Softwareproject Topics

Katja Markert: Topics SS22

Computerlinguistik
Universität Heidelberg
Topic Suggestions

Variations
Most topics can be handled by more than one group via variations of method, language/domains or data. Every group can determine their focus (within reason) themselves. When two groups use the same data, they can also work as if in a “competition”.
Topic Suggestions

1. Topic Markert I: Semi-supervised learning for the automatic resolution of **metonymies**

2. Topic Markert II: Improving **unsupervised sentence summarization and headline generation** with regards to fluency and fidelity

3. Topic Markert III: **Comparative anaphora resolution** as question answering (no slides, if interest will explain on blackboard)
Markert1: Semi-supervised Learning for the Resolution of Metonymies


Frequent (every third sentence). Important for sentiment mining, text simplification, anaphora resolution, geographical IR...
Examples

Metaphors
Use a similarity relationship between two domains (ARGUMENT-IS-WAR)

- He **attacked** my arguments.
- He **bashed** my arguments.

Metonymies
Use a contiguity relation between two domains (PLACE-FOR-EVENT)

- He was traumatized after **Vietnam**
- **Pearl Harbour** still has an effect on our foreign policy

Both types tend to be systematic and generalize over groups of words
Prior Work and Task

Most work focuses on metaphor resolution → this software project is metonymy recognition

- He was traumatized after Vietnam → PLACE-FOR-EVENT
- Brazil lost the quarterfinal → PLACE-FOR-TEAM
- Brazil decided to stop deforestation → PLACE-FOR-GOV
- He lived in Tokyo → LITERAL
- BMW lost 3 points yesterday → ORG-FOR-INDEX
- He worked for IBM → LITERAL
## Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
<th>Type</th>
<th>Annot</th>
<th>literal</th>
<th>metos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semeval-LOC¹</td>
<td>BNC</td>
<td>Countries</td>
<td>Manual</td>
<td>1458</td>
<td>375</td>
</tr>
<tr>
<td>Semeval-ORG²</td>
<td>BNC</td>
<td>Companies</td>
<td>Manual</td>
<td>1211</td>
<td>721</td>
</tr>
<tr>
<td>ReLocar³</td>
<td>Wikipedia</td>
<td>Locations</td>
<td>Manual</td>
<td>995</td>
<td>1031</td>
</tr>
<tr>
<td>ConLL⁴</td>
<td>News</td>
<td>Locations</td>
<td>Manual noisy</td>
<td>4609</td>
<td>2448</td>
</tr>
<tr>
<td>WimCor⁵</td>
<td>Wikipedia</td>
<td>Locations</td>
<td>automatic</td>
<td>154322</td>
<td>51678</td>
</tr>
</tbody>
</table>

1, 2: Markert and Nissim, 2007  
3, 4: Gritta et al., 2017  
5: Mathews and Strube, 2020
State-of-the-Art: Li et al, 2020

Plus **masking of target word** in training and testing to avoid spurious information from rare target word occurrences:

He was traumatized by **Vietnam** → He was traumatised by **X**
Results Li et al (2020) (Accuracy)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>BL</th>
<th>BERT-BASE-MASK</th>
<th>BERT-LG-MASK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semeval-LOC</td>
<td>80.1%</td>
<td>87.1%</td>
<td>88.2%</td>
</tr>
<tr>
<td>Semeval-ORG</td>
<td>62.7%</td>
<td>75.6%</td>
<td>77.2%</td>
</tr>
<tr>
<td>ReLocar</td>
<td>50.8%</td>
<td>93.9%</td>
<td>94.4%</td>
</tr>
<tr>
<td>ConLL</td>
<td>65.3%</td>
<td>93.7%</td>
<td>93.9%</td>
</tr>
<tr>
<td>WimCor</td>
<td>74.9%</td>
<td>95.4%</td>
<td>95.5%</td>
</tr>
</tbody>
</table>
This does not look too bad: what’s the problem?

• Worst results on manually annotated datasets with diversity and **natural distribution**

• **Cross-domain accuracies** much lower: WimCor → Semeval 78.4% (worse than BL), WimCor → ReLocar 64.6%

• Ignores important target word information: Vietnam vs. Solomon Islands as PLACE-FOR-EVENT?

<table>
<thead>
<tr>
<th>Country</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greenland</td>
<td>4/100</td>
</tr>
<tr>
<td>Guyana</td>
<td>5/100</td>
</tr>
<tr>
<td>Japan</td>
<td>18/100</td>
</tr>
<tr>
<td>Hungary</td>
<td>21/100</td>
</tr>
</tbody>
</table>

But good target word info not easy to integrate with such small datasets
This does not look too bad: what’s the problem?

- Worst results on manually annotated datasets with diversity and **natural distribution**
- **Cross-domain accuracies** much lower: WimCor → Semeval 78.4% (worse than BL), WimCor → ReLocar 64.6%
- Ignores important target word information: **Vietnam vs. Solomon Islands** as PLACE-FOR-EVENT?
  
  - Greenland: 4/100
  - Guyana: 5/100
  - ... ...
  - Japan: 18/100
  - Hungary: 21/100

But good target word info not easy to integrate with such small datasets
Semi-supervised learning for Figurative Language

Currently

Almost all work on metaphor or metonymy recognition is fully supervised. As especially the manually annotated metonymy datasets are small, this is a problem.

Recent exception for metaphor: CATE (Lin et al., EMNLP 2021): Use of self-training!
CATE’s approach

- Fine-tuning (and test) data: VUA metaphor corpus (BNC)
- Two contributions: contrastive objective (Stage I) plus self-training (Stage II)
Self-Training

Example for generated metaphor data in self-training

<table>
<thead>
<tr>
<th>Distant supervision instances</th>
<th>pseudo-label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Does it really take seven years to <em>digest</em> gum?</td>
<td>literal</td>
</tr>
<tr>
<td>2. This <em>food</em> is supposed to be easy to <em>digest</em>.</td>
<td>literal</td>
</tr>
<tr>
<td>3. Congee is also considered an ideal <em>food</em> for babies, as it is easily eaten and <em>digested</em>.</td>
<td>literal</td>
</tr>
<tr>
<td>4. <em>the substance</em> of his London lectures is full of the latest information well <em>digested</em>.</td>
<td>metaphorial</td>
</tr>
<tr>
<td>5. Helen Reddy’s voice is easily accepted and <em>digested</em> by large groups of <em>people</em>.</td>
<td>metaphorial</td>
</tr>
</tbody>
</table>

Picture from Lin et al (2021)

- Self-training has the problem of error propagation: CATE’s solution is soft-labeling
- They show that self-training already helps for metaphor even without contrastive objective (simpler Stage 1)
Problems

- Only focuses on target word for dataset expansion, never the context
- Not used for metonymy
- No attempt to match labeled and unlabeled data domain (BNC $\neq$ Wikipedia)
- Only one semi-supervised paradigm
Markert I.1: Metonymy recognition with self-training

- Same self-training with soft labels on metonymies
- Expanded with domain matching (SemEval uses BNC unlabeled examples, Conll news etc).
- Include both context and target word in generating strategy:
  
  He was traumatized by *Pearl Harbour*

**Target word-based**

The attack on *Pearl Harbour*

The consequences of *Pearl Harbour*

... 

**Context-based**

Americans had been traumatized by *Vietnam*

Traumatized by *Madrid*, Pochettino can't sleep anymore

...
Markert I.2: Metaphor/Metonymy Recognition with Active Learning

- Selection strategy crucial: often use examples where classifier is uncertain
- Advantage: added new data not noisy as human in the loop
- Can be simulated by holding out parts of training data as unlabeled data, if you don’t want to annotate anything
Resources and Literature

- All mentioned metonymy/metaphor data is publically available.
MarkertII: Sentence summarization/Headline Generation

The problem
Shorten a sentence or generate a headline from a news sentence, given a target length for the shortened sentence/headline

Example Pair

- ORIG: The world’s biggest miner BHP Billiton announced Tuesday it was dropping its controversial hostile takeover bid for rival Rio Tinto due to the state of the global economy.
- HUMAN REFERENCE SUMMARY: BHP Billiton drops Rio Tinto takeover bid
Supervised vs. Unsupervised Methods

**SUPERVISED**
- Many pairs given
- Seq2Seq models

**UNSUPERVISED**
- No pairs given
- Source text maybe given
- Target text maybe given
Schumann, Lou and Markert (ACL 2020): Unsupervised

• Word-level Extraction
• Greedy Hill-climbing with restarts
Schumann et al: Objective Function

- Source Sentence $x = (x_1, x_2, \ldots, x_n)$
- Output Sentence $y = (y_1, y_2, \ldots, y_m)$
- $s < n$ summary upper bound
- Objective function $f$ maximises for fluency and similarity

$$f(y; x, s) = f_{\text{LM}}(y) \cdot f_{\text{SIM}}(y; x)^\gamma \cdot f_{\text{LEN}}(y; s), \quad (1)$$

- Fluency was measured via inverse perplexity of LSTMs trained on source or target sentences
- Similarity between $y$ and $x$ was measured by Sent2vec
Results

- State-of-the-art at the time for ROUGE score
- Human evaluation with 5 annotators via comparison to previous best models on 100 instances via fidelity and fluency

<table>
<thead>
<tr>
<th>Models</th>
<th>Score (#wins/#ties/#loses)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fidelity</td>
</tr>
<tr>
<td>HC vs. WL</td>
<td>+0.18 (44/30/26)</td>
</tr>
<tr>
<td>HC vs. ZR</td>
<td>+0.05 (35/35/30)</td>
</tr>
</tbody>
</table>

Table: Human evaluation in a pairwise comparison setting on 100 headline generation instances.
Schumann et al: Example output

Good Example

• mubarak was ousted friday after being at the helm of his north african country for nearly 30 years .
• mubarak ousted after being at the helm of his country for years

Bad Example

• A third national security bill has been introduced to allow sharing of information between intelligence agencies and the Australian defence forces , allowing them to potentially target Australian terrorist fighters .
• bill introduced to allow sharing of information between intelligence agencies and terrorist
Schumann et al: Non-comparative performance analysis

- Preliminary annotation study of fluency and fidelity by Eric Kaiser (http://misc.eric-kaiser.net/annotation)
- 266 annotation
- Fidelity
  - Fidelity correct 39.9%
  - Fidelity incorrect 60.2%
- Fluency
  - 1 18.8%
  - 2 10.9%
  - 3 9.8%
  - 4 16.5%
  - 5 44.0%
Markert II Project Ideas: Improve Fluency and/or Fidelity

- Better language models to improve fluency
- Use semantic graph matching methods (such as AMR scoring) as an objective function to improve fidelity

Figure 1: Similar AMRs, with sketched alignments.

Picture from Opitz et al. (2021)
Markert II Project Ideas: Improve Fidelity

The problem occurs in standard single-document summarization:

Figure from Pagnoni et al (2021) on 250 articles and their summaries

Idea
Adapt a suitable factual consistency evaluation metric from standard document summarization, such as FactCC (Kryscinski et al (2019))
### Ressources and Literature

- Raphael Schumann’s code exists and runs
- Gigaword headline generation dataset: [https://github.com/harvardnlp/NAMAS](https://github.com/harvardnlp/NAMAS)