CS 512 Project Presentation

Academic Concept Extractor

Our 2-phase framework to type and extract entities from scientific titles

Adit Krishnan, Aravind Sankar, Shi Zhi (mentor)
Our objective:

Extract Typed Entities corresponding to technique (solution) and application (problem) to summarize key contributions of a scientific article or publication.

**FacetGist** – Types all entity mentions within a paper.
- Hard to rank entity mentions by relevance
- May obtain several irrelevant mentions

<table>
<thead>
<tr>
<th>Technique</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>conditional random field; unsupervised learning; support vector machine; hidden markov model</td>
<td>Document summarization; sequence labeling; statistical classification</td>
</tr>
</tbody>
</table>

**Bootstrapping approach** to type entity mentions
- Depends on noun phrase chunker. Poor quality entity extraction.
- Requires manual effort to create seed set for each domain.
Titles are useful:

- **APPLICATION/PROBLEM MODIFIER**
- **APPLICATION/SOLUTION MODIFIER**
- **PROBLEM MODIFIER**

<table>
<thead>
<tr>
<th>APPLICATION/PROBLEM ENTITIES</th>
<th>RELATION PHRASES</th>
<th>TECHNIQUE/SOLUTION ENTITIES</th>
</tr>
</thead>
<tbody>
<tr>
<td>solving consensus</td>
<td>and semi supervised clustering problem</td>
<td>by using nonnegative matrix factorization</td>
</tr>
</tbody>
</table>

Titles are chosen by the authors to concisely represent their key contributions in the article or publication. Over 95% of all titles in our dataset contain at least one entity of interest.
• **Phase 1 - Coarse Typing**: Entity Phrase “Nonnegative matrix factorization” must be typed as technique or application for this title

• **Phase 2 - Fine Grained Modifier and Entity Extraction**: Matrix and factorization are bad entities, because matrix factorization is atomic.

---

**PhraseType**

• Our solution to coarse typing.

• Input titles are segmented into entity and relation phrases using POS tags.

• Probabilistic generative model to type entity phrases.

**Adaptor Grammar based extraction**

• Rule based extraction

• Data Driven, does not require human annotation effort. Domain independent.

• Few simple universal grammar rules.
Pipeline of our algorithm:


<table>
<thead>
<tr>
<th>PHRASE</th>
<th>UNIGRAM</th>
<th>SIGNIFICANT PHRASES</th>
<th>LEFT RP</th>
<th>RIGHT RP</th>
</tr>
</thead>
<tbody>
<tr>
<td>word embedding</td>
<td>word, embedding</td>
<td>word embedding</td>
<td>None</td>
<td>based</td>
</tr>
<tr>
<td>generalized language</td>
<td>generalized, language, language model</td>
<td>generalized language, language model</td>
<td>based</td>
<td>for</td>
</tr>
<tr>
<td>information retrieval</td>
<td>information, retrieval</td>
<td>information retrieval</td>
<td>for</td>
<td>None</td>
</tr>
</tbody>
</table>

TECHNIQUE
- ENTITY
  - Word embedding

TECHNIQUE
- MOD
  - ENTITY
  - generalized language model

APPLICATION

<table>
<thead>
<tr>
<th>PHRASE</th>
<th>APPLICATION PROBABILITY</th>
<th>TECHNIQUE PROBABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>word embedding</td>
<td>0.37</td>
<td>0.63</td>
</tr>
<tr>
<td>generalized language model</td>
<td>0.39</td>
<td>0.61</td>
</tr>
<tr>
<td>information retrieval</td>
<td>0.73</td>
<td>0.27</td>
</tr>
</tbody>
</table>
Phrase Type Model:

Time Slice - 1  Time Slice - 2  Time Slice - 3
Model Details:

2-topic model, Technique and Application

Each topic is defined by distributions over:

1) Textual Features – Words, Significant Phrases (extracted by TopMine from the corpus)

2) Relation Phrases – Right and Left Relation Phrases

Inferencing:

MCMC Algorithm – Gibbs Sampling until convergence
Data driven phrase parsing:

- Biterm language model for document retrieval
  - TECHNIQUE
    - MOD
      - Biterm language Model
  - APPLICATION
    - ENTITY
      - document retrieval

- Passage retrieval based on language model
  - APPLICATION
    - ENTITY
      - passage retrieval
  - TECHNIQUE
    - ENTITY
      - language model

- Probabilistic document length prior for language model
  - TECHNIQUE
    - MOD
      - Probabilistic document length prior
  - APPLICATION
    - ENTITY
      - language model

- Time based language model
  - TECHNIQUE
    - MOD
      - Time based
    - ENTITY
      - language model
• The method needs to be non parametric. The user cannot be asked to fill in number of entities or length or any other description.

• A PCFG is one way to do it.

Problem?

Fixed probability, Not Data Driven!

<table>
<thead>
<tr>
<th>Rule</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phrase – Modifier Entity</td>
<td>0.3</td>
</tr>
<tr>
<td>Phrase – Entity Modifier</td>
<td>0.2</td>
</tr>
<tr>
<td>Phrase – Modifier Entity Modifier</td>
<td>0.3</td>
</tr>
<tr>
<td>Phrase – Modifier</td>
<td>0.2</td>
</tr>
</tbody>
</table>
**Key Insight:** Capture popular entities and mods statistically, good entities are present in the data several times.

\[
p(z_{n+1} = k|z_{1:n}, \alpha) = \begin{cases} 
\frac{n_k}{n + \alpha} & (k \in \{z_1:n\}) \\
\frac{\alpha}{n + \alpha} & (k = K^+ + 1)
\end{cases}
\]

- Use chinese restaurant process to capture most common entities and their parse trees.

This can be realized using Adaptor Grammars!
Adaptor Grammars:

- Uses CRP on chosen non terminals to compute parse tree probability.
- Best performance by adapting entity. Otherwise may compete with mod.
Results and Current status:

Evaluation on ground truth dataset based on DBLP from different computer science areas between 1970 - 2016 (About 150000 titles totally)

Compared against state of the art **Bootstrap method + SegPhrase**

<table>
<thead>
<tr>
<th>Area/Domain</th>
<th>Accuracy of BM+SegPhrase</th>
<th>Accuracy of ACE (our method)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR (Info Retrieval)</td>
<td>0.5705</td>
<td>0.7812</td>
</tr>
<tr>
<td>DM (Data Mining)</td>
<td>0.5593</td>
<td>0.7341</td>
</tr>
<tr>
<td>ML (Machine Learning)</td>
<td>0.3421</td>
<td>0.6842</td>
</tr>
<tr>
<td>DB (Databases)</td>
<td>0.4096</td>
<td>0.4698</td>
</tr>
</tbody>
</table>

We convincingly outperform the baseline on all domains!

Planning to submit to CIKM 2017.
Thank you