Linguistic Cues for Distinguishing Literal and Non-Literal Usages
Linlin Li and Caroline Sporleder
Cluster of Excellence (MMCI), Saarland University, Germany

Introduction

Nonliteral expressions are a major challenge in NLP
  - Fairly frequent
  - Often behave idiosyncratically

Most approaches have focused on a type-based classification
  - Dividing expressions into idiomatic or not idiomatic
  - Many expressions can be used literally as well as idiomatically (e.g., spill the
    bills, break the ice).

In this work:
  - What features are useful for token-based idiom detection?
  - Whether these features can generalize across idioms?

Global Lexical Context Feature (glc)

- Words which are either semantically related to the literal or the nonliteral
  sense of the target expression
  1. “Gujral will meet Sharif on Monday and discuss bilateral relations,” the Press Trust of
     India added. The minister said Sharif and Gujral would be able to “break the ice” over
     Kashmir.
  2. Meanwhile in Germany, the cold penetrated Cologne cathedral, where worshippers had to
     break the ice on the frozen holy water in the font.
- We implemented two sets of features which encode the global lexical
  context: salient words and related words as described in Li and Sporleder
  (2009).

Local Lexical Context Feature (locCont)

- The word “Literally” preceding an expression often indicates “idiomatic”
  usage:
  - In the documentary the producer literally spills the beans on the real devil behind
    the movie production.
  - The new philosophy is blatantly permissive and literally gives the buck to the House’s
    officer committees.
- There are also some idiom-specific local lexical cues. For example, we found
  that the word just before the expression get ones feet wet tends to indicate
  nonliteral usage.
- The wiki includes a page of tasks suitable for those just getting their feet wet.

Syntactic Feature (allSyn)

- We included a set of syntax features which are extracted from the
  parsed dependency tree. We use MaltParser (Nivre et al., 2006).

Dependency Relation of the Head Verb

- Ross headed back last week to Washington to brief president Bill Clinton on
  the Hebron talks after achieving a breakthrough in
  breaking the ice in the Hebron talks by arranging an Arafat-Netanyahu summit.

Modal Verbs
- Dad had to break the ice on the chicken troughs.

Subjects
- The visit of the minister may break the ice between India and Pakistan.

Verb Subcat
- US defender Alexi Lalas twice went close to forcing an equaliser , first with
  a glancing equaliser from a Paul Caligiuri free kick and then from a
  Wynaolder corner when Prunea dropped the ball [on the ground] only [for
  Tibor Selyme to kick frantically clear].
- “Clinton dropped the ball [on this],” said John Parachini.

Modifiers
- The wiki includes a page of tasks suitable for those just getting their feet wet.

Coordinated Verb
- They may break the ice and fall through.

Discourse Cohesion Feature (dc)

- Aims at lexical cohesion cues between an expression and the surrounding
  discourse
  - Related score and discourse connectivity from Li and Sporleder (2009)
  - The prediction of the cohesion graph proposed by Sporleder and Li (2009).
- These features are more likely to generalize across different idioms.

Other Cues

- Named Entities (ne)
  - Dudayev repeated his frequent warnings that Russia was playing with fire.
- Indicative Terms (ITerm)
  - e.g., literally, proverbially
- Scare Quotes (quote)
  - Do consider “getting your feet wet” online, using some of the technology that is now available to us.

Experiments

- 3795 instances of 13 English expressions (mainly V+PP or V+NP) from
  Sporleder and Li (2009)
- We use SVM classifier
  - Three experiments: (1) train one model for each idiom; (2) train one
    generic model for all idioms; (3) test generic model on unseen idioms.

Idiom Specific Models

<table>
<thead>
<tr>
<th>Avg. literal</th>
<th>Avg. Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>feature</td>
<td>Prec.</td>
</tr>
<tr>
<td>all</td>
<td>89.84</td>
</tr>
<tr>
<td>glc+dc</td>
<td>90.42</td>
</tr>
<tr>
<td>allSync</td>
<td>76.30</td>
</tr>
<tr>
<td>heaSync</td>
<td>76.64</td>
</tr>
<tr>
<td>parSync</td>
<td>76.43</td>
</tr>
<tr>
<td>chSync</td>
<td>76.49</td>
</tr>
<tr>
<td>sbSync</td>
<td>76.27</td>
</tr>
<tr>
<td>locCont</td>
<td>76.51</td>
</tr>
<tr>
<td>ne</td>
<td>76.48</td>
</tr>
<tr>
<td>ITerm</td>
<td>76.51</td>
</tr>
<tr>
<td>quote</td>
<td>76.51</td>
</tr>
<tr>
<td>Baseline</td>
<td>76.71</td>
</tr>
</tbody>
</table>

Generic Models

<table>
<thead>
<tr>
<th>Avg. literal</th>
<th>Avg. Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>feature</td>
<td>Prec.</td>
</tr>
<tr>
<td>all</td>
<td>89.59</td>
</tr>
<tr>
<td>glc+dc</td>
<td>82.53</td>
</tr>
<tr>
<td>allSync</td>
<td>50.83</td>
</tr>
<tr>
<td>heaSync</td>
<td>50.57</td>
</tr>
<tr>
<td>parSync</td>
<td>33.33</td>
</tr>
<tr>
<td>chSync</td>
<td>62.45</td>
</tr>
<tr>
<td>sbSync</td>
<td>40.00</td>
</tr>
<tr>
<td>ne</td>
<td>0.00</td>
</tr>
<tr>
<td>ITerm</td>
<td>0.00</td>
</tr>
<tr>
<td>quote</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Unseen Idioms

- The performance drops compared with the idiom specific model.
- However, the performance still statistically significantly beats the
  majority baseline classifier.

Conclusion

- Idioms on the whole are frequent, but instances of each particular idiom can be
  relatively infrequent. Thus, there is need to explore idiom-independent features.
- Our lexical cohesion-based features can generalize across idioms.
- Although certain linguistic features are intuitively linguistically informative,
  they do not gain performance which may be due to data sparseness.

MCCI, Saarland University

{linlin,csporleder}@coli.uni-sb.de