A Survey of MWE Identification Experiments: The Devil is in the Details

Carlos Ramisch, Abigail Walsh, Thomas Blanchard, and Shiva Taslimipoor
MWE 2023 Workshop
Outline

Introduction and scope

Corpus constitution and selection

Pre- and post-processing

Evaluation metrics

Significance

Error analysis

Conclusions and open issues
Full-text multiword expression identification – sequence annotation
→ Focus of recurrent shared tasks (ST): DiMSUM & PARSEME

Survey’s goal

• Analyses MWE identification papers with experiments on data
• Look at methodological issues often seen as minor or omitted
• Hypothesis: these issues influence results and conclusions
Scope

Selection criteria:

- Available on the ACL Anthology
- Focus on MWE identification (Constant et al. 2017)
- Report experimental results
  - DiMSUM or PARSEME shared task or system description
    - OR
  - report experiments on DiMSUM or PARSEME corpora
Scope

Selection criteria:

- Available on the ACL Anthology
- Focus on MWE identification (Constant et al. 2017)
- Report experimental results
  - DiMSUM or PARSEME shared task or system description
    OR
  - report experiments on DiMSUM or PARSEME corpora

Paper stats

- 40 papers
  - 4 overall ST papers
  - 27 ST system descriptions
  - 9 non-ST system descriptions
Questions

- Data
  - Corpora
  - Pre- and post-processing
  - Sequence label encoding and decoding

- Evaluation
  - Metrics
  - Significance of comparisons
  - Error analysis
<table>
<thead>
<tr>
<th></th>
<th>2 Languages</th>
<th>3 Split of the corpora</th>
<th>4.1 Preprocessing</th>
<th>4.2 How are MWEs handled?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>PARSEME 1.0</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>The PARSEME Shared Task on Automatic Parsing</td>
<td>BG, CS, DE, EL</td>
<td>train/test, no dev</td>
<td>N/A</td>
</tr>
<tr>
<td>3</td>
<td>Parsing and MWE Detection: Fips at the IIT</td>
<td>FR, EN, DE, IT, EL</td>
<td>Not mentioned</td>
<td>VID, LVC, VPC, Transformation</td>
</tr>
<tr>
<td>4</td>
<td>The ATILF-LLF System for Parseme Shared Task</td>
<td>BG, CS, DE, EL</td>
<td>PARSEME data</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>5</td>
<td>Detection of Verbal Multi-Word Expressions</td>
<td>CS, DE, EL, ES</td>
<td>PARSEME data</td>
<td>VPC, LVC, VID</td>
</tr>
<tr>
<td>6</td>
<td>USzeged: Identifying Verbal Multiword Expressions</td>
<td>DE, EL, ES, FR</td>
<td>PARSEME 1.0 (no dev)</td>
<td>Single-token: representation</td>
</tr>
<tr>
<td>7</td>
<td>A data-driven approach to verbal multiword expressions</td>
<td>RO, FR, CS, DE</td>
<td>PARSEME 1.0 - cross</td>
<td>Two steps: Head detection, base forms</td>
</tr>
<tr>
<td>8</td>
<td>Neural Networks for Multi-Word Expressions</td>
<td>BG, CS, DE, EL</td>
<td>PARSEME 1.0</td>
<td>MWE category</td>
</tr>
<tr>
<td>9</td>
<td><strong>PARSEME 1.1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Edition 1.1 of the PARSEME Shared Task on Automatic Parsing</td>
<td>BG, DE, EL, EN</td>
<td>3 languages had no dev</td>
<td>LVC, VID, IRV, V</td>
</tr>
<tr>
<td>11</td>
<td>CRF-Seq and CRF-DepTree at PARSEME 2018</td>
<td>BG, DE, EL, EN</td>
<td>PARSEME 1.1 data</td>
<td>PARSEME 1.1</td>
</tr>
<tr>
<td>12</td>
<td>Deep-BGT at PARSEME Shared Task 2018</td>
<td>BG, DE, ES, FR</td>
<td>PARSEME 1.1 data</td>
<td>All PARSEME 1.1</td>
</tr>
<tr>
<td>13</td>
<td>GBD-NER at PARSEME Shared Task 2018</td>
<td>BG, DE, EL, EN</td>
<td>PARSEME 1.1 (no dev)</td>
<td>All PARSEME 1.1</td>
</tr>
<tr>
<td>14</td>
<td>Mumpitz at PARSEME Shared Task 2018</td>
<td>BG, DE, EL, ES</td>
<td>PARSEME 1.1 (they removed some categories)</td>
<td>Binary, whether a sub-graph</td>
</tr>
<tr>
<td>15</td>
<td>TRAPACC and TRAPACCS at PARSEME 2018</td>
<td>BG, DE, EL, EN</td>
<td>PARSEME 1.1 (param PARSEME 1.1)</td>
<td>Not mentioned</td>
</tr>
<tr>
<td>16</td>
<td>TRAVERSAL at PARSEME Shared Task 2018</td>
<td>BG, DE, EL, EN</td>
<td>PARSEME 1.1 (devel PARSEME 1.1)</td>
<td>Case lifting (chains)</td>
</tr>
<tr>
<td>17</td>
<td>VariDE at PARSEME Shared Task 2018</td>
<td>BG, DE, EL, EN</td>
<td>PARSEME 1.1 (no dev)</td>
<td>All PARSEME 1.1</td>
</tr>
<tr>
<td>18</td>
<td>Veyn at PARSEME Shared Task 2018</td>
<td>BG, DE, EL, EN</td>
<td>PARSEME 1.1 (no tun)</td>
<td>All PARSEME 1.1</td>
</tr>
<tr>
<td>19</td>
<td>SHOMA at Parseme Shared Task on Automatic Parsing</td>
<td>BG, DE, EL, EN</td>
<td>PARSEME 1.1 data</td>
<td>All PARSEME 1.1</td>
</tr>
<tr>
<td>20</td>
<td><strong>PARSEME 1.2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Edition 1.2 of the PARSEME Shared Task on Automatic Parsing</td>
<td>DE, EL, EU, FR</td>
<td>train/dev/test for all languages</td>
<td>LVC, VID, IRV, V</td>
</tr>
<tr>
<td>22</td>
<td>MultiVitaminBooster at PARSEME Shared Task 2018</td>
<td>DE, EL, EU, FR</td>
<td>PARSEME 1.2</td>
<td>All PARSEME 1.1</td>
</tr>
<tr>
<td>23</td>
<td>MTL-B-STRUCT at Parseme 2020</td>
<td>DE, EL, EU, FR</td>
<td>PARSEME 1.2</td>
<td>All PARSEME 1.1</td>
</tr>
</tbody>
</table>
Corpus splits

Shared tasks

- DiMSUM: 3 domains, 1 lang, train + test
- PARSEME 1.0: news, 18 lang, train + test
- PARSEME 1.1: news, 19 lang, train + test + dev (16 lang)
- PARSEME 1.2: news, 14 lang, train + test + dev
  → Biased split: focus on unseen MWEs
Corpus use

- **Training corpus** unused: 4/36 papers
  - External resources (2 papers), other corpora (2 papers)

- **Development corpus** not provided:
  - Custom train-dev set: 6/36 papers
  - Cross-validation: 3/36 papers
  - Dev on another language: 2/36 papers
  - Dev corpus not mentioned: 3/36 papers

Recommendation: Always mention development data
Corpus use

- **Training corpus** unused: 4/36 papers
  - External resources (2 papers), other corpora (2 papers)

- **Development corpus** not provided:
  - Custom train-dev set: 6/36 papers
  - Cross-validation: 3/36 papers
  - Dev on another language: 2/36 papers
  - Dev corpus not mentioned: 3/36 papers

**Recommendation**
Always mention development data
Languages

Number of papers

Chinese
Czech
Maltese
Irish
Croatian
Swedish
Lithuanian
Farsi
Bulgarian
Hindi
Slovene
Hungarian
Basque
Hebrew
Spanish
Turkish
English
Romanian
Portuguese
Polish
Italian
Greek
German
French
Outline

Introduction and scope

Corpus constitution and selection

Pre- and post-processing

Evaluation metrics

Significance

Error analysis

Conclusions and open issues
Pre-processing

Variants of **BIO-style encoding**: 12/36 papers

DiMSUM  The staff leaves a lot to be desired.

O O B b i_ i_ i_ i_ O

PARSEME  I did a lot of study and research.

* 1:LVC;2:LVC * * * 1 * * 2 *

- **Gaps**: 12/36 papers account for gaps
- **Nesting and overlaps**
  - Ignored, handled by modifying BIO-style
  - Kept the tags as they are, dependency graphs
  - No mention (most papers)
Post-processing

Conversion from BIO-style

- Combination **heuristics** (7/36 papers)
  - $\rightarrow$ B-labelled and I-labelled words matched
  - $\rightarrow$ Standalone I-labelled ignored
- **Greedy-matching** algorithm (1/36 paper)
- **Viterbi** decoding (1/36 paper)
- **Conditional random fields** (8/36 papers)
- **Dependency trees** (2/36 papers)
  - $\rightarrow$ Elements of MWE assumed to be nodes in the same subtree

Recommendation
Explicitly report all pre- and post-processing + MWE encoding
Post-processing

Conversion from BIO-style

- Combination **heuristics** (7/36 papers)
  - B-labelled and I-labelled words matched
  - Standalone I-labelled ignored
- **Greedy-matching** algorithm (1/36 paper)
- **Viterbi** decoding (1/36 paper)
- **Conditional random fields** (8/36 papers)
- **Dependency trees** (2/36 papers)
  - Elements of MWE assumed to be nodes in the same subtree

**Recommendation**
Explicitly report all pre- and post-processing + MWE encoding
Outline

Introduction and scope

Corpus constitution and selection

Pre- and post-processing

Evaluation metrics

Significance

Error analysis

Conclusions and open issues
Evaluation metrics

DIMSUM exact match and linked-based P, R and F1

PARSEME MWE-based and token-based P, R and F1

PARSEME focused measures:

- Seen/Unseen: focus of 9 papers
- Diversity: 2 PARSEME papers
- Discontinuity: focus of 5 papers
Evaluation metrics

DIMSUM exact match and linked-based P, R and F1

PARSEME MWE-based and token-based P, R and F1

PARSEME focused measures:

- Seen/Unseen: focus of 9 papers
- Diversity: 2 PARSEME papers
- Discontinuity: focus of 5 papers

Recommendation
Focused measures help highlight system strengths and limitations
Compare systems A and B

- Test set
  - \( x = x^{(1)} \ldots x^{(m)} \) – \( m \) input sentences
  - \( y = y^{(1)} \ldots y^{(m)} \) – \( m \) reference MWE annotations

- Method:
  1. Apply \( A \) to \( x \) to obtain \( \hat{y}_A \), compare to \( y \)
  2. Calculate evaluation metric \( M(A, x, y) \) (e.g. MWE-based F1)
  3. Do the same for \( B \), obtain \( M(B, x, y) \)
  4. Calculate difference (effect)

\[
\delta_{A-B}(x, y) = M(A, x, y) - M(B, x, y)
\]

- \( \delta_{A-B}(x, y) > 0 \implies A \) better than \( B \)?
Hypothesis testing

- \( H_0 : \delta(X, Y) \leq 0 \implies \text{if true, then } A \text{ not better than } B \)
- \( H_1 : \delta(X, Y) > 0 \)

- \( X, Y \rightarrow \text{random variables, all possible test sets} \)
  - Of which \( x, y \) is an \( m \)-sized sample

- Reject \( H_0 \implies \text{significant difference between the systems} \)

- **P-value**: probability of observing \( \delta_{A-B}(x, y) \) while \( H_0 \) is true:
  - \( p-value = P[\delta(X, Y) \geq \delta_{A-B}(x, y)|H_0] \)
  - probability to reject \( H_0 \) when it is true
**Input**

- Test set $x = x^{(1)} \ldots x^{(m)}, y = y^{(1)} \ldots y^{(m)}$,
- Predictions $\hat{y}_A^{(i)}$ and $\hat{y}_B^{(i)}$ of systems $A$ and $B$
- Evaluation metric $M(\cdot)$

```python
1  deltaobs = M(A,x,y) - M(B,x,y)  # observed difference
2  for i in range(R) :              # R constant 10k
3      xsample, ysample = sample(x,y,m)  # m with repetition
4  deltasample = M(A,xsample,ysample) - M(B,xsample,ysample)
5  if deltasample > 2 * deltaobs :  
4      r = r + 1
6  pvalue = r/R                   # % of surprising results
7  return pvalue
```

*Bootstrap p-value* (Berg-Kirkpatrick et al. 2012)
Significance analysis

- Only 2/40 papers report significance
- Our tool estimates p-values for two CUPT predictions
  → https://gitlab.com/parseme/significance
- We compare all system pairs and metrics of PARSEME 1.2
  → 2,728 p-values in total
  → 783 above the $\alpha = 0.05$ threshold (29%)
# P-values for MWE-based F1 in Swedish

<table>
<thead>
<tr>
<th>Systems</th>
<th>TRAVIS-multi</th>
<th>Seen2Unseen</th>
<th>TRAVIS-mono</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTLB-STRUCT</td>
<td>0.7158</td>
<td>0.025</td>
<td>0.0</td>
</tr>
<tr>
<td>TRAVIS-multi</td>
<td>0.6911</td>
<td>0.464</td>
<td>0.081</td>
</tr>
<tr>
<td>Seen2Unseen</td>
<td>0.6892</td>
<td>0.103</td>
<td>0.103</td>
</tr>
</tbody>
</table>

Recommendation:
Systematically calculate/report p-values for model comparison.
## P-values for MWE-based F1 in Swedish

<table>
<thead>
<tr>
<th>Systems</th>
<th>TRAVIS-multi</th>
<th>Seen2Unseen</th>
<th>TRAVIS-mono</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTLB-STRUCT</td>
<td>0.7158</td>
<td>0.025</td>
<td>0.0</td>
</tr>
<tr>
<td>TRAVIS-multi</td>
<td>0.6911</td>
<td>0.464</td>
<td>0.081</td>
</tr>
<tr>
<td>Seen2Unseen</td>
<td>0.6892</td>
<td>0.103</td>
<td>0.103</td>
</tr>
</tbody>
</table>

**Recommendation**  
Systematically calculate/report p-values for model comparison
Outline

Introduction and scope

Corpus constitution and selection

Pre- and post-processing

Evaluation metrics

Significance

Error analysis

Conclusions and open issues
Error Analysis

- 33/36 papers report some error analysis
- 11/36 report MWE category or cross-language analyses
- Heterogeneous analyses
  - Discontinuities, seen/unseen
  - POS sequences, syntactic structure
  - Ablation, role of external lexicons
  - Pre-trained embeddings, tagging schemes

Recommendation
Error analyses uncover interesting phenomena for future work
Error Analysis

- 33/36 papers report some error analysis
- 11/36 report MWE category or cross-language analyses
- Heterogeneous analyses
  - Discontinuities, seen/unseen
  - POS sequences, syntactic structure
  - Ablation, role of external lexicons
  - Pre-trained embeddings, tagging schemes

Recommendation
Error analyses uncover interesting phenomena for future work
Outline

Introduction and scope

Corpus constitution and selection

Pre- and post-processing

Evaluation metrics

Significance

Error analysis

Conclusions and open issues
Recommendations

We advocate **reporting on experimental choices:**

- corpus constitutions and selections
- pre- and post-processing
- evaluation metrics and significance testing of performance
- error analysis

We encourage **focused measures** that facilitate error analysis

We propose a **tool to predict p-values** from 2 CUPT predictions
Open issues

- **Hyper-parameter tuning**
  - Selection of the data
  - Strategy (e.g. grid search, random, etc.)

- **Should manual evaluation** of detected MWEs be performed?

- **New evaluation protocols**
  - e.g. are some MWE categories more important than others?
Thanks! Questions?

Acknowledgments

- SELEXINI project (ANR-21-CE23-0033-01)
- Ulysses PHC grant (45358QD)
- Marius Thorre, Silvio Cordeiro, Agata Savary, reviewers