Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb–noun combinations

Milton King and Paul Cook
University of New Brunswick
Fredericton, Canada
Multiword Expressions

• Expressions of multiple words that can exhibit an idiomatic meaning
  – *Ivory tower*
  – *Hit up*
  – *Take a walk*

• Verb noun combinations
  – *See stars*
  – *Kick the bucket*
Idiomatic vs Literal

• Pull plug
  – (I) They *pulled* the plug on the Department of Health funding
  – (L) Unfortunately someone *pulled* the sink *plug*

• See stars
  – (I) It caught him on the head and he went down *seeing* little sparkling *stars*
  – (L) It’s still dark enough to *see* the brightest *stars*
Idiom Token Classification

• Determine if an MWE instance is idiomatic
  – *They pulled the plug on the project* ➔ [Idiomatic/Literal]

• Applications
  – Machine translation
    • *Kick the bucket* ➔ [*mourir/frapper avec le pied*]
  – Sentence completion
    • *Keegan is ready to pull the plug on* [a deal / the tv]
Overview of Approach

• Supervised approach
• VNC token instances are represented via use of an embedding model
• Embedding models
  – Skip-thoughts
  – Word2vec
  – Siamese CBOW
• SVM classifier
Lexico-Syntactic Fixedness

• The idiomatic meaning of an expression is typically restricted to a small number of lexico-syntactic patterns

• **See star** (Idiomatic)
  – Active voice, no determiner, plural noun
    • *See stars*

• **See star** (Literal)
  – Active voice, determiner, singular noun
    • *See a star*
  – Passive voice, plural noun
    • *Stars were seen*
Patterns

<table>
<thead>
<tr>
<th>Pattern No.</th>
<th>Pattern Signature</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$v_{act}$ det:NULL $n_{sg}$</td>
<td>give money</td>
</tr>
<tr>
<td>2</td>
<td>$v_{act}$ det:a/an $n_{sg}$</td>
<td>give a book</td>
</tr>
<tr>
<td>3</td>
<td>$v_{act}$ det:the $n_{sg}$</td>
<td>give the book</td>
</tr>
<tr>
<td>4</td>
<td>$v_{act}$ det:DEM $n_{sg}$</td>
<td>give this book</td>
</tr>
<tr>
<td>5</td>
<td>$v_{act}$ det:POSS $n_{sg}$</td>
<td>give my book</td>
</tr>
<tr>
<td>6</td>
<td>$v_{act}$ det:NULL $n_{pl}$</td>
<td>give books</td>
</tr>
<tr>
<td>7</td>
<td>$v_{act}$ det:the $n_{pl}$</td>
<td>give the books</td>
</tr>
<tr>
<td>8</td>
<td>$v_{act}$ det:DEM $n_{pl}$</td>
<td>give those books</td>
</tr>
<tr>
<td>9</td>
<td>$v_{act}$ det:POSS $n_{pl}$</td>
<td>give my books</td>
</tr>
<tr>
<td>10</td>
<td>$v_{act}$ det:OTHER $n_{sg,pl}$</td>
<td>give many books</td>
</tr>
<tr>
<td>11</td>
<td>$v_{pass}$ det:ANY $n_{sg,pl}$</td>
<td>a/the/this/my book/books was/were given</td>
</tr>
</tbody>
</table>

Afsaneh Fazly et al. 2009
Canonical Form

• Lexico-syntactic patterns that idiomatic usages tend to occur in

\[ C(v, n) = \{ pt_k \in \mathcal{P} \mid z(v, n, pt_k) > T_z \} \]

\[ z(v, n, pt_k) = \frac{f(v, n, pt_k) - \bar{f}}{s} \]

Afsaneh Fazly et al. 2009
Integrating Canonical Forms

• Unsupervised method used in Fazly et al. to identify canonical forms

• One-dimensional binary vector representing if the expression is in the canonical form
VNC-Tokens Dataset
Cook et al. 2008

• Dev
  – 14 MWEs
  – Training
    • 270 Idiom
    • 179 Literal
  – Testing
    • 92 Idiom
    • 53 Literal

• Test
  – 14 MWEs
  – Training
    • 298 Idiom
    • 172 Literal
  – Testing
    • 90 Idiom
    • 53 Literal
## Accuracy

<table>
<thead>
<tr>
<th>Model</th>
<th>DEV</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>−CF</td>
<td>+CF</td>
</tr>
<tr>
<td>CForm</td>
<td>-</td>
<td>0.721</td>
</tr>
<tr>
<td>Word2vec</td>
<td>0.830</td>
<td>0.854</td>
</tr>
<tr>
<td>Siamese CBOw</td>
<td>0.763</td>
<td>0.774</td>
</tr>
<tr>
<td>Skip-thoughts</td>
<td>0.803</td>
<td>0.827</td>
</tr>
</tbody>
</table>
## Results per class

<table>
<thead>
<tr>
<th>Model</th>
<th>Idiomatic</th>
<th></th>
<th></th>
<th></th>
<th>Literal</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F</td>
<td></td>
<td>P</td>
<td>R</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Word2vec -CF</td>
<td>0.815</td>
<td>0.879</td>
<td>0.830</td>
<td></td>
<td>0.627</td>
<td>0.542</td>
<td>0.556</td>
<td></td>
</tr>
<tr>
<td>Word2vec +CF</td>
<td>0.830</td>
<td>0.892</td>
<td>0.848</td>
<td></td>
<td>0.758</td>
<td>0.676</td>
<td>0.691</td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

• Averaging word2vec embeddings outperforms all other models used
• Canonical form feature improves results
• Future work
  – Unseen MWEs
  – Other embedding models
Thank you

This work was financially supported by NSERC, NBIF, and University of New Brunswick
## Results per class

<table>
<thead>
<tr>
<th>Model</th>
<th>Idiomatic</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
<td>F</td>
<td>P</td>
</tr>
<tr>
<td>CForm</td>
<td>0.766</td>
<td>0.901</td>
<td>0.794</td>
<td></td>
<td>0.668</td>
</tr>
<tr>
<td>Word2vec −CF</td>
<td>0.815</td>
<td>0.879</td>
<td>0.830</td>
<td></td>
<td>0.627</td>
</tr>
<tr>
<td>Word2vec +CF</td>
<td>0.830</td>
<td>0.892</td>
<td>0.848</td>
<td></td>
<td>0.758</td>
</tr>
</tbody>
</table>