Gotta catch ’em all: Crowdsourcing a wide-coverage corpus of idiomatic expressions in English

Hessel Haagsma, Malvina Nissim & Johan Bos

“What I know of Moscow is that if you keep your nose clean and do the work, there’s no hassle.”

↓

“What I know of Moscow is that if you wash your nose and do the work, there’s no hassle.”
or

“What I know of Moscow is that if you behave properly and do the work, there’s no hassle.”

Please read the following sentence(s) carefully.

What is the meaning of “keep your nose clean” in the sentence you have just read? (required)

• Idiomatic

• Literal

• Other

What does ‘keep your nose clean’ mean?

Definition

stay out of trouble, informal

Existing Corpora

<table>
<thead>
<tr>
<th>Name</th>
<th># Idioms</th>
<th># Instances</th>
<th>Corpus Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>VNC-Tokens</td>
<td>53</td>
<td>2,984 BNC</td>
<td>V+NP</td>
</tr>
<tr>
<td>Gigaword</td>
<td>17</td>
<td>3,964 Gigaword</td>
<td>V+NP/PP</td>
</tr>
<tr>
<td>IDIX</td>
<td>52</td>
<td>4,022 BNC</td>
<td>V+NP/PP</td>
</tr>
<tr>
<td>SemEval-2013</td>
<td>65</td>
<td>4,350 ukWaC</td>
<td>unrestricted</td>
</tr>
</tbody>
</table>

Goals

▶ > 1000 types
▶ ≤ 200 instances per type
▶ Unrestricted syntactic patterns
▶ Enable fine-grained evaluation
▶ Facilitate deep learning approaches
▶ Focus on cross-idiom disambiguation

Setup

▶ Automatic pre-extraction of candidates
▶ Source: British National Corpus
▶ Idioms: (Wiki ∩ Oxford) ∪ (UE.com ∩ Oxford)
▶ Platform: FigureEight (f.k.a CrowdFlower)
▶ Test questions from existing PIE Corpus

Idiom Frequencies

Open Questions

▶ Agreement threshold
▶ Number of annotations
▶ Idioms within other idioms
▶ Single-sense idioms
▶ Strong ambiguity
   Commandos were digging in all over the place.
▶ Limits on modification
   [...] jumping on the Tetris/Columns bandwagon.