feature on multiple **text representations** (lowercase, character trigrams, Double Metaphone codes, etc):

INESC-ID@ASSIN

Feature	O	L	S	C	DM	T
LCS	X	X	X	X	X	
Edit Distance	X	X	X	X	X	
Cosine Similarity	X	X	X	X	X	X
Abs Length	X	X	X	X	X	
Max Length	X	X	X	X	X	
Min Length	X	X	X	X	X	
Jaccard	X	X	X	X	X	X
Soft TF-IDF	X	X	X			
NE Overlap	X	X	X	X	X	X
NEG Overlap	X	X	X	X	X	X
Modal Overlap	X	X	X	X	X	X
BLEU-3	X	X	X	X	X	
METEOR	X	X	X	X	X	
ROUGE N	X	X	X	X	X	
ROUGE L	X	X	X	X	X	
ROUGE S	X	X	X	X	X	
TER	X	X	X	X	X	
NCD	X	X	X	X	X	
Numeric	X	X	X			

Setup

- Training data:
  - ASSIN, Portuguese (PT-PT) or Brazilian (PT-BR): 3000 samples each
  - SICK (Marelli et al., 2014), translated in Bing: 9191 samples
- 3 combinations:

1. PT-PT or PT-BR: train only with the same Portuguese sample (European or Brazilian, respectively) of the test (3000 samples).
2. PT: merge datasets of both languages for training, regardless of the intended test (6000 samples).
3. PT+BingSICK: use the full Portuguese dataset and the translated SICK dataset for training (15191 samples, 9191 from SICK).

- Linear and Polynomial kernels

Results

- Portuguese test set
  - Polynomial
  - Linear

	Similarity		RTE	
Training	Pearson	MSE	Accuracy	F1
PT-PT	0.74	0.60	83.55%	0.68
PT	0.74	0.60	83.95%	0.69
PT+BingSICK	0.72	0.68	80.70%	0.59
PT-PT	0.73	0.62	84.90%	0.71
PT	0.74	0.61	84.05%	0.68
PT+BingSICK	0.70	0.73	77.10%	0.47

- Brazilian test set
  - Polynomial
  - Linear

PT-BR	0.73	0.36	85.45%	0.64
PT	0.73	0.36	85.70%	0.66
PT+BingSICK	0.70	0.40	84.30%	0.58
PT-BR	0.73	0.36	85.35%	0.55
PT	0.73	0.36	85.85%	0.66
PT+BingSICK	0.70	0.42	82.60%	0.46

Best features

- RTE
  - Soft TF-IDF, on original tokens
  - Jaccard, on Double Metaphone
  - Jaccard, on stemmed of lowercase tokens
  - Absolute Length, on Double Metaphone
  - LCS, on stemmed of lowercase tokens
  - Numeric, on original tokens
  - NE Overlap, on Double Metaphone
  - ROUGE-N, on original tokens
  - ROUGE-L, on stemmed of lowercase tokens
  - TER, on stemmed of lowercase tokens

Best features

- Semantic relatedness
  - Cosine Similarity, on original tokens
  - Soft TF-IDF, on original tokens
  - Jaccard, on Double Metaphone
  - Jaccard, on stemmed of lowercase tokens
  - Jaccard, on character trigrams
  - Numeric, on stemmed of lowercase tokens
  - NE Overlap, on Double Metaphone
  - ROUGE-N, on original tokens
  - ROUGE-N, on word clusters
  - ROUGE-S, on stemmed of lowercase tokens

Conclusions

- The configuration that most consistently yielded the best results is PT, both for RTE and similarity grading

- Our system performs better on Brazilian inputs

- Understanding a Portuguese variety while only knowing the other is better than using the Bing translated SICK dataset

	Similarity		RTE	
Training	Pearson	MSE	Accuracy	F1
PT-BR	0.73	0.63	82.70%	0.64
PT-PT	0.72	0.37	84.30%	0.66

Future work

- Same ML algorithms on syntactic (parse trees) and/or semantic features (word embeddings), isolated and combined with the current features

- Other ML: Neural Networks

- More knowledge sources, machine translated (SNLI: more language variability than SICK) and/or not (WordNet)