ANEACL: Automated (Named) Entity Annotation for German Domain-Specific Texts

Anastasia Zhukova, Felix Hamborg, Bela Gipp

2nd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE2021)
30 September 2021
Introduction
Introduction

- Named entity recognition (NER) is a well-known NLP task.
- NER datasets contain general categories, e.g., person, location, time, etc.

Problems
1. General NER reflects no categories of the other domains, e.g., technology, production
2. A small number of NLP datasets for German, i.e., a low-resources language
3. Domain NER requires annotating a dataset for training a NER model
   → a very time-consuming task

Goal
- Minimize the time of creating a domain dataset for NER in German by automating the annotation process
Research question

How to use knowledge graphs (e.g., Wiktionary) to automatically
1. extract domain terms (nouns),
2. derive entity categories,
3. annotate these terms into categories?

Zhukova et al. “ANE: Automated (Named) Entity Annotation for German Domain-Specific Texts”
Domain graph
German compound nouns

Sechszylindermotor (six-cylinder motor) = sechs + Zylinder + Motor

A Wiktionary page (WP) matches “Motor”:

dede.wiktionary.org/wiki/Motor

Motor

Bedeutungen: Senses and definitions
[1] Technik: antreibende Maschine
[2] übertragen: etwas oder jemand, der etwas voranbringt; Antreiber, treibende Kraft

Oberbegriffe: Hypernyms
[1] Fahrzeug, Technik

Unterbegriffe: Hyponyms
[1] Abüftmotor, Abtriebsmotor, Aluminiummotor, Antriebsmotor, Außenbordmotor, Austauschmotor, Automotor, Backbordmotor,

Properties from WPs: (1) Hypernyms, (2) Hyponyms, (3) Definitions and areas

→ Use the NPs that were mapped to Wiktionary pages

Zhukova et al. “ANEA: Automated (Named) Entity Annotation for German Domain-Specific Texts”
**Domain graph**

*Domain graph* is a *Wiktionary subgraph* with nodes related to the current domain.

Leaves are the domain terms. Nodes are candidate labels for the entity categories.

Zhukova et al. “ANEA: Automated (Named) Entity Annotation for German Domain-Specific Texts”
Candidate Entity Categories

Specific
(Actor, Politician)

Optimal
(Person)

General
(Humanity)

Word embeddings:
fastText represents well out-of-vocabulary terms and labels

The quality metric:

\[ Q_i = T_i \cdot L_i \cdot O_i \cdot \max(\log_2 |EC_i|, 1) \cdot d_{avg_i} \]

\( T_i \) is a mean cross-term cosine similarity

\( L_i \) is a mean label-terms cosine similarity

\( O_i = T_i + L_i \) is an overall similarity

\(|EC_i|\) is a number of terms in an entity category

\( d_{avg_i} \) is an average of non-zero distances between terms and a label
Optimization steps

1. Candidate filtering
   1) a mean cross-term similarity too small \( (T_i < 0.2) \)
   2) a mean label-terms similarity too small \( (L_i < 0.3) \)
   3) EC is too broad (contains > 15% of all terms-to-annotate)
   4) EC is too narrow (contains < 5 terms)

2. Resolution of full overlaps
   When containing same terms, keep an \( EC_i \) with the largest \( Q_i \)

Zhukova et al. “ANEA: Automated (Named) Entity Annotation for German Domain-Specific Texts”
3. Resolution of the substantial overlaps
When terms overlap ≥ 50%, keep a big $EC_i$ that is a best replacement to a small $EC_j$ and to itself

$$arg \max r_h = A \text{ and } arg \max r_A = A$$

$$arg \max r_F = D \text{ but } arg \max r_D = D$$

4. Resolution of the conflicting terms
Resolve the conflicting terms to “clean” candidate $EC_i$ with the highest overall similarity $O_i$

Remaining conflicting terms
- $F$: $a \rightarrow S_A > S_F \rightarrow A$
- $b \rightarrow S_A > S_F \rightarrow A$
- $c \rightarrow S_A > S_F \rightarrow A$
- $h \rightarrow S_A > S_F \rightarrow A$
- $f \rightarrow S_F > S_C \rightarrow F$
- $g \rightarrow S_F > S_C \rightarrow F$
- $i \rightarrow S_F > S_C \rightarrow F$

“clean” ECs

$C$: [f, g, i]
$A$: [a, b, c, h]
$F$: [d, e]

Conflicting terms

$C$: [f, g, i]
$A$: [a, b, c, h]
$F$: [a, b, c, d, e, h, f, g, i]
Evaluation
No German domain-specific dataset available → performed a user study to evaluate the results

- **4 datasets**: processing industry, software development, databases, and travelling
- **9 native German study participants**: 4 f, 5 m, aged between 23-60
- **2-4 evaluators** per dataset
- **4 various configurations** per dataset: different number of terms-to-annotate
- **2 methods**: ANEA and a hierarchical clustering baseline

**Tasks**
1) Evaluate *cross-term relatedness* within a category:
   0-9 where 9 is the best
2) Evaluate *relatedness of a label to terms* in a category:
   0-9 where 9 is the best

Zhukova et al. “ANEA: Automated (Named) Entity Annotation for German Domain-Specific Texts”
• *Distribution* of the relatedness scores between the datasets *differ*.

• The most frequent score per dataset is used as *thresholds* for creating *silver datasets*.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Databases</th>
<th>Software development</th>
<th>Traveling</th>
<th>Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>All words</td>
<td>8161</td>
<td>8581</td>
<td>6293</td>
<td>7984</td>
</tr>
<tr>
<td>Terms</td>
<td>1209</td>
<td>1041</td>
<td>1040</td>
<td>552</td>
</tr>
<tr>
<td>Heads</td>
<td>713</td>
<td>673</td>
<td>801</td>
<td>328</td>
</tr>
<tr>
<td>Assessors</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

Zhukova et al. “ANEA: Automated (Named) Entity Annotation for German Domain-Specific Texts”
Silver datasets are required to compare configurations of ANEA against it.

Zhukova et al. “ANEA: Automated (Named) Entity Annotation for German Domain-Specific Texts”
Evaluation methods

- Silver dataset
- Hierarchical clustering
  - A baseline for terms relatedness
- ANEA
- ANEA voting
  - A final result is derived in an ensemble/voting strategy of multiple ANEA configurations
### Results

<table>
<thead>
<tr>
<th>Topic</th>
<th>Method</th>
<th># terms-to-annotate</th>
<th># entity categories (ECs)</th>
<th># annotated ECs’ average size</th>
<th>Term similarity</th>
<th>Label similarity</th>
<th>Average similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Databases</td>
<td>silver</td>
<td>420</td>
<td>5</td>
<td>113</td>
<td>23</td>
<td>7.2</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>HC</td>
<td>253</td>
<td>8</td>
<td>52</td>
<td>7</td>
<td>7.2</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>ANEA</td>
<td>253</td>
<td>18</td>
<td>179</td>
<td>10</td>
<td>5.7</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>ANEA voting</td>
<td>253-316</td>
<td>12</td>
<td>122</td>
<td>10</td>
<td>6.3</td>
<td>5.9</td>
</tr>
<tr>
<td>Software dev.</td>
<td>silver</td>
<td>356</td>
<td>6</td>
<td>57</td>
<td>10</td>
<td>6.2</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>HC</td>
<td>303</td>
<td>15</td>
<td>152</td>
<td>10</td>
<td>5.5</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>ANEA</td>
<td>191</td>
<td>10</td>
<td>119</td>
<td>12</td>
<td>5</td>
<td>5.3</td>
</tr>
<tr>
<td></td>
<td>ANEA voting</td>
<td>191-255</td>
<td>4</td>
<td>44</td>
<td>11</td>
<td>5.6</td>
<td>6.5</td>
</tr>
<tr>
<td>Traveling</td>
<td>silver</td>
<td>363</td>
<td>6</td>
<td>115</td>
<td>19</td>
<td>7.8</td>
<td>6.7</td>
</tr>
<tr>
<td></td>
<td>HC</td>
<td>363</td>
<td>19</td>
<td>156</td>
<td>8</td>
<td>7.3</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>ANEA</td>
<td>363</td>
<td>22</td>
<td>239</td>
<td>11</td>
<td>5.4</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>ANEA voting</td>
<td>258-363</td>
<td>12</td>
<td>146</td>
<td>12</td>
<td>6.2</td>
<td>5.6</td>
</tr>
<tr>
<td>Processing</td>
<td>silver</td>
<td>282</td>
<td>7</td>
<td>102</td>
<td>15</td>
<td>6.6</td>
<td>6.2</td>
</tr>
<tr>
<td></td>
<td>HC</td>
<td>183</td>
<td>7</td>
<td>56</td>
<td>8</td>
<td>6.1</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>ANEA</td>
<td>227</td>
<td>16</td>
<td>172</td>
<td>11</td>
<td>5.3</td>
<td>4.9</td>
</tr>
<tr>
<td></td>
<td>ANEA voting</td>
<td>181-282</td>
<td>9</td>
<td>157</td>
<td>17</td>
<td>5.7</td>
<td>5.6</td>
</tr>
</tbody>
</table>

ANEA voting shows **improvement of 13-15%** to the original ANEA average similarity scores.

Zhukova et al. “ANEA: Automated (Named) Entity Annotation for German Domain-Specific Texts”
• ANEA hasn’t achieved the relatedness scores of the silver datasets yet.
• The voting strategy shows a significant improvement to the ANEA results

Recommended configurations for ANEA with voting:
1) \( y = 158 + 0.167x \)
2) \( y + 50 \)
3) \( y - 50 \)

where \( x \) is a number of unique heads among the terms to annotate and \( y \) is a number of terms-to-annotate by ANEA

Zhukova et al. “ANEA: Automated (Named) Entity Annotation for German Domain-Specific Texts”
Conclusion
Conclusion

• Proposed ANEA, i.e., an unsupervised approach for automated creation of a small dataset for domain-specific NER.
• Evaluated ANEA with a user study on four domain datasets.
• The produced entity categories required less than one hour, which is significantly faster than manual annotation.
• The produced entity categories are slightly worse than the silver datasets but a voting strategy improves the scores by 13-15%.

A suggested use case with using ANEA:
(1) annotate a small dataset,
(2) validate and improve the dataset with manual inspection,
(3) use the produced dataset in a semi-supervised or transfer learning
Anastasia Zhukova
zhukova@uni-wuppertal.de