Paraphrase Recognition Using Machine Learning to Combine Similarity Measures

Prodromos Malakasiotis

Department of Informatics
Athens University of Economics and Business
Paraphrase Recognition

• Given a pair of phrases, sentences, or patterns \([S_1, S_2]\)
  decide if they are paraphrases, i.e., if they have
  (almost) the same meaning.
  – “X is the writer of Y” ≈ “X wrote Y” ≈ “X is the author
    of Y”

• Related to, but not the same as textual entailment.
  – “Athens is the capital of Greece” \(\models\) “Athens is
    located in Greece”, but not the reverse.

• Paraphrasing can be seen as bidirectional textual
  entailment.
Paraphrase recognition with Machine Learning

Training stage

Vector Creation

Preprocessing

(f1, f2, ..., fm, 1)1
(f1, f2, ..., fm, 0)2
...
(f1, f2, ..., fm, 1)n

Classifier

Trained Classifier

Classification stage

Vector Creation

Preprocessing

Classifiers

Trained Classifier

(f1, f2, ..., fm, ?)1
(f1, f2, ..., fm, ?)2

Experiments with 3 configurations:
- INIT, INIT+WN, INIT+WN+DEP

(AUEB, Department of Informatics)

http://nlp.cs.aueb.gr/
INIT Configuration

• The input pairs \([S_1, S_2]\) are represented as vectors of similarity scores measured on 4 forms of \([S_1, S_2]\):
  – (1) words, (2) stems, (3) POS-tags, (4) soundex codes

• 9 similarity measures, applied to the 4 forms:
  – Levenshtein (edit distance), Jaro-Winkler, Manhattan, Euclidean distance, cosine similarity, n-gram, matching coefficient, Dice, and Jaccard coefficient (see paper).
  – Similarities are measured in terms of tokens.
Partial Matching Features

While Bolton apparently fell and was immobilized, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

After the other inmate fell, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

• Find $S_1$’s (longer sentence) part that is most similar to $S_2$ (shorter sentence) using a sliding window:
  – At each step, calculate the average of 9 similarity scores.
  – Use the highest average (Avg) and the 9 scores it was computed from as additional features in INIT.
  – Do this for words, stems, POS-tags, and soundex codes.
While Bolton apparently fell and was immobilized, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

After the other inmate fell, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

- Find $S_1$’s (longer sentence) part that is most similar to $S_2$ (shorter sentence) using a sliding window:
  - At each step, calculate the average of 9 similarity scores.
- Use the highest average (Avg) and the 9 scores it was computed from as additional features in INIT.
- Do this for words, stems, POS-tags, and soundex codes.

\[ \text{Avg: 0.71} \]
Partial Matching Features

While Bolton apparently fell and was immobilized, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

After the other inmate fell, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

S1: While Bolton apparently fell and was immobilized, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.
S2: After the other inmate fell, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

Avg: 0.82

Find $S_1$’s (longer sentence) part that is most similar to $S_2$ (shorter sentence) using a sliding window:

- At each step, calculate the average of 9 similarity scores.

- Use the highest average (Avg) and the 9 scores it was computed from as additional features in INIT.

- Do this for words, stems, POS-tags, and soundex codes.
While Bolton apparently fell and was immobilized, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.

After the other inmate fell, Selenski used the mattress to scale a 10-foot, razor-wire fence, Fischi said.
INIT+WN Configuration

- The same as INIT, but:
  - It treats words from $S_1$ and $S_2$ that are synonyms (in WordNet) as identical.

Fewer than a dozen FBI agents were dispatched to secure and analyze evidence.

Fewer than a dozen FBI agents will be sent to Iraq to secure and analyze evidence of the bombing.

INIT’s Avg: 0.73

INIT+WN’s Avg: 0.78
INIT+WN+DEP Configuration

• Same as INIT+WN, but:
  – 3 additional features that measure dependency grammar similarity between $S_1$ and $S_2$:

\[
R_1 = \frac{|\text{common dependencies}|}{|S_1 \text{ dependencies}|}
\]

\[
R_2 = \frac{|\text{common dependencies}|}{|S_2 \text{ dependencies}|}
\]

\[
F_{R_1,R_2} = \frac{2 \cdot R_1 \cdot R_2}{R_1 + R_2}
\]
The dollar was at 116.92 yen against the yen, flat on the session, and at 1.2891 against the Swiss franc, also flat.

The dollar was at 116.78 yen JPY, virtually flat on the session, and at 1.2871 against the Swiss franc CHF, down 0.1 percent.

```
det(dollar-2, The-1)  
nsubj(flat-25, dollar-2)  
...  
dep(at-4, against-7)  
det(yen-9, the-8)  
...  
det(session-14, the-13)  
pobj(on-12, session-14)  
...  
```

```
det(dollar-2, The-1)  
nsubj(was-3, dollar-2)  
...  
det(session-14, the-13)  
prep_on(flat-11, session-14)  
...  
dep(at-17, against-19)  
det(CHF-23, the-20)  
...  
```

Avg = 0.72
R₁ = 0.14
R₂ = 0.16
F_{R₁,R₂} = 0.15
Last week the power station’s US owners, AES Corp, walked away from the plant after banks and bondholders refused to accept its financial restructuring offer.

The news comes after Drax's American owner, AES Corp. AES.N, last week walked away from the plant after banks and bondholders refused to accept its restructuring offer.

amod(week-2, Last-1)
tmod(walked-13, week-2)
...
trt(walked-13, away-14)
...
det(news-2, The-1)
...
amod(week-18, last-17)
dep(walked-19, week-18)
...
trt(walked-19, away-20)
...

Avg = 0.71
R₁ = 0.52
R₂ = 0.59
F_{R₁,R₂} = 0.55
Feature Selection

• Start with an empty feature set.
• Gradually add features:
  – Form new feature sets by adding one feature.
  – Measure the predictive power of the new sets.
  – Keep the best new feature set(s).
  – Tried both hill-climbing and beam-search.
• A lot of redundancy in the full feature set.
  – Feature selection leads to competitive results with much fewer features (10 instead of 136).
• But the full feature set leads to better results.
Experiments

• Microsoft Research (MSR) Paraphrase Corpus:
  – 5,801 pairs of sentences evaluated by judges.
  – 4,076 training pairs.
  – 1,725 testing pairs.

• Baseline (BASE):
  – Use a threshold on edit distance to decide if a pair is positive (paraphrases) or negative.
  – The threshold is tuned on the training pairs.
Results on MSR corpus

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASE</td>
<td>60,00%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INIT</td>
<td>75,00%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INIT + WN</td>
<td>85,00%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INIT + DEP</td>
<td>90,00%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://nlp.cs.aueb.gr/
INIT performs well too!

Accuracy  Precision  Recall  F-measure

INIT + WDN

60,00%  65,00%  70,00%  75,00%

80,00%  85,00%  90,00%

INIT + WDN

http://nlp.cs.aueb.gr/

AUEB Natural Language Processing Group
Conclusions

• INIT is competitive to the best known systems, using fewer resources.
  – Useful for languages where WordNet, reliable dependency parsers etc. are unavailable.

• INIT+WN and INIT+WN+DEP perform even better, but they require more resources and the improvement is small.
  – The differences may be small because of the high lexical overlap of the paraphrases in the MSR corpus.
Thank you!

Questions?