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

Dustin Heckmann, Anette Frank
{heckmann,frank}@cl.uni-heidelberg.de

Matthias Arnold, Peter Gietz, Christian Roth
{arnold,croth}@asia-europe.uni-heidelberg.de,gietz@daasi.de

Department of Computational Linguistics, Heidelberg University
Cluster of Excellence "Asia and Europe", Heidelberg University

Citation Segmentation in a Digital Humanities Context. Bibliographies are an important resource for scientific research. Their storage in (online) bibliographic databases offers efficient search functionalities for wide-spread and timely use in international research communities. For this purpose it is crucial to automatically detect the inherent structure of bibliographic references, by isolating and extracting citation subfields (e.g., author, title, venue).

Previous approaches in citation segmentation strongly rely on language-specific lexical data and multiple occurrences of the same citation entry in online publication repositories. However, when dealing with multilingual data, the use of language-specific knowledge becomes difficult. Moreover, self-contained data sources like printed bibliographies are naturally short of recurring citation entries, and thus cannot rely on data redundancy.

In this work, we present an approach to citation segmentation that operates on sparse and noisy OCR input originating from a single, multilingual bibliography, the *Turkology Annual* (*Turkologischer Anzeiger*).[1] The *Turkology Annual* is a bibliography for Turkology and Ottoman studies, comprising 28 volumes which are only available in print. Citation entries containing multiple languages and scripts, the shortage of citation redundancy, frequent OCR errors and inconsistencies in citation structure impede the use of state-of-the-art statistical approaches for citation segmentation.

Citation Segmentation on Sparse & Noisy Data using Markov Logic Networks. Following [2], our approach builds on Markov Logic Networks (MLN), a framework of statistical relational learning that combines first-order logic with probabilistic modelling [3]. Formalization in first-order logic offers high expressivity and flexibility, and thus makes it possible to tailor citation segmentation to the specific conventions of a given bibliography – in our case the *Turkology Annual*. MLNs can be trained on labeled data in a supervised learning scenario, but can also be applied in an unsupervised way, using structure learning or manually set rule weights. Given the lack of training data and the shortage of data redundancy in bibliographic sources, MLNs offer an attractive framework for citation segmentation using unsupervised methods.

In our paper we develop an approach to citation segmentation using Markov Logic Networks in a joint inference setting. We apply this method to a large multilingual bibliography obtained from noisy OCR output. The particular challenges we address are: noise from OCR, multilinguality, complex citation entry structures, inconsistencies and lack of redundancy. Our joint inference

[1] See [1] and the digitized online resource created on the basis of the present work: http://kjc-fs2.kjc.uni-heidelberg.de:8000/en/.

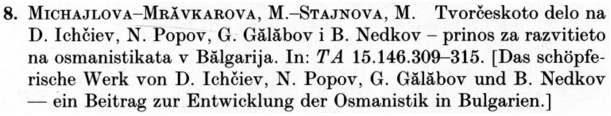
approach extends the scope of prior work in [2] by exploiting redundancy at the field level. By this move, we are able to cope with the lack of citation redundancy.

A basic MLN formalization for individual entries outperforms strong baselines obtained from traditional regular expression based parsing as well as a supervised statistical approach using Conditional Random Fields (CRF). Inclusion of joint inference at the *entity* and *field levels* yields further performance gains in recall and precision, with joint inference over fields yielding the best overall results. Our approach is fully unsupervised, using manually assigned rule weights.

Our evaluation experiments show that in face of the specific challenges found with segmenting references from a digitized bibliography, our MLN formalizations outperform both rule-based and state-of-the-art statistical methods. On our evaluation data set we obtain 88% F_1-score for exact field match, a 24.8% increase over a CRF-based system baseline.

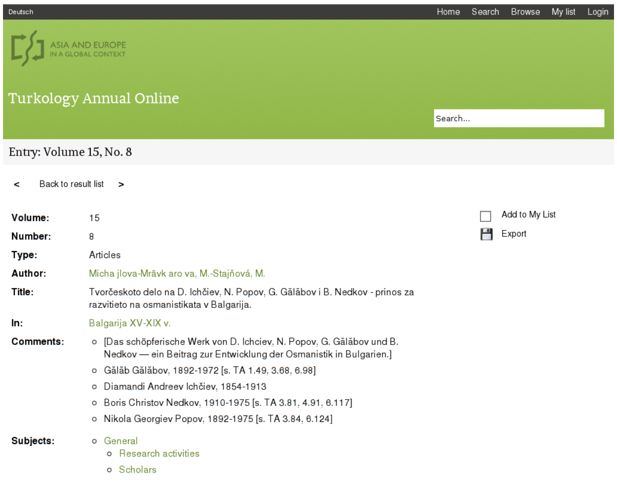
Outcomes. Contrasting with prior work we address a data set from a Digital Humanities context that features sparse and noisy data. Our method extends [2]'s approach by applying *joint inference* at the *field level*. By this move, we are able to cope with the lack of citation redundancy and noise in the data. Our approach is fully unsupervised, hence avoids creating annotated training data. The rule sets we designed can be adapted to other bibliographies, or further types of digitized sources, such as historical dictionaries or encyclopedias.

In practical terms, the results of the project have been made accessible to the public through the *Turkology Annual Online* web interface[2]. It provides functionality for searching and browsing within the TA. Bibliographic sub-fields (e.g. title, author) are stated explicitly and can be used as criteria for searching and sorting. Cross-references are easily accessible using hyperlinks. Citations can be exported in various formats, allowing the use of reference management software like BibTeX. Figure 1 shows a sample citation as obtained from the OCR process. Figure 2 shows the corresponding entry resulting from the segmentation process, as visible on the web site.



8. MICHAJLOVA–MRĂVKAROVA, M.–STAJNOVA, M. Tvorčeskoto delo na D. Ichčiev, N. Popov, G. Gălăbov i B. Nedkov – prinos za razvitieto na osmanistikata v Bălgarija. In: *TA* 15.146.309–315. [Das schöpferische Werk von D. Ichčiev, N. Popov, G. Gălăbov und B. Nedkov — ein Beitrag zur Entwicklung der Osmanistik in Bulgarien.]

Fig. 1. Turkology Annual: Sample scan of references



Entry: Volume 15, No. 8

Volume: 15
Number: 8
Type: Articles
Author: Micha jlova-Mrävk aro va, M.-Stajňová, M.
Title: Tvorčeskoto delo na D. Ichčiev, N. Popov, G. Gălăbov i B. Nedkov - prinos za razvitieto na osmanistikata v Balgarija.
In: Balgarija XV-XIX v.
Comments:
- [Das schöpferische Werk von D. Ichciev, N. Popov, G. Gălăbov und B. Nedkov — ein Beitrag zur Entwicklung der Osmanistik in Bulgarien.]
- Gălăb Gălăbov, 1892-1972 [s. TA 1.49, 3.68, 6.98]
- Diamandi Andreev Ichčiev, 1854-1913
- Boris Christov Nedkov, 1910-1975 [s. TA 3.81, 4.91, 6.117]
- Nikola Georgiev Popov, 1892-1975 [s. TA 3.84, 6.124]

Subjects:
- General
  - Research activities
  - Scholars

Fig. 2. Turkology Annual Online: display of a single entry

[2] http://kjc-fs2.kjc.uni-heidelberg.de:8000/

References

[1] G. Hazai and B. Kellner-Heinkele, Eds., *Turkologischer Anzeiger*. Universität Wien. Institut für Orientalistik and Universität Wien. Orientalisches Institut, 1975–, vol. 1–28. [Online]. Available: http://orientalistik.univie.ac.at/forschung/publikationen/turkologischer-anzeiger

[2] H. Poon and P. Domingos, "Joint Inference in Information Extraction," in *Proceedings of the Twenty-Second National Conference on Artificial Intelligence*. Vancouver, Canada: AAAI Press, 2007.

[3] M. Richardson and P. Domingos, "Markov Logic Networks," *Machine Learning*, vol. 62, no. 1, pp. 107–136, 2006.