An open-source tool for negation detection – a maximum-margin approach

Martine Enger and Erik Velldal and Lilja Øvrelid
Department of Informatics, University of Oslo
{marenger|erikve|liljao}@ifi.uio.no

Workshop on Computational Semantics Beyond Events and Roles
April 4 2017, Valencia, Spain
Examples

- And yet it was **not** quite the last.
- Since we have been so **unfortunate** as to **miss** him and have […]

- Much published research, but hard to find available systems.
- **Goal:** Implement a lean and simple tool with minimal dependencies that
  - open and freely available,
  - easy to adapt (i.e. keep heuristics to a minimum),
  - draws on previous best practices, and
  - has competitive performance.

- Draws heavily on the design of the **UiO$_2$** system (Lapponi et al., 2012) from the *SEM shared task 2012.*
We use the data and evaluation script from the 2012 *SEM shared task on negation detection (Morante & Blanco, 2012).

Training, development and heldout testing based on the CoNLL-style Conan Doyle corpus (Morante & Daelemans, 2012).

We only focus on cues and scopes (not events and focus).

Use Stanford basic dependency representations rather than the provided constituent trees of the Charniak and Johnson (2005) parser.
System design, at a glance

- A maximum-margin learning approach for both cues and scopes.
- Implemented on top of PyStruct.
- Takes parsed (CoNLL-X) or raw text (assumes CoreNLP is installed).
- **Cue** detection: Binary SVM with lexical features
- **Scope** detection: SVM-based CRF with lexical and syntactic features.
Cue identification

- **Closed-class assumption**: only attempt to disambiguate cues seen during training (Velldal, 2011; Read et al., 2012).

- A lexicon of **affixal cues** is also automatically extracted
  - prefixes: \{dis, im, in, ir, un\}
  - infix: less
  - suffix: less
  - The classifier is presented with any words matching an affix pattern, e.g. *im*patient, *im*age and *im*aginary would match the prefix pattern.

- **Features**: Token PoS, form and lemma, as well as lemmas ± 1 position.

- Additional features for affixal candidates: the affix itself and character 5-grams from start/end of ‘base’.

- **Multi-word cues**, e.g. ‘by no means’ or ‘neither...nor’: Post-processing.
Cue identification, results

<table>
<thead>
<tr>
<th></th>
<th>Development</th>
<th></th>
<th></th>
<th></th>
<th>Held-out</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>90.68</td>
<td>84.39</td>
<td>87.42</td>
<td></td>
<td>87.10</td>
<td>92.05</td>
<td>89.51</td>
<td></td>
</tr>
<tr>
<td>UiO₂</td>
<td>93.75</td>
<td>95.38</td>
<td>94.56</td>
<td></td>
<td>89.17</td>
<td>93.56</td>
<td>91.31</td>
<td></td>
</tr>
<tr>
<td>System</td>
<td>91.67</td>
<td>95.38</td>
<td>93.49</td>
<td></td>
<td>90.15</td>
<td>93.56</td>
<td>91.82</td>
<td></td>
</tr>
</tbody>
</table>

- Majority class baseline:
  - Assign each word its most frequent label in the training data.
  - Outperforms 1/3 of the *SEM 2012 shared task systems.

- Slight drop in F1 when moving from the dev. to held-out set.

- Compared to UiO₂: recall is identical, but our system has more stable precision (1 percentage point higher on held-out).

- Would have ranked third in the *SEM 2012 shared task.
Approached as a sequence labeling task using a maximum-margin CRF.

Features and labels inspired by the UiO₂ system from *SEM 2012.

Configuration after tuning (along with the regularization parameter):

- **Surface features**: Form, lemma ($\pm 1$), and PoS ($\pm 1$), cue PoS, cue type, and left/right cue distance.

- **Dependency features**: Graph distance and path from cue.

- **Label set**: Beginning, Inside, Outside, and Cue.
Scope resolution results, for *gold cues*

- **System comparison:**

  - **Scope-level F1 for gold cues** on the development and held-out set.

<table>
<thead>
<tr>
<th></th>
<th>Dev. F1</th>
<th>Test F1</th>
<th>Approach</th>
<th>Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>UiO₂</td>
<td>80.00</td>
<td>–</td>
<td>CRF</td>
<td>no</td>
</tr>
<tr>
<td>UiO₁</td>
<td>82.52</td>
<td>77.26</td>
<td>rules, SVM</td>
<td>no</td>
</tr>
<tr>
<td>Packard 2014</td>
<td>69.30</td>
<td>61.30</td>
<td>rules</td>
<td>scopes</td>
</tr>
<tr>
<td>Packard + UiO₁</td>
<td>82.50</td>
<td><strong>78.70</strong></td>
<td>rules(\times 2), SVM</td>
<td>scopes, partly</td>
</tr>
<tr>
<td>Fancellu 2016</td>
<td>–</td>
<td>77.77</td>
<td>BiLSTM</td>
<td>scopes</td>
</tr>
<tr>
<td>System</td>
<td>77.38</td>
<td>77.26</td>
<td>CRF</td>
<td>scopes + cues</td>
</tr>
</tbody>
</table>
End-to-end results (scope-level metric)

<table>
<thead>
<tr>
<th></th>
<th>Development</th>
<th></th>
<th>Held-out</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
<td>P</td>
</tr>
<tr>
<td>System (gold cues)</td>
<td>100.00</td>
<td>63.10</td>
<td>77.38</td>
<td>98.75</td>
</tr>
<tr>
<td>System</td>
<td>88.14</td>
<td>61.90</td>
<td>72.73</td>
<td>85.00</td>
</tr>
</tbody>
</table>

- **Cue classification errors propagate** to the scope classifier which will predict scopes for FP cues and do nothing for FN cues.
- Mostly affects precision.
- End-to-end system would have ranked 4th in the *SEM 2012* shared task w.r.t. the relevant subtasks.
Scope error analysis (for gold cues)

- Struggles with **discontinuous** scopes, as in:

  *It was not, I must confess, a very alluring prospect.*

- Other types of recurring errors: sentences with multiple negation cues with **overlapping** (gold) scopes.

- Many cases counted as FNs wrt the scope-level measure often just have a **single token wrong**, reflected in the higher token-level scores.
Future work

- Extend the scope resolution with post-processing heuristics for targeting discontinuous scopes.
- Train/test for other tasks, domains and annotation strategies.
- For example; speculation (and negation) detection based on BioScope.
- Pre-requisite: convert annotation format (XML → CoNLL-X).
Conclusion

▶ [https://github.com/marenger/negtool](https://github.com/marenger/negtool)

▶ A simple and open-source tool for detecting negation cues and their in-sentence scopes with competitive performance.

▶ Mostly relies on learned models, based on a maximum-margin approach.

▶ Pre-trained models for English are distributed along with the code, users can also train their own models.


approach to hedge detection in biomedical literature. *Journal of Biomedical Semantics, 2*(5).