AIDA-light: High-Throughput Named-Entity Disambiguation

Ba Dat Nguyen
Johannes Hoffart
Martin Theobald
Gerhard Weikum

Max-Planck-Institut für Informatik
Saarbrücken, Germany
Overview

• Named Entity Disambiguation

• High-performance Accurate Entity Disambiguation
  ▪ Simplifying Expensive Features
  ▪ Categories and Domains
  ▪ Multi-phase Computation

• Experiments
Named Entity Disambiguation (NED)

NED aims to map mentions of ambiguous names in natural language onto a set of known entities (e.g. YAGO or DBpedia).

**Text & Mentions**

Under Fergie, United won the Premier League title 13 times.

- **Fergie (singer)**, an American singer, songwriter, fashion designer, television host and actress.
- **Alex Ferguson**, a former Scottish football manager of Manchester United F.C.
- **Sarah, Duchess of York**, the former wife of Prince Andrew, Duke of York.
- ...

- **United Airlines**, an American major airline.
- **United Airways**, a Bangladeshi airline.
- **Manchester United F.C.**, an English professional football club.
- ...

- **Premier League**, the English professional football league.
- ...

**correct entities**
State-of-the-art NED Systems

• Accurate Systems:
  ▪ AIDA and Illinois Wikifier: use rich contextual features (and joint inference) → emphasis on quality.

• High-performance Systems:
  ▪ DBpedia Spotlight and TagMe: mention-by-mention inference with more lightweight features → emphasis on speed.
AIDA-light

• **Goal**: reconcile efficiency and accuracy.

• **Approach**:
  - simplify expensive features.
  - add novel features with low footprint.
  - multi-phase computation.
Joint Inference over Disambiguation Graph

- Construct an undirected weighted graph between mentions and entities.
- Compute the best joint mapping sub-graph.
Simplify Expensive Features

• **Key-phrases (AIDA):** link anchor texts including categories, citation titles, and external references.

• **Key-tokens:** extracted from all key-phrases except stop words.

• **Example:**
  - AIDA key-phrases: “*U.S. President*”, “*President of the U.S.*”
  - AIDA-light key-tokens: “*President*”, “*U.S.*”
Categories and Domains

- Entity, Categories and Domains
  - For example:
    - Entity: Premier_League
      → Category: Football_Leagues
      → ... → Domain: Football

- Domain-Entity Coherence
  A entity belongs to a domain if it belongs to at least one category of the domain → recompute the mention-entity edge’s weight under the domain.

- Entity-Entity Coherence
  connect entities from the same domain → give higher weight to same-domain entity-entity coherence edges.
Multi-phase Computation

- "Easy" mentions: mentions with very few candidates or with skewed distributions.
- Update the context by chosen entities (with domains).
- Better understanding of the context.
- Reduce the complexity of the later stages.
Experimental Setup

- **Systems under comparison:**
  - AIDA-light
  - AIDA
  - DBpedia Spotlight

- **Performance measures:**
  - All systems take the same mentions as the input.
  - Each mention is mapped to one entity in DBpedia ∩ YAGO.
  - Mapping a mention of in-KB entity to null is a failure.

We apply per-mention precision.
Experimental Corpora

- **CoNLL-YAGO testb**: news articles with long-tail entities.
- **WP**: short contexts with highly ambiguous mentions and long-tail entities.
- **Wikipedia articles**: Wikipedia articles with internal links as mentions.
- **Wiki-links**: long documents with a few mentions.
Results on NED Quality

- Precision on different corpora, statistically significant improvements over Spotlight are marked with an asterisk.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AIDA</th>
<th>AIDA-\textit{light}</th>
<th>Spotlight</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoNLL-YAGO</td>
<td>82.5%*</td>
<td>84.8%*</td>
<td>75.0%</td>
</tr>
<tr>
<td>WP</td>
<td>84.7%*</td>
<td>84.4%*</td>
<td>63.8%</td>
</tr>
<tr>
<td>Wikipedia articles</td>
<td>90.0%</td>
<td>88.3%</td>
<td>89.6%</td>
</tr>
<tr>
<td>Wiki-links</td>
<td>80.3%</td>
<td>85.1%*</td>
<td>80.7%</td>
</tr>
</tbody>
</table>
Results on Run-time

- Average per-document run-time results.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AIDA-light</th>
<th>Spotlight</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoNLL-YAGO</td>
<td>0.47s</td>
<td>0.51s</td>
</tr>
<tr>
<td>WP</td>
<td>0.05s</td>
<td>0.14s</td>
</tr>
<tr>
<td>Wikipedia articles</td>
<td>5.47s</td>
<td>4.22s</td>
</tr>
<tr>
<td>Wiki-links</td>
<td>0.18s</td>
<td>0.32s</td>
</tr>
</tbody>
</table>

- **AIDA** uses a SQL database, not considered here.
Conclusion

• A high-performance accurate NED system
  ▪ First method to consider domain coherence.
  ▪ Judicious choice of high benefit/cost features.

• Experiments: AIDA-light
  ▪ as good as rich-feature systems.
  ▪ as efficient as fastest systems.
AIDA-light source code is available to download at
https://www.mpi-inf.mpg.de/yago-naga/aida/

Thanks!