SherLIiC: A Typed Event-Focused Lexical Inference Benchmark for Evaluating Natural Language Inference

Martin Schmitt and Hinrich Schütze
Center for Information and Language Processing, University of Munich
martin@cis.lmu.de

Summary

(1) We provide a new evaluation benchmark for NLI.
(2) The new challenge is very hard for current models.
(3) Knowledge graph embeddings completely fail.

Introduction

SherLIiC is a controlled yet challenging testbed:

- Binary entailment detection for Lexical Inference in Context (LIiC)
- Abstract context with knowledge graph types
- Very similar sentences
- Distributional similarity of positive and negative examples

For two typed relations \( A, B \subseteq \mathcal{E} \times \mathcal{E} \), we compute three scores:

\[
\text{Relv}(A, B) := \frac{P(B | A)}{P(B)}
\]

\[
\text{esr}(A, B) := \frac{\sum_{H \in [2]} \pi_i(A \cap B)}{2 |A \cap B|}
\]

\[
\sigma(A, B) := \frac{2 |A \cap B|}{\sum_{H \in [2]} \pi_i(A \cap B)} \log(\text{Relv}(A, H))
\]

\( A \Rightarrow B \) is accepted as inference candidate if \( \sigma(A, B) \geq 15 \), \( \text{esr}(A, B) \geq 0.6 \) and \( \forall i \in \{1, 2\} : |\pi_i(A \cap B)| \geq 5 \).

Inference Candidate Collection

- **synonymy**
  - ORGF[A] is supporter of ORGF[B]
  - ORGF[A] is backing ORGF[B]
- **typical actions**
  - AUTH[A] is president of LOC[B]
  - AUTH[A] is representing LOC[B]
- **common sense knowledge**
  - ORGF[A] claims LOC[B]
  - ORGF[A] is wanting LOC[B]
- **directionality**
  - PER[A] is region[B]'s ruler
  - PER[A] is dictator of region[B]
- **antonymy**
  - LOC[A] is fighting with ORGF[B]
  - LOC[A] is allied with ORGF[B]
- **correlation**
  - ORGF[A] is seeking from ORGF[B]
  - ORGF[A] is giving to ORGF[A]

Positive (top) and negative (bottom) examples from SherLIiC-dev.

**Inference Candidate Collection**

- **synonymy**
  - ORGF[A] is supporter of ORGF[B]
  - ORGF[A] is backing ORGF[B]
- **typical actions**
  - AUTH[A] is president of LOC[B]
  - AUTH[A] is representing LOC[B]
- **common sense knowledge**
  - ORGF[A] claims LOC[B]
  - ORGF[A] is wanting LOC[B]
- **directionality**
  - PER[A] is region[B]'s ruler
  - PER[A] is dictator of region[B]
- **antonymy**
  - LOC[A] is fighting with ORGF[B]
  - LOC[A] is allied with ORGF[B]
- **correlation**
  - ORGF[A] is seeking from ORGF[B]
  - ORGF[A] is giving to ORGF[A]

Positive (top) and negative (bottom) examples from SherLIiC-dev.

**Baselines**

<table>
<thead>
<tr>
<th>Baseline</th>
<th>F1 in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lemma</td>
<td>16.1</td>
</tr>
<tr>
<td>Database Lookup</td>
<td>44.4</td>
</tr>
<tr>
<td>Always yes</td>
<td>49.9</td>
</tr>
<tr>
<td>ESIM</td>
<td>53.1</td>
</tr>
<tr>
<td>word2vec</td>
<td>55.9</td>
</tr>
<tr>
<td>typed_rel_emb</td>
<td>50.8</td>
</tr>
<tr>
<td>untyped_rel_emb</td>
<td>57.2</td>
</tr>
</tbody>
</table>

**State of the Art**

<table>
<thead>
<tr>
<th>Baseline</th>
<th>F1 in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>w2v+tag_rel_emb</td>
<td>60.5</td>
</tr>
<tr>
<td>w2v+untyped_rel</td>
<td>60.0</td>
</tr>
<tr>
<td>w2v+typed_rel</td>
<td>59.4</td>
</tr>
<tr>
<td>TransE (typed)</td>
<td>49.8</td>
</tr>
<tr>
<td>TransE (untyped)</td>
<td>49.1</td>
</tr>
<tr>
<td>ComplEx (typed)</td>
<td>49.7</td>
</tr>
<tr>
<td>ComplEx (untyped)</td>
<td>49.3</td>
</tr>
</tbody>
</table>

MLU

Ludwig-Maximilians-Universität München

Linguistic Inference Benchmark (LIiC)