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

Dustin Heckmann[1] Anette Frank[1]
Matthias Arnold[2] Peter Gietz[2] Christian Roth[2]

[1]Department of Computational Linguistics, Heidelberg University

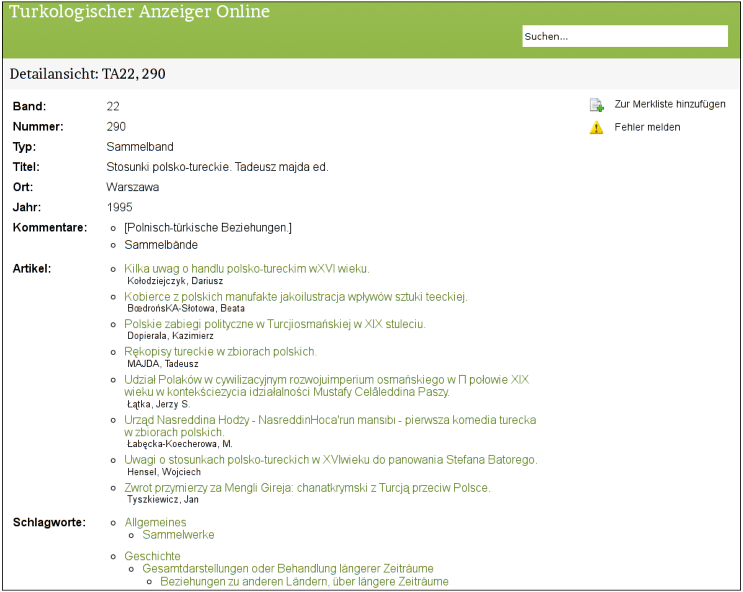
[2]Cluster of Excellence "Asia and Europe", Heidelberg University

November 19th 2013

Turkology Annual - A Showcase for Digital Humanities Research

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



Detailansicht: TA22, 290

Band: 22

Nummer: 290

Typ: Sammelband

Titel: Stosunki polsko-tureckie. Tadeusz majda ed.

Ort: Warszawa

Jahr: 1995

Kommentare:

- [Polnisch-türkische Beziehungen.]
- Sammelbände

Artikel:

- Kilka uwag o handlu polsko-tureckim wXVI wieku.
  Kołodziejczyk, Dariusz
- Kobierce z polskich manufakte jakoilustracja wpływów sztuki teeckiej.
  BœdrońsKA-Słotowa, Beata
- Polskie zabiegi polityczne w Turcjiosmańskiej w XIX stuleciu.
  Dopierala, Kazimierz
- Rękopisy tureckie w zbiorach polskich.
  MAJDA, Tadeusz
- Udział Polaków w cywilizacyjnym rozwojuimperium osmańskiego w Π połowie XIX wieku w kontekściezycia idziałalności Mustafy Celâleddina Paszy.
  Łątka, Jerzy S.
- Urząd Nasreddina Hodży - NasreddinHoca'run mansıbı - pierwsza komedia turecka w zbiorach polskich.
  Łabęcka-Koecherowa, M.
- Uwagi o stosunkach polsko-tureckich w XVIwieku do panowania Stefana Batorego.
  Hensel, Wojciech
- Zwrot przymierzy za Mengli Gireja: chanatkrymski z Turcją przeciw Polsce.
  Tyszkiewicz, Jan

Schlagworte:

- Allgemeines
  - Sammelwerke
- Geschichte
  - Gesamtdarstellungen oder Behandlung längerer 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

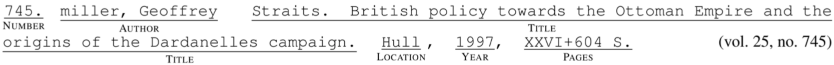
3. Citation Segmentation using Joint Inference and Markov Logic
    - Markov Logic Rules
    - Experiments
    - Discussion

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 → Optical Character Recognition (OCR) → **Citation Segmentation** → Database population → Web interface

Citation Segmentation

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



745. miller, Geoffrey Straits. British policy towards the Ottoman Empire and the
origins of the Dardanelles campaign. Hull , 1997, XXVI+604 S. (vol. 25, no. 745)

NUMBER
AUTHOR
TITLE
TITLE
LOCATION
YEAR
PAGES

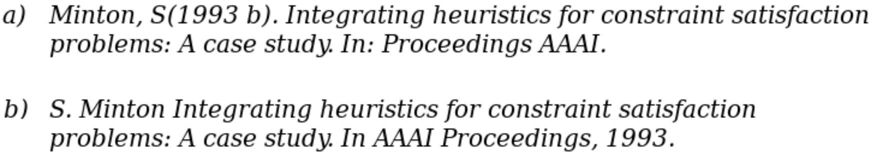
- Challenges:
  - Noise from OCR
  - Lack of redundant citations
  - Complex citation structures
  - Multilinguality
  - Inconsistencies

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 $\rightarrow$ *handling uncertainty*
- Weights can be learned from training data or assigned manually
- We assigned manual weights to hand-written rules $\rightarrow$ *unsupervised*

Joint Inference

- Machine learning technique
- Exploiting redundant information



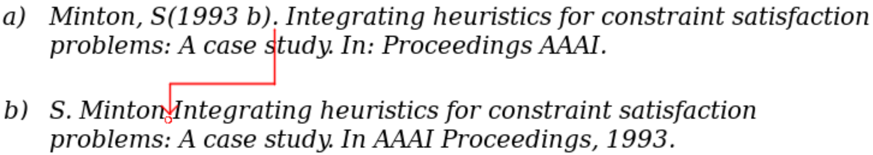
a) *Minton, S(1993 b). Integrating heuristics for constraint satisfaction problems: A case study. In: Proceedings AAAI.*

b) *S. Minton Integrating heuristics for constraint satisfaction problems: A case study. In AAAI Proceedings, 1993.*

Two citations of the same article.

Joint Inference

- Machine learning technique
- Exploiting redundant information



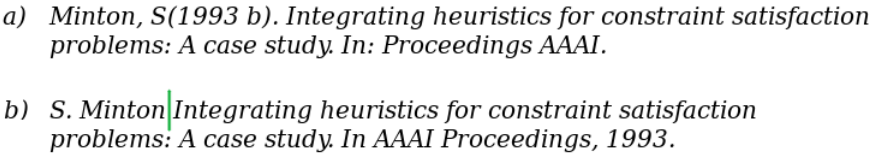
a) Minton, S(1993 b). Integrating heuristics for constraint satisfaction problems: A case study. In: Proceedings AAAI.

b) S. Minton Integrating heuristics for constraint satisfaction problems: A case study. In AAAI Proceedings, 1993.

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

Joint Inference

- Machine learning technique
- Exploiting redundant information



a) *Minton, S(1993 b). Integrating heuristics for constraint satisfaction problems: A case study. In: Proceedings AAAI.*

b) *S. Minton Integrating heuristics for constraint satisfaction problems: A case study. In AAAI Proceedings, 1993.*

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:

    ```
    Type(c,Article) => !InField(c,Editor,i).
    ```
- Local characteristics of fields and delimiters:
  - Special key word delimiters ("ed.", "In:")
  - Characteristics of tokens, e.g. year must consist of digits:

    ```
    InField(c,Year,i), Token(t,i,c) => IsNumeric(t).
    ```

Markov Logic Rules II

- Joint inference rules:
  - Exploiting redundancy at the field level
  - Making use of recurrent entities (authors, editors)
  - Example:
  - `474. Germano-turcica. Zur Geschichte des Türkisch-Lernens in den deutschsprachigen Ländern. Klaus Kreiser ed. Bamberg, 1987, 161 S.`
  - `2137. Kreiser, Klaus Edirne im 17. Jahrhundert nach Evliya Çelebi. Ein Beitrag zur Kenntnis der osmanischen Stadt. Freiburg/Breisgau, 1975, XXXIII + 289 S. [...]`
    - 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*

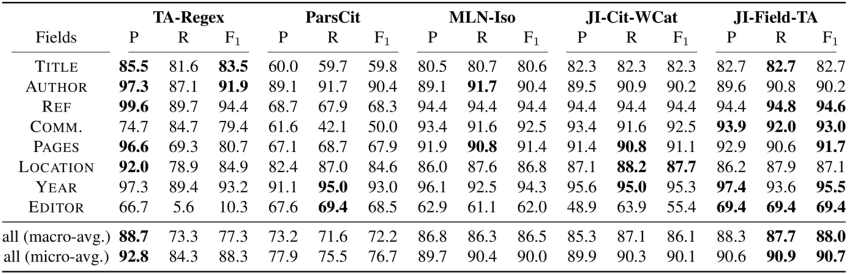
    → They are also labeled as author in citation *b*
- 70 rules

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

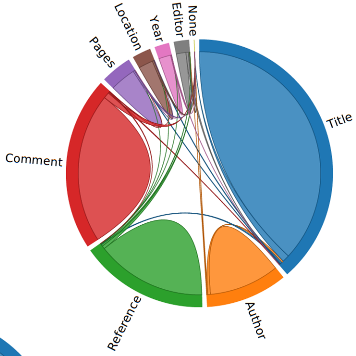
Excact field match:



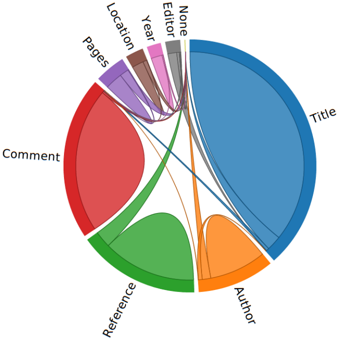
Fields	TA-Regex			ParsCit			MLN-Iso			JI-Cit-WCat			JI-Field-TA		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Title	**85.5**	81.6	**83.5**	60.0	59.7	59.8	80.5	80.7	80.6	82.3	82.3	82.3	82.7	**82.7**	82.7
Author	**97.3**	87.1	**91.9**	89.1	91.7	90.4	89.1	**91.7**	90.4	89.5	90.9	90.2	89.6	90.8	90.2
Ref	**99.6**	89.7	94.4	68.7	67.9	68.3	94.4	94.4	94.4	94.4	94.4	94.4	94.4	**94.8**	**94.6**
Comm.	74.7	84.7	79.4	61.6	42.1	50.0	93.4	91.6	92.5	93.4	91.6	92.5	**93.9**	**92.0**	**93.0**
Pages	**96.6**	69.3	80.7	67.1	68.7	67.9	91.9	**90.8**	91.4	91.4	**90.8**	91.1	92.9	90.6	**91.7**
Location	**92.0**	78.9	84.9	82.4	87.0	84.6	86.0	87.6	86.8	87.1	**88.2**	**87.7**	86.2	87.9	87.1
Year	97.3	89.4	93.2	91.1	**95.0**	93.0	96.1	92.5	94.3	95.6	**95.0**	95.3	**97.4**	93.6	**95.5**
Editor	66.7	5.6	10.3	67.6	**69.4**	68.5	62.9	61.1	62.0	48.9	63.9	55.4	**69.4**	**69.4**	**69.4**
all (macro-avg.)	**88.7**	73.3	77.3	73.2	71.6	72.2	86.8	86.3	86.5	85.3	87.1	86.1	88.3	**87.7**	**88.0**
all (micro-avg.)	**92.8**	84.3	88.3	77.9	75.5	76.7	89.7	90.4	90.0	89.9	90.3	90.1	90.6	**90.9**	**90.7**

Precision, Recall and F_1-Score by fields, macro-average, micro-average

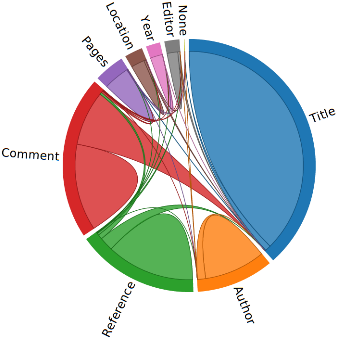
Confusion Graphs



MLN-Iso



TA-Regex



ParsCit

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
  In Proceedings of LREC 2008, Marrakech, pp. 661-667.

- Domingos, P. and Lowd, D.
  Markov Logic. An Interface Layer for Artificial Intelligence
  In R. R. Brachmann & T. Dietterich, eds. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Playpool, 2009

- 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