1. Motivation

- Recognizing lexical inferences between term-pairs is typically performed within a given context.
- Such context-sensitive inferences have to consider:
  - Term meaning in context:
    \[
    S_1: \text{Which teams competed in the Super Bowl?} \\
    S_2: \text{The national anthem is played at international football matches.}
    \]
    \text{ compete # play}
  - The fine-grained relation between the terms:
    \[
    S_1: \text{My iPhone’s battery is low} \Rightarrow \text{My phone’s battery is low} \\
    S_2: \text{Talking on the phone is prohibited} \Rightarrow \text{Talking on the iPhone is prohibited}
    \]
    \text{iPhone \# phone, S_1: upward-monotone, S_2: downward-monotone}
- Required: datasets annotated with fine-grained semantic relations in-context.
- Existing datasets: annotated either out-of-context or to coarse-grained relations (e.g. similarity).

2. Generating Candidates

- PPDB 2.0 [1] contains the semantic relations between paraphrases, following natural logic [2]:

<table>
<thead>
<tr>
<th>Relation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>equivalence</td>
<td>is the same as</td>
</tr>
<tr>
<td>forward entailment</td>
<td>is more specific than</td>
</tr>
<tr>
<td>reverse entailment</td>
<td>is more general than</td>
</tr>
<tr>
<td>alternation</td>
<td>is mutually exclusive with</td>
</tr>
<tr>
<td>other-related</td>
<td>is related in some other way to</td>
</tr>
<tr>
<td>independence</td>
<td>is not related to</td>
</tr>
</tbody>
</table>

- We use their manually-annotated term-pairs (PPDB-fine-human), filtering them accordingly:
  - Leave out independent terms
  - Filter out ungrammatical phrases
  - Remove trivial pairs:
    * inflections (Iraq, Iraqi)
    * alternate spellings (center, centre)
    * same lemma
  - Remove determiners (a kid, the boy) \(\Rightarrow\) (kid, boy)
- We sample a subset of the filtered term-pairs.
- For each \((x, y)\) term-pair, we add 10 random sentence-pairs \((s_x, s_y)\) from WikiNews, such that \(x \in s_x\) and \(y \in s_y\).

3. Annotation Task

- We re-annotate term-pairs with respect to given contexts, using Amazon Mechanical Turk:
  - Leave out independent terms
  - Filter out ungrammatical phrases
  - Remove trivial pairs:
    * inflections (Iraq, Iraqi)
    * alternate spellings (center, centre)
    * same lemma
  - Remove determiners (a kid, the boy) \(\Rightarrow\) (kid, boy)
  - We sample a subset of the filtered term-pairs.
  - For each \((x, y)\) term-pair, we add 10 random sentence-pairs \((s_x, s_y)\) from WikiNews, such that \(x \in s_x\) and \(y \in s_y\).

- Each HIT was assigned to 5 workers, and gold labels were selected using the majority rule.
- Moderate levels of agreement: Fleiss’ Kappa \(\kappa = 0.51\). Out-of-context annotation agreement is lower \((\kappa = 0.46)\).
- We split the dataset to 70% train, 25% test, and 5% validation sets.

4. Analysis

<table>
<thead>
<tr>
<th>Relation</th>
<th>in-context</th>
<th>out-of-context</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\equiv)</td>
<td>60.54</td>
<td>6.9</td>
</tr>
<tr>
<td>(\sqsubseteq)</td>
<td>6.9</td>
<td>1.41</td>
</tr>
<tr>
<td>(\sqsupseteq)</td>
<td>1.41</td>
<td>2.08</td>
</tr>
<tr>
<td>(\sim)</td>
<td>2.08</td>
<td>11.13</td>
</tr>
<tr>
<td>#</td>
<td>11.13</td>
<td>39.17</td>
</tr>
</tbody>
</table>
- The original relation holds in many of the contexts.
- Term-pairs often become independent:
  \(S_1: \text{Roughly 1,500 gold and silver pieces were found} \ldots \) \(S_2: \text{... he will pull out of the race} \rightarrow \text{the president of Zimbabwe.}\)
- Sometimes the semantic relation is changed:
  \(S_1: \text{3 countries withdrew from the competition} \text{: Germany, Spain and Switzerland.} \quad S_2: \text{... he will pull out of the race} \rightarrow \text{race}\)

5. Baseline Results

- Baseline performance on the test set (mean over all classes):
  \[
  \begin{array}{l|cccc}
  & \text{precision} & \text{recall} & F_1 \vspace{-0.2cm} \\
  \hline
  \text{PPDB-fine-human} & 0.722 & 0.380 & 0.288 \\
  \text{PPDB2 classifier} & 0.611 & 0.565 & 0.556 \\
  \text{in-context classifier} & 0.677 & 0.685 & 0.670 \\
  \end{array}
  \]
- Our classifier contains PPDB 2.0 features + 3 context features that measure similarities between each term and the other term’s context, and between the contexts.
- The context-sensitive method notably outperforms context insensitive baselines.
- Our dataset is effective for developing fine-grained context-sensitive lexical inference methods.
- We need to develop better methods!

6. References


Get the Dataset & Contact Us
The dataset and annotation guidelines are available at: [http://goo.gl/QD7j1q](http://goo.gl/QD7j1q)
Email vered1986@gmail.com