DISTANT SUPERVISED RELATION EXTRACTION

Marcel Ackermann
Overview

- Introduction
- Method
- Evaluation
- Conclusion
Introduction: Motivation

- Relation Extraction
  - “the task of recognizing the assertion of a particular relationship between two or more entities in text” (Banko & Etzioni, 2008)
  - “Juan Ramón Jiménez was born in Moguer”
  - relation(partA, partB)

- Distant Supervision
  - Label training corpus (Wikipedia) with a source of relational data (Freebase)

- Why RE?
  - Information Retrieval, Information Extraction, Text Summarization, Question Answering, Paraphrasing, Word Sense Disambiguation…

- Why Distant Supervision?
  - Better than the alternatives!
Introduction: Supervised

- supervised
  - hand-labeled corpora, e.g. SemEval-2 task 8*: 9 relations, 1k labeled sentences
  - any kind of resources: large corpora, dictionaries, lexical-semantic resources
    - any kind of features: lexical, syntactic, semantic
  - state-of-the-art: Rink and Harabagiu (2010), F1 of .82
    - features: context words, hypernyms, POS, dependencies, semantic roles, paraphrases...
- disadvantages
  - labeling: time consuming, expansive
  - few relations, small corpus
  - does not scale
  - domain-dependent
  - -> fails on web corpora and narrow-domain (company) documents

* Multi-Way Classification of Semantic Relations Between Pairs of Nominals
Introduction: Unsupervised & Distant Supervised

- **unsupervised**
  - Zhu et al. (2009): Pattern learning, F1 of .76
  - disadvantages
    - mapping results on existing knowledge bases hard
    - results questionable: Supervised subcomponents (NER, tagger, parser)

- **distant supervised**
  - Mintz et al. (2009): Label Wikipedia with Freebase
  - Distant Supervision Assumption: “If two entities participate in a relation, all sentences that mention these two entities express that relation.”
  - Yao et al. (2010) F1 intra-domain .4, extra-domain .25

- **why our work?**
  - totally distant supervised, no supervised subcomponents, therefore
  - language & domain independent, scales to web size corpora, direct mapping!
Method: Architecture Overview

- Split Freebase instances in test & training
- Training
  - NER on Wikipedia sentences containing relation parts
  - Relation Recognizer (RR) on sentences with
    - both parts of one instance (positive)
    - parts of different instances (negative)
- Extraction
  - Conjunction of NER and RR on all sentences
- Test
  - Extracted relations vs. held-out test data
  - Manual
Method: Data & Components

- **Data**
  - Unsupervised POS tagged Wikipedia sentences: „Juan/6 Ramón/6 Jiménez/10 was/222 born/3 in/3 the/350 house/2 number/2 two/262 the/350 street/2 from/3 the/350 Ribera/1 de/157 Moguer/8“
  - Freebase instances: person/place_of_birth(Juan Ramón Jiménez, Moguer)

- **Named Entitiy Recognizer (NER)**
  - Stanford NER: Conditional Random Fields
  - Features: all combinations of words, tags, positions, classes and their ngrams
  - „Juan/A Ramón/A Jiménez/A was/O born/O in/O the/O house/O number/O two/O the/O street/O from/O the/O Ribera/O de/O Moguer/B”

- **Relation Recognizer (RR)**
  - 1) characterizing (discriminating) n-grams
  - 2) characterizing patterns: before (A,B) between (A,B) after
    - 12 features: \{word,pos\}×\{AB,BA\}×\{before,between,after\}
Evaluation

<table>
<thead>
<tr>
<th>unigrams</th>
<th>bigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Childhood</td>
<td>Hall of</td>
</tr>
<tr>
<td>Poetry</td>
<td>the University</td>
</tr>
<tr>
<td>Career</td>
<td>Early life</td>
</tr>
<tr>
<td>inducted</td>
<td>was born</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>trigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>in the village</td>
</tr>
<tr>
<td>the University of</td>
</tr>
<tr>
<td>grew up in</td>
</tr>
<tr>
<td>was born in</td>
</tr>
</tbody>
</table>

Characterizing n-grams

<table>
<thead>
<tr>
<th>pos</th>
<th>words</th>
<th>mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 A 3 B 2:32</td>
<td>A born B of:6</td>
<td>1 A in B :16</td>
</tr>
<tr>
<td>1 A 3 B:29</td>
<td>A born B was:5</td>
<td>1 A born B</td>
</tr>
<tr>
<td></td>
<td>2:16</td>
<td></td>
</tr>
<tr>
<td>A 3 B 2:23</td>
<td>Personal A was B 6:5</td>
<td></td>
</tr>
<tr>
<td>1 A 3 B 2:20</td>
<td>life A was B:5</td>
<td>2 A 3 B</td>
</tr>
<tr>
<td></td>
<td>1 A was B</td>
<td><em>NUM</em>:17</td>
</tr>
<tr>
<td>1 A 3 B 6:20</td>
<td>Early A was B:51</td>
<td></td>
</tr>
<tr>
<td>1 A 3 B of:15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unigram patterns with highest support

<table>
<thead>
<tr>
<th>P</th>
<th>R</th>
<th>F1</th>
<th>acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>NER</td>
<td>.16</td>
<td>.091</td>
<td>.116</td>
</tr>
<tr>
<td>RR with Patterns</td>
<td>.136</td>
<td>.893</td>
<td>.236</td>
</tr>
<tr>
<td>RR with n-grams</td>
<td>.23</td>
<td>.5</td>
<td>.315</td>
</tr>
</tbody>
</table>

Precision, recall, F1 & accuracy for NER & RR
Conclusion & Future Work

- in theory we have a
  - purely distant supervised
    - Named Entity Recognizer, Relation Recognizer, Relation Extractor
  - that scales to Web size corpora
  - is language and domain independent
  - fast and inexpensive

- in reality
  - lot of work todo
  - mainly improving the NER, idea: Mallet with partially labeled data

- lessons learned
  - look at your data. at first. and again and again…
  - try simple methods first, might be better than getting stuck in the fancy ones
  - use frameworks that support your architecture & docu, maybe ClearTK
References

Thank you!

QUESTIONS & ANSWERS