Softwareproject Topics

Katja Markert: Topics WS22

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: Data augmentation for the automatic resolution of metonymies
2. Topic Markert II: Saints-Memory: Matching saints in a German historical encyclopedia to German Wikipedia
Markert1: Data Augmentation for the Resolution of Metonymies

“**Trope:** [. . .] jede Form der Rede, die das Gemeinte nicht direkt und sachlich durch das eigentl. Wort ausspricht, sondern [. . .] durch e. Anderes, Naheliegendes, e. “übertragenen” Ausdruck wiedergibt.”

*Gero von Wilpert (1989): Sachwörterbuch der Literatur*

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-LOC(^1)</td>
<td>BNC</td>
<td>Countries</td>
<td>Manual</td>
<td>1458</td>
<td>375</td>
</tr>
<tr>
<td>Semeval-ORG(^2)</td>
<td>BNC</td>
<td>Companies</td>
<td>Manual</td>
<td>1211</td>
<td>721</td>
</tr>
<tr>
<td>ReLocar(^3)</td>
<td>Wikipedia</td>
<td>Locations</td>
<td>Manual</td>
<td>995</td>
<td>1031</td>
</tr>
<tr>
<td>ConLL(^4)</td>
<td>News</td>
<td>Locations</td>
<td>Manual noisy</td>
<td>4609</td>
<td>2448</td>
</tr>
<tr>
<td>WimCor(^5)</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%
- Overfitting to Training Set
- Realistic datasets are too small and often too unbalanced
Current learning for Figurative Language

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

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

Recent exception for metonymy: SWP Summer 2022: use of self-training with promising results
Semi-supervised learning vs. Data Augmentation

Self-Training

Data Augmentation: Add variations to input data (or feature-space) that are label-preserving or predicably label-reversing

By Backtranslation
IBM rose 4 points yesterday. → IBM stieg gestern um 4 Punkte. → IBM increased by 4 points yesterday.

Data Augmentation Possibilities

Figure 1: Taxonomy and grouping for different data augmentation methods.

Examples of data space data augmentation methods for metonymy recognition

Starting from *IBM rose 4 points yesterday.*

- By Backtranslation from/to German: *IBM increased by 4 points yesterday.*
- Grammar Variations: *IBM rises 4 points yesterday*
- Antonym or Synonym substitution: *IBM gained 4 points yesterday*
- Entity transformation: *IBM Corp. rose 4 points yesterday*
- Yoda Transformation: *Rose by 4 points yesterday, IBM did.*
- Noise transformations: *IBM rose 4 pints yesterday.*

Starting from *IBM shares rose 5 points yesterday* (literal) \(\rightarrow\) *IBM rose 5 points yesterday* (metonymy)
Feature-space data augmentation

Interpolation with SMOTE or Mix-Up by interpolating, for example BERT layers:

\[
\tilde{x} = \lambda x_i + (1 - \lambda)x_j
\]

\[
\tilde{y} = \lambda y_i + (1 - \lambda)y_j
\]

Figure 5: Illustration of the interpolation method SMOTE.

Advantages and challenges of Data Augmentation

• Not dependent on original classifier and its quality
• Many variations
• Tradeoff between meaning preservation and diversity, especially if your baseline model is a large language model
Challenges and Possibilities

- Identify label-preserving and predictably label-inverting transformations for metonymy
- Explore data augmentation strategies implemented in the NL-Augmenter (Dhole et al, 2021) to improve metonymy resolution
- Add own data augmentation strategies to the NL-Augmenter
- Enhance diversity of augmentation strategies

Challenges:
- Use knowledge gained in modules such as Statistical Methods, Syntax, Semantics to come up with sensible data augmentation strategies
- Implement or re-implement such strategies
- Integrate with a standard language model metonymy resolution baseline
Resources and Literature: data augmentation


Resources and Literature: Metonymy


Saints-Memory

https://encycnet.github.io/: aims to create a new semantic resource for historical German in form of a knowledge graph.

- Currently 22 historic German encyclopedias
- Attempt to use DBSpotlight to automatically match entries to DBPedia (and GermaNet)

<table>
<thead>
<tr>
<th></th>
<th>Germanet</th>
<th>DBPedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brockhaus 1809</td>
<td>32.80</td>
<td>51.15</td>
</tr>
<tr>
<td>Eisler Philosophie 1904</td>
<td>46.06</td>
<td>25.40</td>
</tr>
<tr>
<td>Wander Sprichtwort 1867</td>
<td>43.75</td>
<td>16.06</td>
</tr>
<tr>
<td>Roell Eisenbahnen 1912</td>
<td>24.91</td>
<td>27.88</td>
</tr>
<tr>
<td>Heiligenlexikon 1858</td>
<td>2.34</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Use cases:
- Finding gaps in Wikipedia
- Finding mismatched information
S. Bilhildis, (27. Nov.), Wittwe und Stifterin des Klosters Altmünster (<i>Altum Monasterium B. V. M</i>.) war die Tochter christlicher Eheleute von vornehmer Abkunft, Namens Iberius und Mechildis (Mechtildis, Mathildis) und wurde zu Hochheim am Main um das Jahr 625 oder 626 geboren. Was dieß für ein Hochheim am Main sei, ob der nicht weit von Wirzburg gelegene Ort, gewöhnlich Veitshöchheim genannt, ... 

Von ihrer Base zu Wirzburg in aller Gottseligkeit erzogen, ward sie in jungen Jahren, etwa 16 oder 17 Jahre alt, an den heidnischen Herzog Hettan (in Thüringen) vermählt, ...

Die Zeit, wann sie das Zeitliche segnete, ist nicht zu ermitteln; Einige jedoch setzen ihren Tod in das Jahr 630. (<i>El., Buc</i>.)
Bild von Joachim Schäfer - https://www.heiligenlexikon.de Ökumenisches Heiligenlexikon, Creative Commons CC BY-NC-SA 4.0
That was the simple case...

- Unambiguous and successful name match in Wikipedia
- Matching of feast day ("Gedenktag")
- Even then we can see: name variations *Hettan* - *Heden*, different birth or death dates, uncertain information on very early saints
Normally...

- Obscure and unmatched saints:
  
  `<b>Boderius</b>, (22. Mai), wird in einigen Orten als Martyrer verehrt`

- Several names and several or ambiguous feast days

- Very ambiguous saints where name and day is not enough
  
  1. Bernardus (1): cistercian, abbot of Clairveaux, approx. 1100
  2. Bernardus (2): arch bishop of Vienne died 842, also named Barcar, or Barnar
  3. Bernardus (3): bishop of Carinola, died 1109
  4. . .
  5. Bernardus (64): simple monk of the Capucines, died 1540

- Normally no infoboxes in Wikipedia
Project Idea

- Define information extraction templates for saints

<table>
<thead>
<tr>
<th>saint</th>
<th>relation</th>
<th>possible filler</th>
</tr>
</thead>
<tbody>
<tr>
<td>saint</td>
<td>has-name</td>
<td>any string</td>
</tr>
<tr>
<td>saint</td>
<td>feast-day</td>
<td>date</td>
</tr>
<tr>
<td>saint</td>
<td>has-job</td>
<td>martyr, abbot, bishop, arch-bishop</td>
</tr>
<tr>
<td>saint</td>
<td>is-born</td>
<td>date</td>
</tr>
<tr>
<td>saint</td>
<td>has-died</td>
<td>date</td>
</tr>
<tr>
<td>saint</td>
<td>located-in</td>
<td>location</td>
</tr>
</tbody>
</table>

- Algorithms for template filling inspired by IE work for Heiligenlexikon → Template 1
- Wiki API for approximate name match plus filtering
- Template filling from Wikipedia matches and Wikidata → Template(s) 2
- Template match
- Evaluation
Template filling

- There is no gold data, so unsupervised and probably no or little standard machine learning
- Programming from scratch
- Possibility:
  1. Preprocessing with Heideltime and German Stanford Core NLP
  2. Template 1: Some relations can be filled by (approximate) regular expression match and tag restrictions (names, job titles, feast days)
  3. Template 2: REs or Wikidata
     https://www.wikidata.org/wiki/Q477895
  4. Perform simple, unambiguous saint matches
  5. Bootstrapping for natural language patterns from these seeds: iterative algorithm
  6. Enhanced with semantic similarity of texts
Extensions

- Some relations are shared with standard IE on news. Use existing algorithms trained on news for domain transfer.
- Use additional resources such as Wikipedias in different languages or https://www.heiligenlexikon.de/
Resources and Literature

- Encyc-Net for the Heiligenlexikon: https://encycnet.github.io/