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

Martin Schmitt and Hinrich Schütze

CIS, LMU Munich

July 29, 2019
Natural Language Inference (NLI) is the task of recognizing \textbf{entailment, contradiction} or \textbf{neutrality} for a pair of sentences (Dagan et al., 2013; Williams et al., 2018).

\textbf{Example from SNLI (Bowman et al., 2015)}

\textbf{Premise:} Two men on bicycles competing in a race.

\textbf{Hypotheses:}

1. People are riding bikes. \textbf{entailment}
2. Men are riding bicycles on the street. \textbf{neutral}
3. A few people are catching fish. \textbf{contradiction}

- a lot of different linguistic phenomena involved
- role of lexical knowledge often neglected
  (Gururangan et al., 2018; Glockner et al., 2018)
Specifically: Verbal semantics in context

<table>
<thead>
<tr>
<th>Condition 1</th>
<th>Condition 2</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>If PERSON[A] is running ORG[B], then PERSON[A] is leading ORG[B].</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>If COMPUTER[A] is running SOFTWARE[B], then COMPUTER[A] is using SOFTWARE[B].</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>If COMPUTER[A] is running SOFTWARE[B], then COMPUTER[A] is leading SOFTWARE[B].</td>
<td>✗</td>
<td></td>
</tr>
</tbody>
</table>
Specifically: Verbal semantics in context

<table>
<thead>
<tr>
<th>If</th>
<th>PERSON[A]</th>
<th>is running</th>
<th>ORG[B],</th>
</tr>
</thead>
<tbody>
<tr>
<td>then</td>
<td>PERSON[A]</td>
<td>is leading</td>
<td>ORG[B].</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>If</th>
<th>COMPUTER[A]</th>
<th>is running</th>
<th>SOFTWARE[B],</th>
</tr>
</thead>
<tbody>
<tr>
<td>then</td>
<td>COMPUTER[A]</td>
<td>is using</td>
<td>SOFTWARE[B].</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>If</th>
<th>COMPUTER[A]</th>
<th>is running</th>
<th>SOFTWARE[B],</th>
</tr>
</thead>
<tbody>
<tr>
<td>then</td>
<td>COMPUTER[A]</td>
<td>is leading</td>
<td>SOFTWARE[B].</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>×</td>
</tr>
</tbody>
</table>

Compared to general NLI: controlled yet challenging

😊 Task: Binary entailment detection
😊 Abstract context with knowledge graph types (Freebase)
⚠️ Very similar sentences
⚠️ Distributional similarity of positive and negative examples
Components of SherLliC

**Typed event graph**

- United States of America
- Germany
- Barack Obama
- Angela Merkel

~190k typed relations between Freebase entities

~17M triples

**Inference candidates**

- PER be forced to leave LOC
- ⇒ PER leave LOC
- ORG be led by PER
- ⇒ PER manage ORG
- POLITICIAN meet with PER
- ⇒ POLITICIAN interact with PER

~960k pairs of Freebase-typed relations with high distributional overlap

**Annotated dev and test**

- ~4k annotated inference candidates
- split 25/75 in dev and test
Creation of SherLliC (I)

SherLliC Typed Event Graph (TEG)

- Preprocess the large entity-linked corpus ClueWeb09 with a dependency parser
- Extract shortest paths between entities in the dependency graphs ⇒ relations
- Type heterogeneous relations by finding largest typable subsets

United States of America 领先地位 → Barack Obama 与 Angela Merkel 会面

1. nsubj leader possess
   nsubj lead dobj
   nsubj chancellor prep of dobj

2. nsubj meet prep with dobj
   nsubj interact prep with dobj
   nsubj support dobj policy possess

Germany
Creation of SherLI\textsc{i}C (II)

SherLI\textsc{i}C-InfCands

- Score all pairs of relations according to statistical relevance, significance and entity overlap (distributional features)
- Best-scoring relations pairs become inference candidates

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

$$\text{Relv}(A, B) := \frac{P(B \mid A)}{P(B)}$$

$$\text{esr}(A, B) := \frac{\left| \bigcup_{i \in \{1, 2\}} \pi_i(A \cap B) \right|}{2 |A \cap B|}$$

$$\sigma(A, B) := 2 |A \cap B| \sum_{H \in \{B, \neg B\}} P(H \mid A) \log(\text{Relv}(A, H))$$
Creation of SherLIiC (III)

SherLIiC-dev and SherLIiC-test

- Annotate a random subset of SherLIiC-InfCands on Amazon Mechanical Turk
- Collect at least 5 annotations per InfCand
- Filter annotators with a qualification test and confidence values

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of annotated InfCands</td>
<td>3985</td>
</tr>
<tr>
<td>Balance yes/no</td>
<td>33% / 67%</td>
</tr>
<tr>
<td>Pairs with unanimous gold label</td>
<td>53.0%</td>
</tr>
<tr>
<td>Pairs with 1 disagreeing annotation</td>
<td>27.4%</td>
</tr>
<tr>
<td>Pairs with 2 disagreeing annotations</td>
<td>19.6%</td>
</tr>
<tr>
<td>Individual label = gold label</td>
<td>86.7%</td>
</tr>
</tbody>
</table>
Creation of SherLIiC (III)

SherLIiC-dev and SherLIiC-test

- Annotate a random subset of SherLIiC-InfCands on Amazon Mechanical Turk
- Collect at least 5 annotations per InfCand
- Filter annotators with a qualification test and confidence values

![Bar chart showing the number of annotations disagreeing with the majority.]

<table>
<thead>
<tr>
<th>Class label</th>
<th>Number of annotations disagreeing with the majority</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>0</td>
</tr>
<tr>
<td>yes</td>
<td>1, 2</td>
</tr>
</tbody>
</table>
Main Insights

- Knowledge graph embeddings do not capture necessary information
- Supervised NLI model ESIM is fooled by sentence similarity
- Best system combines word2vec and type-informed relation embeddings

<table>
<thead>
<tr>
<th>Baseline</th>
<th>P in %</th>
<th>R in %</th>
<th>F1 in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lemma</td>
<td>90.7</td>
<td>8.9</td>
<td>16.1</td>
</tr>
<tr>
<td>Always yes</td>
<td>33.3</td>
<td>100.0</td>
<td>49.9</td>
</tr>
<tr>
<td>TransE (typed)</td>
<td>33.3</td>
<td>99.1</td>
<td>49.8</td>
</tr>
<tr>
<td>ComplEx (typed)</td>
<td>33.7</td>
<td>94.9</td>
<td>49.7</td>
</tr>
<tr>
<td>ESIM</td>
<td>39.0</td>
<td>83.3</td>
<td>53.1</td>
</tr>
<tr>
<td>w2v+tsg_rel_emb</td>
<td>51.8</td>
<td>72.7</td>
<td>60.5</td>
</tr>
</tbody>
</table>
Analysis of entailment scores on SherLIiC-dev

<table>
<thead>
<tr>
<th>Model</th>
<th>Entailment</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>TransE (typed)</td>
<td>yes</td>
<td>1.0</td>
</tr>
<tr>
<td>ComplEx (typed)</td>
<td>yes</td>
<td>0.5</td>
</tr>
<tr>
<td>word2vec</td>
<td>no</td>
<td>0.0</td>
</tr>
<tr>
<td>typed_rel_emb</td>
<td>no</td>
<td>0.5</td>
</tr>
<tr>
<td>w2v+tsg_rel_emb</td>
<td>no</td>
<td>1.0</td>
</tr>
</tbody>
</table>
How can I use SherLliC?

SherLliC-TEG, -InfCands, -dev and -test are publicly available: https://github.com/mnschmit/SherLliC

Summary

- **SherLliC-TEG**: knowledge graph with event-like relations
  - combines knowledge graph relations with textual relations
  - large resource for relation inference in knowledge graphs

- **SherLliC-InfCand**: large collection of unlabeled samples
  - similar enough to SherLliC-dev and -test for transfer learning
  - contains noisy labels from best baseline w2v+tsg_rel_emb

- **SherLliC-test and -dev**:
  - Finetune on SherLliC-dev
  - Evaluate models of natural language inference (NLI) and/or lexical semantics
  - Evaluate relation and graph embedding techniques


Example sentence from Freebase-linked ClueWeb09

In Japan[03_3d], the PS3[067gh] is slowly closing the huge sales gap with Wii[026kds].
In m.067gh is slowly closing the huge sales with m.026kds

- m.067gh nsubj closing dobj gap prep with obj m.026kds
- m.067gh nsubj closing prep In obj m.03_3d
- m.03_3d obj In prep closing dobj gap prep with obj m.026kds
Typing the Event Graph (SherLliC-TEG)

**Type a relation $R$**

- For each argument slot, identify the $k$ entity types that induce the largest subsets.
- Consider the $k^2$ typed subrelations of $R$ constructed by restricting arguments to one of the types from the previous step.
- Accept a typed relation if it contains at least $\vartheta_{\min}$ entity pairs.

```
arg1
nsubj lead dobj

arg2
```
Supervised combination of typed and untyped embeddings

<table>
<thead>
<tr>
<th>If</th>
<th>is explaining in WRITTEN_WORK</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>then</td>
<td>is writing in WRITTEN_WORK</td>
<td>✓</td>
</tr>
</tbody>
</table>

detected by \(w2v+\text{typed\_rel}\), but not by \(w2v+\text{untyped\_rel}\)

\(w2v+\text{tsg\_rel\_emb}\)

**Idea:** Learn for each type signature if typed or untyped works better.

**Implementation:**
- Count on the dev set for each type signature how often typed or untyped embeddings are more accurate
- Pick typed or untyped according to these counts for seen type signatures
- For unseen type signatures, count individual types as well
Meta rule discovery

A Simple Meta Rule Algorithm

1. Consider InfCands where the premise or the hypothesis is contained by the other
2. Mask the common part in both relations by $X$
3. Simplify $\text{pobj}$, $\text{dobj}$ and $\text{iobj}$ to $\text{obj}$
4. Count patterns found in this way

Example

A is followed by $B$

$B$ is following $A$

$nsubjpass\_X\_prep\_by\_pobj \Rightarrow dobj\_X\_nsubj$

$nsubjpass\_X\_prep\_by\_obj \Rightarrow obj\_X\_nsubj$
### Example meta rules

<table>
<thead>
<tr>
<th>Rule Type</th>
<th>Rule</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>nsubj-X-poss</td>
<td>$\iff$ nsubj-X-prep-of-obj</td>
<td>$A$ is $B$’s ally $\iff$ $A$ is an ally of $B$</td>
</tr>
<tr>
<td>nsubjpass-X-prep-by-obj</td>
<td>$\iff$ obj-X-nsubj</td>
<td>$A$ is followed by $B$ $\iff$ $B$ follows $A$</td>
</tr>
<tr>
<td>nsubj-Xer-prep-of-obj</td>
<td>$\iff$ nsubj-X-obj</td>
<td>$A$ is a teacher of $B$ $\iff$ $A$ teaches $B$</td>
</tr>
<tr>
<td>nsubj-reX-obj</td>
<td>$\Rightarrow$ nsubj-X-obj</td>
<td>$A$ rewrites $B$ $\Rightarrow$ $A$ writes $B$</td>
</tr>
<tr>
<td>nsubj-agree-xcomp-X-obj</td>
<td>$\Rightarrow$ nsubj-X-obj</td>
<td>$A$ agrees to <em>buy</em> $B$ $\Rightarrow$ $A$ buys $B$</td>
</tr>
<tr>
<td>nsubjpass-force-xcomp-X-obj</td>
<td>$\Rightarrow$ nsubj-X-obj</td>
<td>$A$ is forced to <em>leave</em> $B$ $\Rightarrow$ $A$ leaves $B$</td>
</tr>
<tr>
<td>nsubj-decide-xcomp-X-obj</td>
<td>$\Rightarrow$ nsubj-X-obj</td>
<td>$A$ decides to <em>move to</em> $B$ $\Rightarrow$ $A$ moves to $B$</td>
</tr>
</tbody>
</table>