Citation Segmentation from Sparse & Noisy Data: An Unsupervised Joint Inference Approach with Markov Logic Networks

Dustin Heckmann\textsuperscript{1} Anette Frank\textsuperscript{1}
Matthias Arnold\textsuperscript{2} Peter Gietz\textsuperscript{2} Christian Roth\textsuperscript{2}

\textsuperscript{1}Department of Computational Linguistics, Heidelberg University
\textsuperscript{2}Cluster of Excellence “Asia and Europe”, Heidelberg University

November 19th 2013
Performing automatic citation segmentation

- for a *highly multilingual* bibliography for Ottoman Studies
- operating on *sparse* and *noisy* OCR input
- following an *unsupervised* approach using probabilistic Markov Logic Networks
Turkologischer Anzeiger Online

Detailansicht: TA22, 290

Band: 22
Nummer: 290
Typ: Sammelband
Titel: Stosunki polsko-tureckie. Tadeusz majda ed.
Ort: Warszawa
Jahr: 1995
Kommentare:
- [Polsisch-türkische Beziehungen.]
- Sammelbande

Artikel:
- Kilka uwag o handlu polsko-tureckim w X VI wieku.
  Koładziejczyk, Danusz
- Kobierce z polskich manufaktur jako ilustracja wpływów sztuki tureckiej.
  Bielawska-Słotowa, Beata
- Polskie zabiegi polityczne w Turcji osmańskiej w XIX stuleciu.
  Dąbrowska, Kazimierz
- Rekopyry tureckie w zbiorach polskich.
  MAJDA, Tadeusz
- Udział Polaków w cywilizacyjnym rozwój imperium osmańskiego w połowie XIX wieku w kontekście życia działalności Mustafy Celâeddina Paszy
  Łąk, Jerzy S.
- Urząd Nasreddina Hodży - NasreddinHoca'run mansibi - pierwsza komedia turecka
  w zbiorach polskich
  Łąbczy-Koszterowa, M.
- Uwagi o stosunkach polsko-tureckich w XV wieku do panowania Stefana Batorego.
  Hensel, Wojciech
- Zwrot przymierzy za Mengii Gireja: chanatkrymski z Turcją przeciw Polsce.
  Tyszkiewicz, Jan

Schlagwörter:
- Allgemeines
  - Sammelwerke
- Geschichte
  - Gesamt und Länder oder Einzelheiten längere Zeiträume
  - Beziehungen zu anderen Ländern, über längere Zeiträume
1 Introduction
   - Turkology Annual Online
   - Citation Segmentation

2 Markov Logic Networks and Joint Inference
   - Markov Logic Networks
   - Joint Inference

3 Citation Segmentation using Joint Inference and Markov Logic
   - Markov Logic Rules
   - Experiments
   - Discussion
Turkology Annual Online

- Digitization project at the Cluster of Excellence „Asia and Europe in a Global Context“
- Turkology Annual (TA)
  - Bibliography for Turkology and Ottoman Studies
  - Department of Oriental Studies, University of Vienna
  - Highly multilingual, more than 20 different languages
  - 28 volumes, only appeared in printed form
- Scanning $\rightarrow$ Optical Character Recognition (OCR) $\rightarrow$ Citation Segmentation $\rightarrow$ Database population $\rightarrow$ Web interface
Citation Segmentation

- **Citation**: set of bibliographic information (fields)
- **Citation Segmentation**:
  - Extraction of field instances

<table>
<thead>
<tr>
<th>Number</th>
<th>Author</th>
<th>Title</th>
<th>Location</th>
<th>Year</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>745</td>
<td>miller, Geoffrey</td>
<td>Straits. British policy towards the Ottoman Empire and the origins of the Dardanelles campaign.</td>
<td>Hull</td>
<td>1997</td>
<td>XXVI+604 S.</td>
</tr>
</tbody>
</table>

- **Challenges**:
  - Noise from OCR
  - Lack of redundant citations
  - Complex citation structures
  - Multilinguality
  - Inconsistencies
Markov Logic Networks

- Probabilistic extension of first-order logic
- *Weighted* first-order clauses over knowledge base
- Allow for concise statement of constraints
- Constraints can be violated → *handling uncertainty*
- Weights can be learned from training data or assigned manually
- We assigned manual weights to hand-written rules → *unsupervised*
Joint Inference

- Machine learning technique
- Exploiting redundant information


Two citations of the same article.
Joint Inference

- Machine learning technique
- Exploiting redundant information


In a) author and title are separated, b) lacks a clear separation
Joint Inference

- Machine learning technique
- Exploiting redundant information


We use knowledge extracted from a) to infer a **field separation** in b)
Joint Inference in Information Extraction

- Prior work by Poon & Domingos, 2007:
  - Exploiting recurring citation variants
  - Redundancy of full citation entries
  - Modeled fields: title, author, venue
  - CiteSeer data set

- Our approach:
  - TA does not contain fully redundant citations
    → Instead, we exploit recurring *fields* (authors, editors, locations)
  - Modeled fields: title, author, editor, location, reference, comment, year, pages
Markov Logic Rules I

- Global definitions of citation types and their field structure:
  - Different citation types (articles, monographs, anthologies)
  - Expected fields depend on citation type, e.g. articles do not contain editor:
    \[\text{Type}(c,\text{Article}) \Rightarrow \neg \text{InField}(c,\text{Editor},i).\]

- Local characteristics of fields and delimiters:
  - Special key word delimiters ("ed.", "In:")
  - Characteristics of tokens, e.g. year must consist of digits:
    \[\text{InField}(c,\text{Year},i), \text{Token}(t,i,c) \Rightarrow \text{IsNumeric}(t).\]
Joint inference rules:

- Exploiting redundancy at the field level
- Making use of recurrent entities (authors, editors)
- Example:
    - If two tokens are separated by comma and they are assigned the author field in citation a and they appear next to each other in citation b
      → They are also labeled as author in citation b

- 70 rules
Experiments

- 3 variants of the MLN system, unsupervised, Tuffy:
  - **MLN-Iso**: segmentation on the basis of local citations only
  - **JI-Cit-WCat**: extends MLN-Iso by joint inference exploiting citation-level redundancy
    - Redundant citations extracted from online bibliographic database WorldCat
  - **JI-Field-TA**: extends MLN-Iso by joint inference rules at the field level

- 2 baseline systems:
  - **TA-Regex**: Regular expression based system
  - **ParsCit**: Supervised CRF-based system, small training size

- Evaluation against gold standard:
  - 425 manually annotated citations, 2 annotators
  - Inter-annotator agreement: $\kappa = 0.97$
Field Match

Exact field match:

<table>
<thead>
<tr>
<th>Fields</th>
<th>TA-Regex</th>
<th>ParsCit</th>
<th>MLN-Iso</th>
<th>JI-Cit-WCat</th>
<th>JI-Field-TA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F₁</td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>TITLE</td>
<td>85.5</td>
<td>81.6</td>
<td>83.5</td>
<td>60.0</td>
<td>59.7</td>
</tr>
<tr>
<td>AUTHOR</td>
<td>97.3</td>
<td>87.1</td>
<td>91.9</td>
<td>89.1</td>
<td>91.7</td>
</tr>
<tr>
<td>REF</td>
<td>99.6</td>
<td>89.7</td>
<td>94.4</td>
<td>68.7</td>
<td>67.9</td>
</tr>
<tr>
<td>COMM.</td>
<td>74.7</td>
<td>84.7</td>
<td>79.4</td>
<td>61.6</td>
<td>42.1</td>
</tr>
<tr>
<td>PAGES</td>
<td>96.6</td>
<td>69.3</td>
<td>80.7</td>
<td>67.1</td>
<td>68.7</td>
</tr>
<tr>
<td>LOCATION</td>
<td>92.0</td>
<td>78.9</td>
<td>84.9</td>
<td>82.4</td>
<td>87.0</td>
</tr>
<tr>
<td>YEAR</td>
<td>97.3</td>
<td>89.4</td>
<td>93.2</td>
<td>91.1</td>
<td>95.0</td>
</tr>
<tr>
<td>EDITOR</td>
<td>66.7</td>
<td>5.6</td>
<td>10.3</td>
<td>67.6</td>
<td>69.4</td>
</tr>
<tr>
<td>all (macro-avg.)</td>
<td>88.7</td>
<td>73.3</td>
<td>77.3</td>
<td>73.2</td>
<td>71.6</td>
</tr>
<tr>
<td>all (micro-avg.)</td>
<td>92.8</td>
<td>84.3</td>
<td>88.3</td>
<td>77.9</td>
<td>75.5</td>
</tr>
</tbody>
</table>

Precision, Recall and F₁-Score by fields, macro-average, micro-average
Confusion Graphs
Discussion

- All MLN formalizations clearly outperform supervised CRF-based and rule-based methods on the TA data set
- Clear gains in recall with largely comparable precision
- Joint Inference over fields (JI-Field-TA) yields best overall results
- ParsCit scores lowest overall
- MLN Approach: unsupervised
Conclusion

Joint Inference with Markov Logic Networks for citation segmentation on sparse & noisy data

- Local and global constraints for addressing noise and sparse data
- Generalization and mutual resolution of field structure
- Knowledge-based rule encoding with probabilistic inference
- Efficient and unsupervised approach for small, non-redundant and noisy data sets
- Easily adaptable to novel data sets and domains
- Supplemented by a web-based search interface for Turkology and Ottoman Studies
References

Councill, I.G., Giles, C.L. and Kan, M.-Y.
ParsCit: An open-source CRF reference string parsing package

Domingos, P. and Lowd, D.
Markov Logic. An Interface Layer for Artificial Intelligence

Hazai, G. and Kellner-Heinkele, B. eds.
Turkology Annual
Universität Wien. Institut für Orientalistik, 1975ff

Poon, H. and Domingos, P.
Joint Inference in Information Extraction
In Proceedings of the national conference on Artificial Intelligence, 2007