The MWE 2008 Shared Task: Ranking MWE Candidates

Stefan Evert
University of Osnabrück

stefan.evert@uos.de | purl.org/stefan.evert
MWE 2008 Shared Task

A first exploratory shared task, using some of the data sets contributed to MWE 2008

Focus on candidate ranking

- well-established task & evaluation
- essential step for MWE extraction

More sophisticated challenges to be tackled in future years

The MWE 2008 Shared Task: Ranking MWE Candidates (closed)

Submission: Short paper (4 pages) describing evaluation results (submission details)
Deadline: Wednesday, 19 March 2008 (extended)

Overview

This first, exploratory evaluation campaign focusses on the straightforward task of ranking MWE candidates, in order to enable a wide range of systems (any that include a ranking component) to participate in the shared task, bypass lengthy discussions on evaluation paradigms, and give teams more time to adapt their algorithms. As a consequence, only a few of the submitted resources were included in the shared task. We expect a full-fledged competition with a more varied range of tasks to be organised in the near future, making use of the other data sets from this workshop.

Participating teams have to apply their MWE candidate ranking algorithm to at least three different data sets, provided in the form of a list of MWE candidates that have been manually annotated as true positives (TP) and false positives (FP). Evaluation is carried out by the teams themselves in terms of n-best precision, precision-recall graphs and average precision. Further analysis and experiments are strongly encouraged (e.g. on different subsets of the candidates, or for different subtypes of MWE). Baseline results for standard association measures will be presented by the workshop organisers.

Teams planning to participate in the shared task should join the multword mailing list for up to date information and discussions, and send a brief expression of interest to the list.

Shared task participants will submit a short paper (4 pages) with the complete evaluation results of their ranking algorithm on at least three different data sets. Submissions are not anonymous and will be briefly reviewed by the workshop organisers for compliance with the shared task requirements (submission details). In addition, participants should make available their full candidate ranking and, if possible, an implementation of their ranking algorithm, to encourage further experiments. Details will be discussed on the multword...
The 4 subtasks

1. **English verb-particle combinations** (Baldwin, EN-VPC)
   - 3078 candidates, 440 TPs, baseline precision = 14.29%
   - TPs subdivided into transitive and intransitive VPC

2. **German PP-verb combinations** (Krenn, DE-PNV)
   - 5102 candidates (FR-30 subset), 566 TPs, baseline = 11.09%
   - TPs are figurative expressions and support-verb constructions

3. **German adjective-noun collocations** (Evert, DE-AN)
   - 1252 candidates, 520 TPs (cat. 1+2), baseline = 41.53%

4. **Czech dependency bigrams** (Pecina, CZ-MWE)
   - 12232 candidates, 2572 unanimous TPs, baseline = 21.03%
1. **English verb-particle combinations** (Baldwin, EN-VPC)
   - no frequency data provided
   - participants used BNC (fragment) and Web frequencies

2. **German PP-verb combinations** (Krenn, DE-PNV)
   - frequency data from *Frankfurter Rundschau* (FR) corpus provided
   - verb + nearest PP (based on TreeTagger & YAC chunking)

3. **German adjective-noun collocations** (Evert, DE-AN)
   - frequency data from FR corpus (TreeTagger & YAC chunking)

4. **Czech dependency bigrams** (Pecina, CZ-MWE)
   - frequency data from Prague Dependency Treebank provided
**Shared task participants**

- **Baseline** results (Evert)
  - all 4 subtasks, 6 standard association measures

- **Ramisch, Schreiner, Idiart & Villavicencio**
  - EN-VPC, DE-PNV, DE-AN
  - two standard association measures + permutation entropy

- **Pecina**
  - DE-PNV, DE-AN, CZ-MWE
  - 57 association measures + machine-learning techniques
  - experiments with different TP definitions and subsets
Evaluation methodology

- Algorithm produces ranking of given set of candidates (in the form of decreasing scores)

- Compute **precision** & **recall** (of TPs) for different n-best lists

- Can be visualised as a precision-recall graph

- Baseline precision: proportion of TPs in data set

- Overall quality measure: average precision
Evaluation methodology

- Algorithm produces ranking of given set of candidates (in the form of decreasing scores)
- Compute **precision** & **recall** (of TPs) for different n-best lists
- Can be visualised as a precision-recall graph
- Baseline precision: proportion of TPs in data set
- Overall quality measure: average precision

**Baseline precision** = 11.09%
Evaluation methodology

- Algorithm produces ranking of given set of candidates (in the form of decreasing scores)
- Compute **precision** & **recall** (of TPs) for different n-best lists
- Can be visualised as a precision-recall graph
- Baseline precision: proportion of TPs in data set
- Overall quality measure: average precision

Baseline precision = 11.09%

Average precision = 39.79%
Baseline results for 6 widely-used association measures

- log-likelihood ($G^2$)
- chi-squared ($X^2$) with Yates' correction
- t-score ($t$)
- Mutual Information (MI)
- Dice coefficient (Dice)
- frequency ranking ($f$)

NB: Ramisch et al.'s MI = $G^2$
baseline results: EN-VPC

- frequency data from full BNC (adjacent verb + particle)
- missing data are ranked at bottom
- baseline: 14.29%
- best AM: t-score (AP = 29.94%)
- frequency ranking: AP = 29.01%
Baseline results: DE-PNV

FR-30 subset with 5102 candidates

TPs = figurative expressions + SVC

baseline: 11.09%

best AM: t-score (AP = 39.79%)

frequency ranking: AP = 33.88%
Baseline results: DE-AN

- TPs = cat. 1+2
- baseline: 41.53%
- best AM: Dice (AP = 58.84%)
- frequency ranking: AP = 46.90%
Baseline results: CZ-MWE

- TPs = unanimous among 3 judges
- baseline: 21.03%
- best AM: chi-sq. (AP = 64.86%)
- frequency: AP = 21.70%
- log-likelihood much worse than on other data sets
Result overview: EN-VPC

<table>
<thead>
<tr>
<th>Method</th>
<th>Baseline</th>
<th>G2</th>
<th>X2</th>
<th>t-score</th>
<th>MI</th>
<th>Dice</th>
<th>freq</th>
<th>PE</th>
<th>EPI</th>
<th>Best AM</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.3</td>
<td>29.3</td>
<td>28.6</td>
<td>29.9</td>
<td>22.0</td>
<td>27.3</td>
<td>29.0</td>
<td>18.0</td>
<td>19.3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Graph showing performance metrics for different methods.
Result overview: DE-AN
Result overview: CZ-MWE

<table>
<thead>
<tr>
<th>Method</th>
<th>Baseline</th>
<th>G2</th>
<th>X2</th>
<th>t-score</th>
<th>MI</th>
<th>Dice</th>
<th>freq</th>
<th>PE</th>
<th>EPI</th>
<th>Best AM</th>
<th>ML</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>21.0</td>
<td>42.5</td>
<td>65.0</td>
<td>24.6</td>
<td>64.9</td>
<td>62.6</td>
<td>21.7</td>
<td>65.0</td>
<td>65.6</td>
<td>79.5</td>
<td>79.5</td>
</tr>
</tbody>
</table>

The table above shows the result overview for CZ-MWE with various metrics.
Result overview: combined
And now on to the participants …