Citation segmentation from sparse & noisy data: a joint inference approach with Markov logic networks

Dustin Heckmann and Anette Frank
Department of Computational Linguistics, Heidelberg University, Germany

Matthias Arnold, Peter Gietz and Christian Roth
Cluster of Excellence ''Asia and Europe'', Heidelberg University, Germany

Abstract

This article presents an approach to citation segmentation that addresses special challenges as typically found in Digital Humanities applications. We perform citation segmentation from Optical Character Recognition (OCR) input obtained from volumes of a printed bibliography, the *Turkology Annual*. This showcase application features serious difficulties for state-of-the-art techniques in citation segmentation: *multilingual citation entries*, *lack of data redundancy*, *inconsistencies*, and *noise from OCR input*. Our approach is based on Markov logic networks (MLN) (Richardson and Domingos, Markov logic networks. *Machine Learning*, **62**(1): 107–36, 2006), a framework of statistical relational learning that combines first-order logic with probabilistic modeling. Formalization in first-order logic offers high expressivity and flexibility, and makes it possible to tailor segmentation to specific conventions of a given bibliography. We show that in face of the specific difficulties found with segmenting references from a digitized bibliography, our MLN formalizations outperform state-of-the-art statistical methods. We obtain 88% F_1-score for exact field match, a 24.8% increase over a conditional random fields-based system baseline. In contrast to prior work, we address a data set featuring sparse and noisy data. Our method extends Poon and Domingos (Joint Inference in information extraction. In *Proceedings of the Twenty-Second National Conference on Artificial Intelligence*. Vancouver, Canada: AAAI Press, 2007)'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 can be characterized as knowledge based and hence does not rely on 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.

Correspondence:
Anette Frank,
Heidelberg University,
Department of
Computational Linguistics,
Im Neuenheimer Feld 325,
69120 Heidelberg, Germany
E-mail: frank@cl.uni-heidelberg.de

Digital Scholarship in the Humanities © The Author 2014. Published by Oxford University Press on behalf of EADH.
All rights reserved. For Permissions, please email: journals.permissions@oup.com
doi:10.1093/llc/fqu061

1 Introduction

Bibliographies are an important resource for scientific research. Their storage in (online) bibliographic databases offers efficient search functionalities for widespread and timely use in international research communities. Today, bibliographic references are available in standardized structured formats [e. g. Metadata Object Description Schema[1]] that can be searched or mapped to databases and electronic catalogs. But this is not the situation we typically find in the Humanities, where bibliographies are often only available in print. For making such resources searchable in electronic catalogs, it is crucial to automatically detect the inherent structure of bibliographic references, by isolating and extracting citation subfields (e.g. author, title, venue). That is, the task of *citation segmentation,* which is central to the present work, is to segment full bibliographic references in a given bibliographic style and to extract from it structured *field information* that can be stored in a bibliographic database, as illustrated below:

Type:	Article
Author:	Cüceloğlu, Doğan
Title:	Türkçe türetme sistemi üzerinde psikolinguistik bir çalışma.
In:	HSBBD 7.1-2.1975.35-58.
Comment:	[Eine psycholinguistische Studie über das türkische Wort bildungssystem.]

Previous approaches to 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, making use of language-specific knowledge becomes difficult. Moreover, self-contained data sources such as 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 Optical Character Recognition (OCR) input originating from a single, multilingual bibliography, the *Turkology Annual* (*Turkologischer Anzeiger*).[2] The *Turkology Annual* is a bibliography for Turkology and Ottoman studies, comprising 28 volumes which were 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.

Following Poon and Domingos (2007), our approach builds on Markov logic networks (MLNs), a framework of statistical relational learning that combines first-order logic with probabilistic modeling (Richardson and Domingos, 2006). 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 assigned 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.

Our contributions in this article are the following. We present an approach to citation segmentation using MLNs 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 as follows: noise from OCR, multilinguality, complex citation entry structures, inconsistencies, and lack of redundancy. Our joint inference approach extends the scope of prior work in Poon and Domingos (2007) by exploiting redundancy at the field level. By this move, we are able to cope with the lack of citation redundancy.

Our approach does not rely on labeled training data, but relies on a carefully designed rule set that models the citation entry structures. 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.

The structure of this article is as follows. In Section 2, we discuss the challenges related to citation segmentation in a Digital Humanities setting. Section 3 reviews related work on citation segmentation, including recent work using statistical relational learning and joint inference techniques. Section 4 introduces our method for inducing high-quality citation segmentations from noisy and sparse data in self-contained bibliographic resources, using a joint inference approach in the framework of MLNs. Section 5 presents our evaluation experiments. We offer detailed error analysis and discussion of results. In Section 6 we present our conclusions.

2 **Citation Segmentation in a Digital Humanities Context**

The Turkology Annual as a showcase. The *Turkology Annual, henceforth TA* (Hazai and Kellner-Heinkele, 1975) is the leading systematic bibliography for Turkology and Ottoman studies. Since 1975, 28 volumes have appeared containing about 62,000 references[3] to articles, books, and reviews. They were collected by contributors from more than 30 countries and are written in more than 20 languages including Russian, Turkish, Arabic, and Japanese. Some of the languages using non-Latin scripts appear in transcription (e.g. Russian, Arabic), others appear in their original script (e.g. Greek).

While this bibliography represents a unique tool indispensable for study and research, its use is quite complicated. Besides its being available in printed form only, volumes 1–9 appeared as annual addendum to the *Wiener Zeitschrift für die Kunde des Morgenlandes* which means that in most libraries, the *TA* volumes are stored in two different places. Whether one wants to perform a topic-driven search using the *TA*'s system of categories or a search using its index, whether one wants to check reviews for a given title or to overview the articles of a conference volume, one will end up spending hours following complex internal references and browsing back and forth in a huge pile of books. Thus, there was an urgent need for turning this bibliography into an online database with nonlinear access, which could also offer additional benefits, such as an interactive interface with crosslinks, cross-volume searching or subject-browsing, saving or exporting lists of references, or library search, in order to look up references in local bibliographic catalogs.[4]

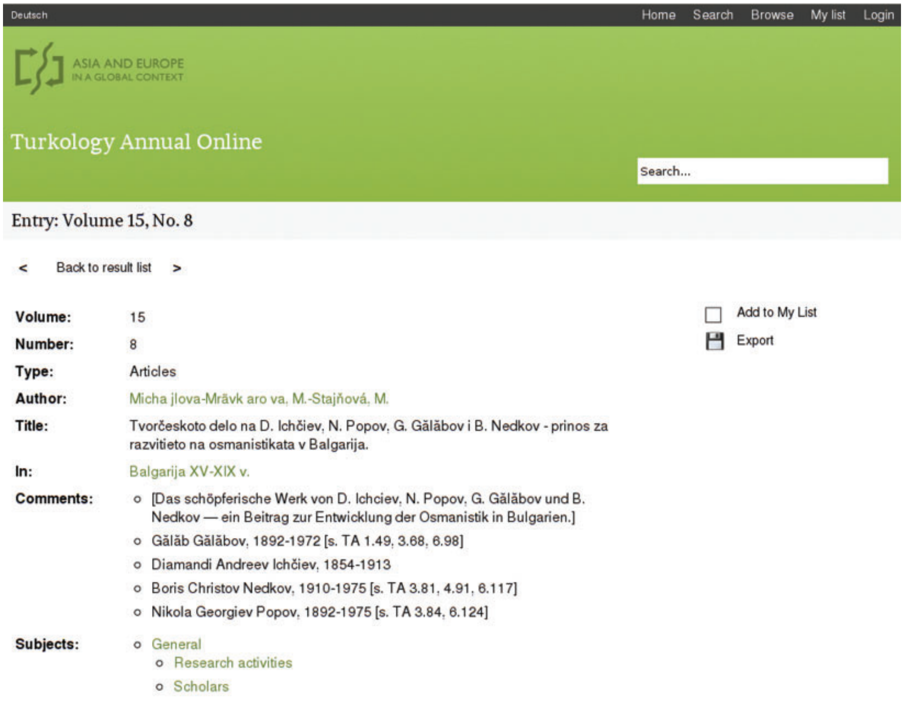
The plan of converting the *TA* to a searchable database was supported by the *TA* editorial board, the heir of the late Prof. Dr. Andreas Tietze who founded the series in 1975, by the Department of Oriental Studies of the University of Vienna as its publisher, and the Hungarian Academy of Sciences, maintaining close contact and cooperation. The project has also revived considerations to continue the *TA* in entirely electronic form.[5]

Digitizing the TA. The process consisted of the following main tasks:

(i) *Scanning: Producing high-quality scans of all available TA volumes in the Cluster's Media Lab,*[6]
(ii) *Image preprocessing and OCR*;
(iii) *Design of a database schema for storage and querying of entries;*
(iv) *Parsing, i.e., syntactic analysis of the OCRed data, including citation segmentation of individuated references;*
(v) *Storage: population of the database with the results of syntactic analysis;*
(vi) *Search: construction of a Web-based search interface for searching and browsing within the database.*

The results of the project have been made accessible to the public through the *Turkology Annual Online* Web interface[7] (cf. task vi). It provides functionality for searching and browsing within the TA. Bibliographic subfields (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 into various formats, allowing the use of reference management software such as BibTeX. Figure 1 shows an example entry as visible on the Web site.

The core of this article concentrates on task (iv), and in particular, on the subtask of *citation segmentation* applied to individuated references on the scanned and OCRed pages of the full volumes.



Turkology Annual Online

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. 1 *TA* Online: display of a single entry

We further discuss the preprocessing stages that are relevant for preparing the input for citation segmentation. A sample scan of the TA entry corresponding to the citation in Figure 1 is displayed in Figure 2.

Challenges. *Scanning and OCR.* Comparison of the OCR results to existing preprint files for the more recent *TA* volumes showed that with proper fine-tuning the *ABBYY FineReader 9.0* software produced reasonably good results. *Syntactic analysis*, however, soon proved to be difficult due to a number of aspects:

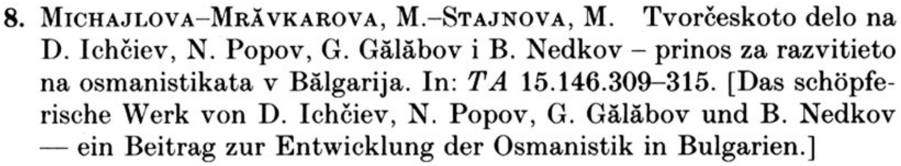
Complexity and inconsistency of entry structures. The *TA* features a variety of entry types with different data structures. These are only implicitly marked, and some of these different data structures and their markings even change from volume to volume.

Noise. The printed *TA* volumes contain human-made syntactic errors such as missing punctuation marks as well as typing errors. OCR errors imposed further noise by disrupted and otherwise ill-formed entry structures.

Word-level processing and multilinguality. As typically found with bibliographies, the *TA* entries are made up of short chunks, such as author, title, editors, locations, etc. The *TA* is an international bibliography without a single reference language, only titles in less well-known languages are translated into German, English, or French. As a result, single entries containing field chunks coded in several different languages are frequently encountered, and since the fields are short, we could not rely on Abbyy's language recognition. For the same reason, the use of existing natural language part-of-speech taggers and parsers was thus not a promising option.

In the face of these problems, using a conventional approach using regular expression-based parsing was not successful: it was not possible to achieve a sufficient balance of precision and robustness.

In fact, state-of–the-art methods in citation segmentation are achieving decent performance on standard data sets, such as the widely used



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. 2 *TA*: Sample scan of references

CiteSeer[8] or Cora[9] data sets.[10] Such data sets are typically created from electronically available document sources, cleaned-up, and most importantly, they feature a considerable amount of redundancy on the citation level. For instance, in the CiteSeer data, about one third of the citations are non-singletons, with up to 21 citations per paper; in the Cora data set, almost every citation occurs multiple times, with up to 54 citations per paper (cf. Poon and Domingos, 2007).

As is well-known from other problems in computational language processing, such as part-of-speech tagging or parsing, many application contexts in Digital Humanities offer less ideal circumstances, and if so, the performance of methods developed on 'clean' and well-behaved data sets drastically decreases. Compared to the natural sciences, multilinguality is a greater challenge in the Humanities, since not only primary sources from any language of the world are studied, but also because secondary literature is more often not being written in English, but in a number of other languages. Also multilingual XML mark-up has been discussed (Bia *et al.*, 2006). We find a similar situation for citation segmentation when applied in a Digital Humanities context as described in this work (cf. footnote 10): The task is applied to OCR input obtained from a manually maintained multilingual bibliography, and hence is noisy and difficult to process. The citation structures are complex and partly inconsistent. Most notably, bibliographies are short of recurrent entries, therefore redundancy-oriented methods suffer from data sparsity. To date, there is little work on citation segmentation in Digital Humanities contexts (cf. Section 3). Nevertheless, our work builds on and extends prior art. We will be using MLNs for the segmentation task. Unlike Poon and Domingos (2007), we will dispense with supervision, using manually assigned or uniform rule weights. We further apply joint inference on the field level, in order to exploit redundancy at the sub-citation level.

The approach used here for bibliographic citations can well be generalized to other types of semi-structured texts, such as dictionaries,[11] encyclopedias, fieldbook entries,[12] archival descriptions, etc. Especially within the Digital Humanities, where numerous such texts still have to be digitized via OCR, we see a large range of use cases for our approach.

3 Related Work

Citation (or *Reference*) *Extraction and Segmentation*[13] are a special application of *Information Extraction* (IE) (Cowie and Lehnert, 1996) tied to the field of Information Science. The task consists of identifying citations within unstructured input data, and segmenting them into predefined fields that can be mapped to database records used in (online) bibliographical databases and citation networks.

Prior work in this area is divided in two main strands, which focus on different aspects of the task:

(i) *Citation Extraction* or *Mining* concentrates on the extraction subtask (see Cortez *et al.*, 2007; Afzal *et al.*, 2010; Park *et al.*, 2012). Here, references embedded in running text need to be identified and separated from the surrounding textual material. (ii) *Citation Chunking* or *Parsing* (see Councill *et al.*, 2008; Groza *et al.*, 2012) refers to the task of citation segmentation proper. Here, entries identified in the extraction phase are automatically segmented into subfields, such as author, title, volume and further bibliographic information. It is at this stage that duplicate citation references must be identified as referring to the same bibliographic entity. This latter process is referred to as *Entity Resolution*.

Citation indexing refers to the extended task of extracting (i) and resolving (ii) citations, and furthermore (iii) building complete citation indices and linking articles that are referenced by citations. One of the earliest approaches to *citation indexing* was the seminal work of Giles *et al.* (1998), which led to the citation database *CiteSeer*, now *CiteSeerX*. Here, the focus was on identifying citations and their citation contexts in scientific documents and resolving all citations that refer to the same article (*entity resolution*). Citation segmentation was performed using a heuristic *invariants first* parsing strategy in order to cope with structural variants of citations.

State of the art in citation segmentation. Since that time, the techniques for citation extraction and segmentation have evolved considerably. The major approaches are rule-based, template-based, and machine learning-based methods. Prototypical for **rule-based methods** is the *CiteSeer* system of Giles *et al.* (1998) that makes use of heuristic, regular expression-based citation parsing strategies. References are parsed using a greedy 'invariants first' method: using observations from overall data characteristics, secure and stable citation elements are analyzed first, taking into account field characteristics that can serve as secure indicators. Characteristics may be field-specific (keywords, data type, or lexical constraints) as well as positional. Here, (possibly domain-specific) knowledge about citation structure and layout is directly encoded in the parsing algorithm. **Template-based approaches** start from a knowledge base of templates designed to match complete extracted citations. A prototypical template-based system is *ParaCite* (Jewell, 2000). The system makes use of a set of predefined reference templates with individuated fields that may be characterized using regular expressions. References to be resolved are matched against these templates, where the different fields are weighted in computing an overall matching score. The highest scoring template is used to split the target reference into fields. These methods are closely related, but address the problem from slightly different perspectives: rule-based methods address the problem from an algorithmic perspective, whereas template-based methods focus on a more declarative, knowledge-based modeling approach. Both methods require extensive modeling, are lacking robustness against noise, and thus suffer from coverage problems when confronted with noise. Both methods require highly articulated rule sets (of hundreds of rules) that are expensive to construct and difficult to maintain.

As an alternative, most current methods are employing **supervised machine learning methods** to tackle the segmentation task. Most prominent are sequence labeling approaches, such as Hidden Markov Models (HMM) and CRF. Some works investigate Support Vector Machine (SVM) and maximum entropy models as learning frameworks. Similar to other sequence labeling tasks, such as part-of-tagging or chunking, these models perform citation segmentation by assigning specific field labels to individual tokens of a citation. That is, the field labels serve as classes the model needs to assign to tokens, and the learner determines optimal weights for characterizing features of the field classes using labeled training data. Features include character-level surface properties of fields as well as positional information and keywords (e.g. frequent location or publisher names stored in a gazetteer).

Among the earliest probabilistic accounts was Seymore *et al.* (1999), using a HMM model with citation fields as hidden states. Peng and McCallum (2004) proposed CRFs as graphical models that are particularly suited for sequence labeling problems. CRF models have been successfully applied to the citation segmentation task with continuous improvements by optimizing feature sets (see i.a. Peng and McCallum, 2004; Councill *et al.*, 2008; Groza *et al.*, 2012). *ParsCit* (Councill *et al.*, 2008) is available as an open-source package and will be used as a baseline in our experiments. The systems typically make use of large gazetteers for names, locations, journals, publishers, etc. CRF-based systems achieve high accuracy rates for field segmentation. However, as supervised systems they are dependent on domain-specific training sets.

Unsupervised learning for citation segmentation has been attempted in different ways. Cortez *et al.* (2007) makes use of a knowledge base that is automatically constructed using existing sample citations, so that the system can recognize component fields. Grenager *et al.* (2005) investigate unsupervised HMM models, making use of prior knowledge for guiding the search space. They obtain moderate performance levels that can approach those of supervised approaches only if their training sets are restricted to a size of about one third of what is typically used for state of the art supervised systems. Thus, the performance of unsupervised learning approaches for citation segmentation is still unsatisfactory.

Poon and Domingos (2007) developed a novel approach for citation segmentation using **joint inference** techniques. Here, citation segmentation and entity resolution are performed jointly. The motivation for combining these tasks is that well-structured citation entries can support the entity resolution task. Also, in a joint approach, insecure field separation decisions can be supported by independent evidence from entity resolution, thereby avoiding early and potentially erroneous insecure decisions in the field segmentation. Poon and Domingos (2007) define joint segmentation and entity resolution using MLNs in a supervised setting, with rule weights estimated from training data. The system outperforms previous systems on the CiteSeer and Cora data sets. Using the MLN framework, segmentation and resolution rules are transparently defined in first-order logic formulas and can be flexibly redesigned. Given probabilistic rule weighting, the system is robust toward noise and idiosyncrasies in the data.

Citation segmentation on noisy data and in Digital Humanities contexts. To our knowledge, there is little work on citation segmentation from noisy OCRed data sources. Besagni and Belaid (2004) process OCRed citations for 140 journals from pharmacology, with rather homogeneous citation structure. Their method is based on part-of-speech (PoS) information, and thus not suited for multilingual bibliographies.[14] Takasu (2003) combines a HMM model for citation segmentation with OCR error patterns. The approach relies on a specified syntactic structure of citation entries. The system is supervised using large amounts of training data. Finally, Kim *et al.* (2012a,b) present a CRF model for bibliographic reference field annotation applied to heterogeneous citation forms in a large online publication platform in a Digital Humanities project. They independently train and apply specialized CRF models for three citation types and obtain good performance using large annotated training sets.

Casinius and Sporleder (2007) approach a related task of segmenting zoological fieldbook entries with an unsupervised HMM approach. To compensate for the lack of training data, they train a language model for field segments from existing structured database entries of the relevant domain and estimate sequential HMM probabilities on a small seed set. They could show that unsupervised HMM modeling in combination with supervision from a structured knowledge base yields best performance.

The approach by Canisius and Sporleder relies on a preexisting structured database for the target application to train the HMM model, and it is strongly lexicalized. A number of manual rules are applied to compensate for divergences between the database fields and the unsegmented input, to allow the model to recognize field-internal separation markers. In this respect, the difficult nature of our data set—inconsistent or missing separation markers, multilingual entries and heterogeneous domains—presents serious challenges for such an approach, and it is unclear whether high-quality segmentations could be achieved.

Our approach will be targeted to apply to very general segmentation task settings that are lacking preexisting structured databases or larger amounts of labeled training data, and instead focuses on knowledge-based modeling of the segmentation problem.

4 Citation Segmentation on Sparse & Noisy Data using MLNs

Our approach to citation segmentation is situated in a classical Digital Humanities context. The major differences of our setting compared to mainstream approaches to citation segmentation are the following:

(i) the bibliography and individual citations are multilingual and involve different scripts;
(ii) the input to citation segmentation is obtained from OCR, with noisy recognition of characters, including field demarcation markup;
(iii) the citation structure of the *TA* is complex and contains numerous inconsistencies;
(iv) as a self-contained bibliography, the TA features little data redundancy at the citation level.

This setting is typical for citation and related segmentation tasks in a Digital Humanities context (see Section 2). These specific difficulties require methods that are *robust toward noise*, are able to *deal with inconsistencies*, and *lack of data redundancy*. Being confronted with application-specific entry structures and a considerable cost for producing annotated data,[15] we target methods that produce high-quality results using little or no annotated data, that are intelligible for human encoding and that offer high flexibility for adapting existing rule sets to comparable data sets.

Given these requirements, we propose an approach to citation segmentation for sparse and noisy data using the framework of MLNs and joint inference techniques.

4.1 MLNs and joint inference

MLNs (Richardson and Domingos, 2006) represent a powerful formalism for solving problems in Artificial Intelligence. They allow for concise statement of (interacting) constraints using first-order logic formulae over knowledge bases in a given domain; at the same time, they encode probabilistic graphical models, or Markov Networks, that define probability distributions over possible worlds. In this way, MLNs offer a powerful tool for the encoding of domain knowledge, while allowing for weak and violable constraints in inference.[16]

More specifically, a MLN represents a first-order knowledge base with a probabilistic weight attached to each formula of the knowledge base. Joint with a set of constants that represent the objects in a domain, it defines a probability distribution over possible worlds, where each world corresponds to an assignment of truth values to all possible ground atoms. In this ground Markov Network the probability of a possible world x is defined as a log-linear model as follows,

$$P(X + x) = \frac{1}{Z}\exp\Big(\sum w_i n_i(x)\Big) \qquad (1)$$

where Z is a normalization constant, w_i is the weight of the ith formula, and $n_i(x)$ specifies the number of true groundings of the formula F_i in world x (cf. Domingos and Lowd, 2009).

The weights of the formulae of a MLN can be estimated from labeled data by optimizing the conditional probability of the query predicate y that corresponds to the annotated label given the possible worlds x where the evidence predicates hold. Normalization is over all worlds x that are consistent with the evidence.

$$P(Y = y|X = x) = \frac{1}{Z_x}\exp\Big(\sum w_i n_i(x, y)\Big) \qquad (2)$$

Besides supervised learning scenarios, MLNs can be applied in an unsupervised manner by manually assigning rule weights. Finally, MLNs can be applied to structure learning, performing unsupervised induction of rules.

Joint inference. Current research in machine learning investigates inference techniques that are able to capture and exploit dependencies between individual component models. The key idea of joint inference is to explicitly model dependencies between separate component analysis levels, to overcome restricted performance of the individual component models.

In natural language processing (NLP) and IE tasks, we are typically dealing with uncertainty and contradictions which are likely to arise in automated processing. A promising strategy to account for these problems is to perform joint inference over individual phenomena in order to obtain a globally optimal solution, overcoming the limitations of traditional pipeline models.

Starting with early work by Roth and Yih (2004), joint learning has become an established framework in NLP research. Examples include models that combine semantic entity type information from a named entity classifier with decisions of a coreference system (Denis and Baldridge, 2009), or a syntactic parsing model (Finkel and Manning, 2010). Joint inference is also increasingly used in classical IE tasks. Such approaches exploit data redundancy to infer target information in cases of insecure knowledge or noise in the data (Poon and Domingos, 2007). They also define complex data structures as joint predictions from component analyses, as in Poon and Vanderwende (2010). Li *et al.* (2011) apply similar techniques in a cross-document setting.

Joint inference frameworks make use of powerful inference techniques using MLNs, or Integer Linear Programming (ILP). Due to the logics-based formalizations of MLNs, joint models can be flexibly defined and modified. ILP formalizations require slightly more abstract definitions using binary trigger variables.

Joint inference in citation segmentation. Poon and Domingos (2007) are the first to apply joint inference in citation segmentation, by jointly resolving segmentation and entity resolution. They exploit recurring citation variants of one and the same entity to resolve insecurities in field segmentation. The approach crucially relies on redundancy of full citation entries. Redundancy is assumed if two citations share similar values within two fields. Field values for titles, e.g. are considered similar if they share substrings that start with the same token trigram and end with the same token. We will apply joint inference techniques in a sparse data setting, and derive segmentation information across *distinct* citation entities. This will be achieved by exploiting dependencies at the field level.

4.2 MLN formalization(s)

This section introduces the MLN formalization for segmenting the OCRed citation input of the *TA*.

We start by defining the basic data structures, i.e. different types of citation entries, and a set of rules that define citation field structures locally. We then define additional rules that capture dependencies across citations. They will allow us to perform joint inference in the presence of sparse data and noise.

4.2.1 *Citation entry types and modeled fields*

The *TA* features the following citation types: articles (75%), monographs (18%), collections (5%), and conference proceedings (2%).[17] Figure 3 illustrates the structure of entries for the three major entry types.[18]

Overall, we distinguish 10 field types. Table 1 gives an overview of the types together with a description and their occurrence frequency in our evaluation data set.

4.2.2 *Segmentation rules*

Rules encode different aspects of the data: (i) global and (ii) local structural properties, as well as (iii) dependencies across citations. Rules can be stated to hold without exception (using a closing period),[19] or will be assigned learned or manually determined weights.

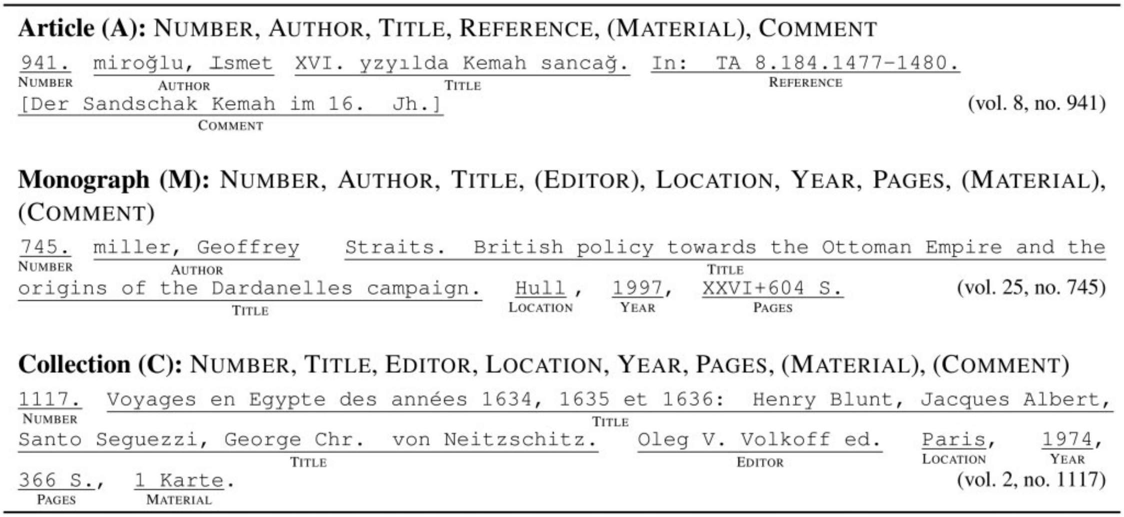
(i) Global definitions of citation types and their field structure.

A first set of definitions and rules defines the different entry types and which fields are admitted for the respective types. We define the 2-place predicate `Type(c,t)` with constants `t=Tarticle`, `Tmonograph` or `Tcollection`. `Type(c,t)` is true if a citation `c` is an instance of citation type `t`.

Field options can be stated by enumeration or by negative constraints on fields that are illicit for a given type. In order to compactly deal with optional fields, we prefer negative field specifications, as illustrated below. Possible field types are defined as constants `Fauthor`, `Ftitle`, `Fyear`, etc. Following Poon and Domingos (2007), we formalize the segmentation problem as the task of assigning field labels to tokens. This is expressed via the 3-place predicate `InField(c,f,i)` which is true if the token at position `i` in citation `c` belongs to an instance of the field `f`, where `f ∈{Fnumber, Ftitle, Fauthor,...}`.

Some fields do not appear in citations of a certain type. For instance, the editor field does not appear in article citations. This restriction is enforced by the following rule, using the negation operator “!”.[20]

```
Type(c,Tarticle) => !InField(c,Feditor,i).
```



Article (A): NUMBER, AUTHOR, TITLE, REFERENCE, (MATERIAL), COMMENT

941. miroğlu, Ismet XVI. yzyılda Kemah sancağ. In: TA 8.184.1477-1480.
[Der Sandschak Kemah im 16. Jh.] (vol. 8, no. 941)

Number
Author
Title
Reference
Comment

Monograph (M): NUMBER, AUTHOR, TITLE, (EDITOR), LOCATION, YEAR, PAGES, (MATERIAL), (COMMENT)

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)

Collection (C): NUMBER, TITLE, EDITOR, LOCATION, YEAR, PAGES, (MATERIAL), (COMMENT)

1117. Voyages en Egypte des années 1634, 1635 et 1636: Henry Blunt, Jacques Albert,
Santo Seguezzi, George Chr. von Neitzschitz. Oleg V. Volkoff ed. Paris, 1974,
366 S., 1 Karte. (vol. 2, no. 1117)

Fig. 3 Sample entries for: article (A), monograph (M), and collection (C)

Table 1 Modeled field types of the *Turkology Annual*; A = article, M = monograph, C = collection, ordered by frequency of occurrence (measured on our evaluation data set)

Field type	Citation type	Frequency of occurrence	Description
Number	A, M, C	425	Running citation number within each volume
Title	A, M, C	424	Title
Author	A, M	374	Name(s) of author(s)
Reference	A	252	Specifies journal, issue, year and pages for articles
Comment	A, M, C	202	May state any kind of information, often translation of title
Pages	M, C	163	Page numbers (absolute, range)
Location	M, C	161	City of publication
Year	M, C	161	Year of publication
Editor	M, C	36	Name(s) of editor(s)
Material	A, M, C	30	Material contained in publication: photographs, maps, . . .

In addition, we define a number of general constraints: (i) that citation types are disjunctive; (ii) that every entry realizes exactly one of the predefined types; and (iii) that fields must be contiguous. These properties are defined as hard constraints.

(i) `Type(c,Tarticle) v Type(c,Tmonograph) v Type(c,Tcollection)`

(ii) `Type(c,t1), t1 != t2 =>!Type(c,t2)`

(iii) `InField(c,f,h), InField(c,f,j), h<i<j => InField(c,f,i)`

Field sequencing. The order of fields within an entry is usually fixed. Positional information for a field is controlled by referring to the position index of tokens filling this field, by the `InField` predicate with index variable `i`.

Again, we avoid fully specified order constraints, and instead resort to *partial order constraints* if possible. For instance, if the author field precedes the title field, we define a constraint that prevents `Ftitle` to precede `Fauthor`. If the order is restricted to a specific entry type, this is specified as an additional constraint using the `Type` predicate (see below).

```
InField(c,Fauthor,j), j > i
  => !InField(c,Ftitle,i)
```

(ii) Local characteristics of fields and delimiters.

Besides global constraints, we encode local characteristics of the different citation entry types. To this end, we employ a number of predefined predicates (cf. Table 2).

Positional information. For example, monographs and articles start with the author field, while collections are headed by the title field (all following the citation index number, at position 0, see below).

```
Type(c,Tmonograph) =>
  InField(c,Fauthor,1)
Type(c,Tarticle) => InField(c,Fauthor,1)
Type(c,Tcollection) =>
  InField(c,Ftitle,1)
```

Field delimiters: punctuation and indentation. Punctuation and indentation are in general no secure field delimiters. Punctuation may be found within fields as abbreviations (e.g. of names), or as delimiters within titles, to separate subtitles. The position of delimiters is encoded by the predicate `FollowedBy(c,i,d)`, which is true if some delimiter `d` is found between the positions `i` and `i + 1` in citation c.[21]

For some field types, we can safely assume punctuation as field delimiters, as in the case of the field year, which is surrounded by commas.[22]

For indentation, we define a rule that captures the TA convention that separates author and title fields by indentation. Indentation is difficult to distinguish from ordinary space delimiters in OCRed input, thus the corresponding rules are subject to exceptions.

```
FollowedBy(c,h,COMMA), Token(t,i,c),
  IsYear(t), FollowedBy(c,i,COMMA),
  i = h + 1 => InField(c,Fyear,i)
FollowedBy(c,i,INDENTATION) =>
  InField(c,Fauthor,i)
FollowedBy(c,i,INDENTATION), j = i + 1 =>
  InField(c,Ftitle,j)
```

Table 2 Features encoding characteristics of tokens

Name	Description
IsSmallCaps	Small caps
IsItalic	Italics
IsYear	4-digit number
IsAlphaChar	Single alphabetical character
IsDigit	Digit
IsCapital	Capital letter
IsMaterialToken	Tables, figures, . . .
FollowedBy	Followed by . . .
INDENTATION	Indentation
PERIOD	Period
COMMA	Comma
LBRACKET	Opening square bracket
RBRACKET	Closing square bracket

Characteristics of tokens. Finally, we define field-specific lexical characteristics of tokens.[23] The rule below models the token properties of author names, which are typically capitalized, consist of alphanumeric characters, or may be in small caps font. The TA uses the dash to link multiple authors which is thus admitted as a token of the field `Fauthor`.

```
InField(c,Fauthor,i), Token(t,i,c) =>
  IsCapital(t) v IsAlphaChar(t) v
  IsSmallCaps(c,i) v t = "–".
```

Special key word delimiters. Some fields are marked by special delimiters or key words. For instance, the reference field in articles is introduced by the token `In:`. This key word can be used to resolve the citation type to article.

```
Token("In", i, c), FollowedBy(c,i,COLON)
  => Type(c,Tarticle)
```

Following this model, we defined a local 'grammar' consisting of 65 rules that segment the three major TA citation entry types into subfields (see Table 1 for an overview of the extracted fields).

(iii) Joint inference rules.

In addition to global and local segmentation rules that define the properties of individual entries, we employ two types of joint inference rules. These are intended to ensure globally optimal segmentation decisions in the presence of insecure local decisions due to noisy OCR input as well as inconsistencies in citation entries.

We employ two types of rules: rules that exploit information from recurrent full citation entries, as well as dependencies cross distinct citation entries at the field level.

Citation-level redundancy. In the TA, cases of recurring citations appear as ‘repetitions’: Publications referring to publications (e.g. reviews) mentioned in an earlier TA volume are marked with the string `s. TA x.y`, where `x` is the referred citation’s volume number and `y` its index number (e.g. `s. TA 14.302`). However, the repetition only contains the original citation’s abbreviated author name and the beginning of the title, therefore it is not fully redundant. Still, we can exploit this knowledge in order to identify the boundary delimiting *Author* and *Title* fields, which is notoriously difficult to locate. In our evaluation data set, for 118 out of 425 entries, we find a TA-internal reference, i.e. in 27.8% of all instances.

To this end, we define the predicate `MatchingTokens(o,o1,o2,r,r1,r2)` that is true in case the token sequence from positions `o1` to `o2` in citation `o` matches the token sequence from positions `r1` to `r2` in citation `r`. The underlying string overlap is detected in preprocessing and can be used as evidence for inference.[24] The following rule is based on the assumption that matching token sequences in original citation and repetition most likely belong to the title field. A match is considered secure if it comprises at least three tokens.

```
IsRepeated(orig,rep),
  MatchingTokens(orig,o2,o4,rep,r2,r4),
  o3 >= o2, o3 < o4 => InField(c,Ftitle,o3).
```

Importing entries from WorldCat Catalogue. As an alternative to address the sparseness of citation recurrences, we will measure the relative impact of recurrent full entries on segmentation quality by importing citation entries from the WorldCat[25] Catalogue.

WorldCat is a large online citation database and is accessible through an API. We retrieved records similar to TA citations by sending automatically generated queries to the API. The queries contained a citation’s first six tokens with a minimal character length of 3. For each retrieved citation, we compute a rudimental relevance value. It is defined as the number of common tokens divided by the number of tokens in the source citation. Only the two best-scoring citations with relevance >30% are retrieved. For 89 out of 425 citations in our evaluation set we find WCat matches (20.9%). In contrast to TA-internal references, these entries contain full bibliographical information. For 74 out of 425 citations we find both TA-internal references and WCat matches (17%).

Field-level redundancy. Unlike Poon and Domingos (2007), in our application, we cannot rely on a large number of recurrent citation entries. Instead, we design joint inference rules that operate at the field level, to capture repeated occurrences of field values, such as author and editor fields, that may be identified in secure positions within distinct citation entries. That is, we rely on the (reasonably plausible) assumption that the author of some cited work has written or published further articles or books. Similar rules can be designed for names of journals or publishers. Since publishers do not appear in the TA and names of journals are easily recognized by the preceding string ‘In:’, we focus on names of authors and editors.

Rules of this type work as follows: Suppose we are processing citation entry CURRENT and are unsure about the name separation. In such cases, we may look for alternative citations that involve the current (hypothetical) author or editor name. If we can detect corresponding tokens in securely identifiable positions (e.g. at the beginning of a citation), we may conclude that the chosen values can indeed be defined as values of the `Fauthor` or `Feditor` field.[26]

```
RangeInField(c,Fauthor,i,j):-
Token(lastName,1,bib),
FollowedBy(bib,1,COMMA),
Token(firstName,2,bib),
FollowedBy(bib,2,INDENTATION),
Token(lastName,i,CURRENT),
Token(firstName,j,CURRENT), Next(j,i),
Next(k,j), bib != CURRENT.
```

We adopt two such rules for the *author* and *editor* fields, with two variants each, to account for positional alternatives in the different citation types.[27]

5 Experiments and Results

This section describes the data, experiments, and results of citation segmentation using the MLN formalization presented above, when applied to sparse and noisy data from OCRed input of the *TA* multilingual bibliography (cf. Section 2).

5.1 Data and experimental setup

Volume of data. The project worked on the first 26 volumes of the TA, which comprises a data volume of 6,179 scanned pages and 61,939 single database entries.

Scanning, OCR, and tokenization. All volumes were scanned to 600 ppi grayscale uncompressed TIFF files using two Plustek OpticBook A300 book scanners. To reduce shine-through disturbances, black paper was inserted behind each page before scanning. The images then went through a semi-automatic batch file postprocessing workflow where systematic file naming, cropping, and a slight contrast enhancing were applied.

The resulting optimized TIFF files were loaded in the OCR software *Abbyy Finereader 9.0* for character recognition. A number of tests were necessary to achieve optimal OCR results. Since so many different languages occur overall and within individual entries and since the fields are short, we could not rely on Abbyy's recognition capabilities for special characters. We therefore developed a custom pattern file that contained most of the special characters[28] and assigned it to a language that would most probably not be encountered (Xhosa as a sort of garbage category). The OCR engine was set to recognize as many of the relevant languages as possible and to Xhosa where the language recognition failed (uncertain characters). To be able to use the language identification from Abbyy Finereader for parsing, we decided to use MS Word XML (wordml) as output format.

The resulting OCR output was preprocessed by applying simple tokenization rules without any linguistic knowledge, using blanks and punctuation as token delimiters. On the basis of the obtained tokenization, we measure 23.6 ($s = 9.0$) tokens/citation in average (on the entire data set).

Annotation of an evaluation data set. In order to establish a data set for evaluation, we manually annotated 425 citations, which were extracted from five different volumes, distributed evenly over consecutive production phases of the bibliography. We tried to approximate the real distribution of entry types that is representative for the TA as a whole, which we estimated from the successfully parsed entries using the rule-based citation parser: 75% articles, 18% monographs, and 5% collections. The distribution shown in Table 3 was obtained on the basis of the sampled annotation data set. For 20% of the data (85 citations), we performed independent annotation by two annotators, one of them an expert in the domain. This pilot annotation yielded a high inter-annotator agreement of 0.97 (Cohen's Kappa: Cohen, 1960) measured on field annotations. Given this high agreement, the remaining data were annotated by a single (the non-expert) annotator.

For experimentation and rule development, we made use of a small *development set* consisting of 150 citations that were manually annotated. In contrast to the evaluation data set, we made sure we have sufficient data for the less frequent and more challenging citation types, namely monographs and collections.

Inference framework. As inference framework we are using Tuffy (Niu *et al.*, 2011). Tuffy offers an efficient architecture for MLN inference that is combined with a relational database system. In our experiments, we apply marginal inference using the MC-SAT algorithm and choose the most likely field for each token.

Setting rule weights. The rule weights are manually assigned. The ranges for soft constraints are chosen from a range between 1 and 25 and are manually optimized on the development data set. Some rules are realized as hard constraints, such as rules stating that every citation starts with an index number. Internally, these rules are assigned very high weights, dynamically computed by the inference framework Tuffy. We also evaluate system performance with uniform rule weight settings.

Baselines. We compare our results against two baselines. One baseline is constituted by the regex-based parser (**TA-Regex**) that was developed on the

Table 3 Evaluation, development, and training sets: citation types

Item	Development set	Evaluation set	Training set 1	Training set 2 (learning curve)
Articles	54 (63%)	252 (59%)	94 (47%)	218 (54%)
Monographs	76 (50%)	125 (30%)	83 (41%)	143 (36%)
Collections	20 (14%)	48 (11%)	23 (12%)	39 (10%)
All	150 (100%)	425 (100%)	200 (100%)	400 (100%)
Average tokens/citation	23.6 (±9.0)	23.6 (±9.0)	23.5 (±8.9)	23.6 (±8.9)

basis of one particular volume (Volume 19). Given some variation in the structure of entries across different volumes, this baseline is representative for rule-based approaches, which typically achieve high precision, but are sensitive to noise.

As an additional baseline, and for more direct reference to the state of the art in probabilistic approaches to citation segmentation, we train and evaluate the CRF-based *ParsCit* package (Councill *et al.*, 2008) on our own data (**ParsCit**). For training of *ParsCit*, we used 125 citations from our (small) development set (cf. Table 3) plus an additional 75 entries disjoint from the evaluation set.[29] All citations were randomly extracted from different TA volumes.

ParsCit learning curve. In order to conduct a learning curve experiment, we annotated a set of 200 additional training instances. These were sampled aiming at a similar citation type distribution as can be found in the TA and were annotated by the same annotator that had labeled the previous data sets. The statistics of the data set is included in Table 3. ParsCit was run on increasing sizes of training set sizes, ranging from 40 to 400 instances, with step size 40.

Evaluation measures. We apply standard evaluation measures as described in Peng and McCallum (2004): *token F_1-score* for measuring correct field assignment at the token level, as well as *complete instance (= citation) accuracy* for evaluating the overall segmentation performance at the citation level. In addition, we report a more rigorous evaluation for field-specific performance in terms of precision (P), recall (R), and F_1-score (F_1),[30] measuring performance in terms of *exact field match*, i.e. completely predicted fields against the gold standard, as opposed to assignment at the token level (token F_1-score), which only evaluates the correct field labeling of individual tokens.

Error types. For the analysis of error types, we group errors into four types and compute their distribution:

(I) actual field contains assigned field
(II) assigned field contains actual field
(III) other kind of overlap
(IV) no overlap

Class (IV) is considered the most serious error type, followed by class (III). We prefer errors of classes (I) and (II) over those of class (III) and (IV).

5.2 Experiments and results

Experiment I: MLN model variants. We conduct experiments with different MLN segmentation rule sets.

(i) **MLN-Local** employs a MLN formalization that defines segmentation on the basis of local citations only.
(ii) **JI-Cit** extends MLN-Local by adopting joint inference rules that exploit citation-level redundancy within the TA or by import from WorldCat (cf. Section 4). We refer to these settings as **JI-Cit-TA** and **JI-Cit-WCat**, respectively.
(iii) **JI-Field-TA** extends MLN-Local by joint inference rules at the field level, with dependencies defined for *Author* and *Editor* field values.[31]

Using these rule sets, we conduct *segmentation experiments* with manually assigned rule weights optimized on the development set (see Section 5.1). For evaluation, we employ the manually annotated evaluation data set of 425 citations, described in Section 5.1, cf. Table 3.

Experiment II: Uniform rule weights. Next to manually assigned rule weights, we conduct the same experiment with uniform rule weights.

As an underlying MLN variant for these experiments, we chose **JI-Field-TA**, which performed best in Experiment I. We set rule weights to 1 for uniform soft constraints.

Evaluation results for Experiment I. We compare the results of **MLN-Local**, **JI-Cit**,[32] **and JI-Field** against the two baselines: our regex-based parser (**TA-Regex**) and the CRF-based system *ParsCit* (**ParsCit**), trained on the held-out data set of 200 entries, a size comparable to earlier evaluation of supervised systems on standardized data sets.

Table 4 reports the evaluation figures for *exact field match*.[33] Next to the exact field match, we also evaluate precision, recall, and F1-score for correct *token* assignment as well as results for *full citation match*. This is summarized in Table 5 for evaluation across all field types.

Comparison to baselines. Looking at the exact field match in Table 4, both baselines yield moderate F_1-scores of 77.3% (**TA-Regex**) and 72.2% (**ParsCit**). The manually created regex-based citation parser outperforms the CRF system with a great advance in precision (+15.5 percentage points, pp.) and even small gains in recall (+1.7 pp.). Compared to its performance on standard citation datasets,[34] the weakness of the statistical system clearly shows the difficulty of the task, due to a combination of lack of redundancy in the data, noise, and the complexity of the citation structures, with insecure field separators.

Our *local* MLN formalization **MLN-Local** clearly outperforms both baselines, by a margin of +9.2 (**TA-Regex**) and +14.3 (**ParsCit**) pp. F_1-score. Compared to the stronger **TA-Regex** baseline, we observe important gains in recall (+13 pp.) with only a slight drop in precision (−1.9 pp.).

For the **joint inference** formalizations **JI-Cit-TA** and **JI-Cit-WCat** that target redundancy at the *citation level*, the results are very close, with P: 87.4/R: 86.8/F_1: 87.1 for JI-Cit-TA and P: 85.3/R: 87.1/F_1: 86.1 for JI-Cit-WCat, i.e. with a slight advance of **JI-Cit-TA** in precision, and **JI-Cit-WCat** in recall. We note especially high precision gains when using TA-internal references in **JI-Cit-TA** for the *Year* and *Editor* fields.[35] Both variants outperform the baselines, but yield only small performance differences compared to the local **MLN-Local** setting: +0.6/+0.5/+0.6 P/R/F_1 for **JI-Cit-TA** and −1.5/+0.8/−0.4 P/R/F_1 for **JI-Cit-WCat** (cf. fn 33).

The performance difference in F_1-score for field match between the **TA-Regex** and **ParsCit** baselines and the **MLN-variants** are all statistically significant at p-level $p = 0.00001$ against **TA-Regex** and p-level $p = 0.02$ against **ParsCit**, using a two-tailed approximate randomization permutation test (Yeh, 2000; Padó, 2006). The differences between the different MLN-variants **MLN-Local**, **JI-Cit-TA** and **JI-Field-TA** cannot be determined as statistically significant (at p-level 0.05) on the basis of the given evaluation set. However, we note a superiority of the joint formulations **JI-Cit-TA** and **JI-Field-TA** over **MLN-Local**, as these both outperform the **ParsCit** baseline at $p = 0.01$ in contrast to **MLN-Local** with a significance level of $p = 0.02$.

Note in particular that **JI-Field-TA** approaches the strong precision baseline of **TA-Regex** by a small margin of −0.4 pp. We display the results of **JI-Cit-TA** for overall comparison of settings in Table 4, as it performs better than **JI-Cit-WCat** in precision and overall (+1pp.) (cf. fn. 33). Finally, **joint inference** at the *field level* in **JI-Field-TA** yields the best overall results, with highest F_1-score of 88.0% and highest overall recall of 87.7% among all system variants. We also obtain further gains in precision over **MLN-Local** and **JI-Cit-TA**, of +1.5 and +0.9pp, with overall strongest precision results for *Editor* and *Comment* fields. Joint inference over fields was defined for the *Author* and *Editor* fields, and indeed we find precision gains for both fields (though only marginally for *Author*). Thus, joint inference on the field level is effective for difficult fields, such as *Editor*, and in addition can affect recognition of other fields, here *Comment*.

Regarding **individual citation fields**, we observe a number of strong divergences of the MLN-formalizations compared to the baselines. Most prominent are *Editor*, *Title*, *Pages*, and *Comment* fields, with a rise in the F_1-score of +59.1 for *Editor* against

Table 4 Experiment I: Segmentation results on evaluation set for baselines, local and joint inference

Fields	TA-Regex			ParsCit			MLN-Local			JI-Cit-TA			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	81.2	81.4	81.3	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	88.9	90.1	89.5	89.6	90.8	90.2
Reference	**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**
Comment	74.7	84.7	79.4	61.6	42.1	50.0	93.4	91.6	92.5	93.0	91.6	92.3	**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	93.1	**90.8**	**91.9**	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	84.1	85.7	84.9	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	**98.1**	93.8	**95.9**	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	66.7	66.7	66.7	**69.4**	**69.4**	**69.4**
All (macro-average)	**88.7**	73.3	77.3	73.2	71.6	72.2	86.8	86.3	86.5	87.4	86.8	87.1	88.3	**87.7**	**88.0**
All (micro-average)	**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**

We report P/R/F_1-score for *exact field match* for individual fields and at citation level (all).

Table 5 Experiment I

Measures	TA-Regex				ParsCit				MLN-Local				JI-Cit-TA				JI-Field-TA			
	P	R	F_1	Acc.	P	R	F_1	Acc.	P	R	F_1	Acc.	P	R	F_1	Acc.	P	R	F_1	Acc.
Token	90.8	74.1	79.1		87.6	86.8	86.6		92.2	**90.4**	91.2		92.1	89.3	90.5		**92.9**	89.9	**91.3**	
Field (macro-average)	**88.7**	73.3	77.3		69.9	71.5	70.5		86.8	86.3	86.5		87.4	86.8	87.1		88.3	**87.7**	**88.0**	
Field (micro-average)	**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**	
Citation (exact)				63.3				44.7				73.4				73.4				**75.8**

Segmentation results on evaluation set: different evaluation measures: token (= word) assignment, exact field match, and exact citation match

TA-Regex, and +22.9 for *Title*, +23.8 for *Pages* and +43.0 for *Comment* against **ParsCit**.

Token, field and full citation match. Table 5 completes the evaluation by measuring precision, recall, and F_1-score for *token assignment* as well as *full citation accuracy*.[36] Here we obtain overall similar results, with constant performance gains from baselines over local MLN to joint inference settings, again with joint inference over fields yielding best overall results in F_1-score. Highest results for exact field match are obtained for **JI-Field-TA** in F_1-score and recall, only slightly beaten by **TA-Regex** in precision (+0.4 pp.). **JI-Field-TA** also achieves the best F_1-score and precision for *token assignment* and best accuracy for *full citation match*.[37]

Learning curve experiment. Given the close performance results of the supervised *ParsCit* baseline and the MLN formalizations, and the relatively small training set provided to the learning-based system, we performed a learning curve experiment by providing more training instances to the CRF-based *ParsCit* model. A plot of *ParsCit* performance on increasing data set sizes in Figure 4 clearly shows that the system's performance quickly levels out: overall performance reaches its peak with 160 training instances; *Author* and *Title* fields reach a maximum at training set size of 280 and 360; *Location* and *Year* reach their highest values at training size 400, but with a small overall increase for *Year*, and a flattening curve for *Location*. Overall, the performance of *ParsCit* clearly drops from its peak at 77.3 F_1 (ParsCit160) to 74.0 F_1 (ParsCit400), which may be a result of over-fitting.[38]

Error analysis for Experiment I. In addition to the learning curve experiment, we substantiate the results obtained in Table 4 by two types of error

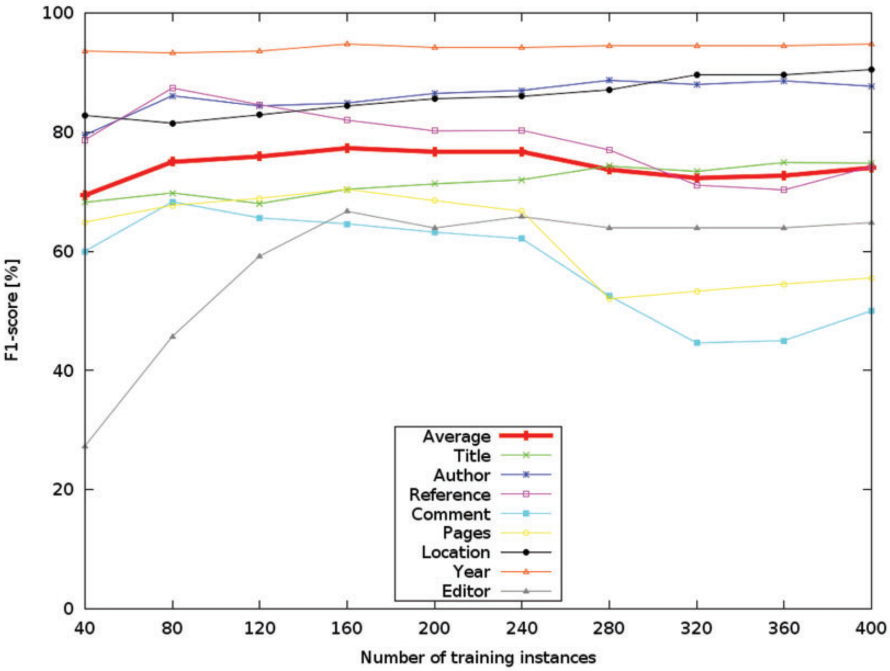


Fig. 4 ParsCit learning curve: F_1 for field match (individual fields and macro-average)

analysis: analysis of misclassifications in confusion matrices and distribution of error type classes.

Confusion matrices for **TA-Regex**, **ParsCit**, and **JI-Field-TA** outputs (–8, respectively) show clear differences. We find a relatively low variation within columns in the **TA-Regex** confusion matrix (Table 6). This is in line with the high performance in token assignment seen in Table 5. However, this also results in a large number of tokens that are not assigned any field (recorded in right-most column) and thus a notably lower recall. This coincides with the intuition that regex-based methods are well suited for precisely describing textual patterns but lack the flexibility needed to cope with noise and uncertainty.

The **ParsCit** classifications, shown in Table 7, yield a different picture. Almost every token is assigned to a field, which results in lower precision. Particularly striking is the large number of misclassifications seen in the Title column. This could be influenced by the very fact that Title, which is present in all entry types, occurs more frequently than any other field and is thus the most probable field assignment (cf. Table 1).

JI-Field-TA (cf. Table 8) shows a more balanced approach to precision and recall (cf. last column). Also, misclassifications are mainly restricted to adjacent fields (e.g. Title and Author/Editor)—in contrast to **ParsCit**, which very frequently confuses Title with Comment, two nonadjacent fields that exhibit similar internal structure.[39] This can be taken as evidence that the MLN account, by stating *global* constraints on the field structure of complete citation entries[40] can considerably improve the results, by exploiting information that complements the characteristics of the internal structure of fields. Note that, similarly, we find few misclassifications of non-neighboring fields with **TA-Regex**. This is due to the sequentially oriented parsing of entries using regular expressions and greedy heuristics, but suffers from recall. In contrast, **JI-Field-TA**, using freely interacting constraints of different nature (absolute and relative positioning information, internal field properties, as well as specific boundary constraints),

Table 6 Confusion matrix of TA-Regex

Actual \ Predicted	Title	Author	Reference	Comment	Pages	Location	Year	Editor	Material	-
Title	3377	0	0	38	0	0	0	10	0	198
Author	51	758	0	3	0	0	0	0	0	114
Ref.	0	0	1247	0	0	0	0	0	0	222
Comment	15	0	0	1925	0	0	0	0	0	23
Pages	0	0	0	10	232	0	0	0	0	121
Location	0	0	0	5	0	155	0	0	0	52
Year	0	0	0	3	0	0	147	0	0	15
Editor	119	0	0	10	0	0	0	13	0	32

Table 7 Confusion matrix of ParsCit

Actual \ Predicted	Title	Author	Reference	Comment	Pages	Location	Year	Editor	Material	–
Title	3534	6	0	0	4	0	0	1	78	0
Author	109	813	0	0	0	0	0	3	1	0
Reference	185	1	1176	62	7	6	6	26	0	0
Comment	731	1	29	1158	21	7	5	6	2	3
Pages	4	1	2	1	344	0	3	0	6	2
Location	24	0	0	4	0	179	0	5	0	0
Year	1	0	0	0	1	2	161	0	0	0
Editor	31	0	0	0	0	0	0	143	0	0
Material	19	0	6	2	5	0	0	0	43	2

Table 8 Confusion matrix of JI-Field-TA

Actual \ Predicted	Title	Author	Reference	Comment	Pages	Location	Year	Editor	Material	-
Title	3456	23	19	5	8	11	0	0	12	89
Author	33	876	0	0	0	0	0	3	3	11
Reference	10	0	1375	0	3	4	2	20	17	38
Comment	12	0	9	1859	4	1	0	0	4	74
Pages	8	0	0	0	320	2	0	0	16	17
Location	9	0	0	0	0	180	0	1	5	17
Year	5	0	0	0	1	0	149	0	4	6
Editor	28	0	0	0	0	8	0	120	6	12
Material	0	0	3	0	0	0	0	0	74	0

we obtain highly informed classifications that integrate contextual information across large distances, and even across complete citation entries.

Error type classes. These insights are corroborated by analyzing differences in the distribution of **error type classes** across the different systems.[41]

Table 9 compares the distribution of error classes overall and for different fields in **ParsCit** compared to the **MLN-Local** and the joint MLN formalizations.[42] Comparing error classes in the local versus joint MLN formalizations, we observe important reductions of errors of type (IV) for the fields Author, Pages, and Location. We note a higher error rate of type (IV) for Materials in all MLN variants, however at much higher correct assignment rates. Overall, all error classes remain pretty stable across these systems across all fields. What is striking, however, is a strong divergence

Table 9 Distribution of error types (in percent)

Fields	ParsCit					MLN-Local					JI-Cit-TA					JI-Field-TA				
	Corr.	I	II	III	IV	Corr.	I	II	III	IV	Corr.	I	II	III	IV	Corr.	I	II	III	IV
Title	60.2	1.0	33.3	0.5	5.0	80.0	4.2	11.8	2.6	1.4	80.7	2.8	12.7	2.6	1.2	82.0	2.9	10.6	2.9	1.7
Author	76.1	9.3	1.1	0.0	13.5	86.8	6.0	1.3	0.3	5.7	87.9	5.5	2.4	0.0	4.2	87.7	5.3	2.1	0.0	4.8
Reference	68.7	18.1	9.2	0.4	3.6	94.4	3.6	0.4	0.4	1.2	94.4	4.0	0.4	0.0	1.2	94.4	3.6	0.4	0.4	1.2
Comment	61.6	15.9	15.9	2.2	4.3	93.4	3.0	1.0	0.0	2.5	93.0	3.0	1.0	0.5	2.5	93.9	2.0	1.0	0.5	2.5
Pages	56.9	6.6	13.2	0.6	22.8	90.1	0.6	0.0	0.6	8.7	90.6	1.9	0.0	0.0	7.5	90.3	1.9	0.6	0.0	7.1
Location	82.4	8.8	1.8	1.2	5.9	87.6	5.0	2.5	0.6	4.3	85.7	7.5	3.1	0.0	3.7	87.3	5.7	2.5	0.0	4.4
Year	90.5	1.2	1.8	0.6	6.0	94.8	1.9	0.0	0.0	3.2	97.4	0.7	0.0	0.0	2.0	96.0	0.7	0.7	0.0	2.6
Editor	67.6	2.7	10.8	2.7	16.2	62.9	17.1	2.9	2.9	14.3	63.9	16.7	0.0	2.8	16.7	66.7	13.9	0.0	2.8	16.7
Material	24.1	20.7	3.4	27.6	24.1	45.7	0.0	17.4	0.0	37.0	46.7	0.0	17.8	0.0	35.6	47.7	0.0	18.2	0.0	34.1
All	69.1	8.0	12.8	1.1	8.9	86.5	4.1	3.9	0.9	4.6	87.0	3.9	4.4	0.7	4.0	87.5	3.5	3.9	0.8	4.3

I: actual field contains assigned field, II: assigned field contains actual field, III: other kind of overlap, IV: no overlap.
Corr. indicates the percentage of correct assignments.

of error type distribution between the MLN formalizations and the *ParsCit* baseline, which exhibits a strong tendency for errors of type (II): 12.8% versus 3.9% in **JI-Field-TA**. These errors are particularly pronounced for Title, Comment, Pages, and Editor fields. This nicely illustrates the capacity of the MLN formalizations to resolve difficult field boundary decisions (especially white space/tabular delimiters) that separate Title from other fields.

Improved delimitation of these fields results in overall gains of *exact field match* scores of the MLN setups, with strong performance gains in both recall and precision (cf. Tables 4 and 5).

Evaluation Results for Experiment II. The rule set in Experiment I was manually assigned rule weights that were tested on the development set. Since grid search for optimization is costly when dealing with substantial rule set sizes, we contrast the performance of MLN-based segmentation with uniform rule weighting. In **MLN-uniform**, we treat all rules as soft rules with weight 1; the results are shown in Table 10. In this setting, we obtain performance results approximately similar to supervised training with the CRF-based baseline setting **ParsCit** (cf. Table 5).

5.3 Discussion

In summary, our MLN formalizations clearly outperform state-of-the-art statistical and rule-based methods for citation segmentation on a data set that is specific to a Digital Humanities context. With citation entries obtained from a manually type-set and later OCRed multilingual bibliography, this data set features specific challenges: (i) low redundancy, (ii) noise from OCR as well as (iii) structural inconsistencies.

State-of-the-art methods in a special DH setting. The two **baselines** we provide highlight the difficulties of this data set: *rule-based parsing* using regular expressions achieves high precision, but low recall at an overall moderate level. This is typical and especially prominent in the current setting, due to noise and inconsistencies in the data. This traditional method is costly and cumbersome, suffers from lack of robustness, and cannot be easily ported to related application scenarios.

Applying the *CRF-based ParsCit* system to our data, we obtain the worst overall results with lowest precision and recall. This still holds if the system is trained on a doubled training set size. The dramatic performance drop of *ParsCit* compared to its performance on standard data sets such as *Cora* highlights the difficulty of the present data set (cf. Councill *et al.*, 2008).[43] Our error analysis points to a particular weakness of CRF-based approaches such as *ParsCit*, which cannot exploit contextual evidence of a more global nature, thus are often misguided by non-contextualized local field characteristics.

MLN for segmentation with sparse and noisy data. Compared to these two strong baseline systems, all MLN formalizations, *local* and *joint*, show statistically significant improvements by margins of

Table 10 Experiment II

Measures	**JI-Field-TA**				**MLN-uniform**			
	P	**R**	**F_1**	**Acc.**	**P**	**R**	**F_1**	**Acc.**
Token	92.9	89.9	91.3		88.7	83.8	85.3	
Field (macro-average)	88.3	87.7	88.0		71.3	72.7	72.0	
Field (micro-average)	90.6	90.9	90.7		80.2	82.2	81.2	
Citation (exact)				75.8				71.5

Segmentation results on evaluation set for MLN-Uniform (compared to JI-Field-TA). Different evaluation measures: P/R/F_1 for token (= word) assignment, exact field match, and exact citation match

+9.2, +8.8, and +10.7 pp. F_1-score against **TA-Regex** and of +14.3, +13.9, and +15.8 pp. F_1-score against **ParsCit** in a competitive training setting.

Lack of data redundancy and noise are clearly factors for the low performance of the baselines, affecting both rule-based and CRF-based models. The MLN formalizations, in contrast, are more robust against noise, due to the global nature of constraints combined with statistical inference. By encoding knowledge about specific citation types using partially specified field order, field-specific features, and constraints, as well as selected (secure) demarcation signals, even a model that formalizes citation structure by looking at individual citations only achieves significant performance gains, both in recall and precision. The knowledge-based encoding of citation structure seems well suited for low-redundancy data, and can achieve high precision without requiring training data. On the other hand, since individual rules can be violated in view of contrasting evidence stemming from complementary rules, the formalization is robust against noise.

Joint inference alleviates sparsity and noise. The *joint inference* settings further improve this performance, by importing knowledge from limited data redundancy at the *citation level* (TA-internally or imported from an external bibliographic source) but more clearly by exploiting global knowledge across entries at the *field level*.

6 Conclusion

In this article, we address the problem of citation segmentation on a novel data set that is typical for Digital Humanities applications. Using the *TA* as a showcase, we present a *knowledge-based statistical inference approach* for citation segmentation applied to sparse and noisy data from an OCRed multilingual bibliography.

Our account is based on *MLNs* and *joint inference* and extends prior work of Poon and Domingos (2007) by applying joint inference at the field level. We show that in the face of sparse and noisy data, this method clearly outperforms classical methods for citation segmentation, including a supervised account using *ParsCit* as a state-of-the-art CRF-based baseline.

Our account is knowledge based and thus requires a formalization of the (rather complex) citation structure of the *TA*. Markov logic, a formalism of first-order predicate logic, is very suitable for this task, as it allows for a transparent encoding using modular and declarative rules. We designed rules of different types that identify citation fields locally within references, using relative, absolute, and partial ordering information, characterizing features and restrictions for particular fields, and also encoding of secure boundary information. We further included rules that import additional knowledge stemming from redundancies at the citation or field levels using joint inference.

The MLN formalizations we offer comprise only 65 rules and 16 defining predicates. This is a modest size, given the considerable complexity of the *TA*'s citation structures. These rules can be flexibly adapted to citation structures of other bibliographic sources, or structured entries of digitized dictionaries or encyclopedias.

Our MLN formalizations achieve high-performance results for citation segmentation, measured in terms of *exact field match*. Given the difficult nature of the data, an overall F_1-score of 88.0% with precision at 88.3% can be considered as highly satisfying. In fact, the segmentation results for the entire *TA* have been compiled to a relational database and build the back end for a searchable Web interface to the *TA*.[44]

In summary, citation segmentation using MLNs is a powerful method that proves particularly suited for digitization challenges encountered in Digital Humanities contexts, being confronted with *noise*, *multilinguality*, *and lack of consistency and data redundancy*. The MLN framework is particularly attractive in this setting, as it allows for transparent formalization in a first-order logic formalism and flexible adaptation to related settings—while proving robust against noise and inconsistencies.

The system we propose is not dependent on manually labeled training data, requiring only little annotated data for system development and setting of rule weights. As a *knowledge-based method*, it avoids costly and brittle parser development as well as tedious manual annotation of large training data sets. It allows for transparent rule design with flexible adaptation from existing rule sets when addressing related problems. In fact, we anticipate that quite a number of digitization projects in the Humanities that are based on structured data such as bibliographies, or further types of digitized sources such as historical dictionaries or encyclopedias can very well profit from using the technologies described in this article.

Acknowledgements

This work was partly supported by a grant from the Cluster of Excellence: ‘Asia and Europe in a Global Context’ at Heidelberg University. Special thanks go to the project’s research assistants Mateusz Dolata and Nicolas Bellm for their work on rule-based citation parsing and database design and population.

References

Afzal, M. T., Maurer, H., Balke, W. T., and Kulathuramaiyer, N. (2010). Rule-based autonomous citation mining with TIERL. *Journal of Digital Information Management*, **8**(3): 196–204.

Besagni, D. and Belaid, A. (2004). Citation recognition for scientific publications in digital libraries. In *Proceedings of the First International Workshop on Document Image Analysis for Libraries, DIAL’04*. Palo Alto, CA, pp. 244–52.

Bia, A., Malonda, J., and Botella, F. (2006). A multilingual markup translation web-service - an entry level solution to internationalize XML markup vocabularies. In Cordeiro, J. A. M. *et al.* (eds), *WEBIST (1)* INSTICC Press, pp. 63–68. http://dblp.uni-trier.de/db/conf/webist/webist2006-1.html#BiaMB06 (accessed 14 July 2013).

Casinius, S. and Sporleder, C. (2007). Bootstrapping Information Extraction from Field Books. In *Proceedings of EMNLP-CoNLL 2007*. Prague: Czech Republic, pp. 827–36.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**: 37–46.

Cortez, E., da Silva, A. S., Gonçalves, M. A., Mesquita, F., and de Moura, E. S. (2007). FLUX-CIM: flexible unsupervised extraction of citation metadata. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*. Vancouver, Canada, pp. 215–24.

Councill, I. G., Giles, C. L., and Kan, M.-Y. (2008). ParsCit: an open-source CRF reference string parsing package. In *Proceedings of LREC 2008*. Marrakech, pp. 661–7.

Cowie, J. R. and Lehnert, W. G. (1996). Information extraction. *Communications of the ACM*, **39**(1): 80–91.

Denis, P. and Baldridge, J. (2009). Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, **42**: 87–96.

Domingos, P. and Lowd, D. (2009). Markov Logic. An interface layer for artificial intelligence. In Brachmann, R. R. and Dietterich, T. (eds), *Synthesis Lectures on Artificial Intelligence and Machine Learning*. San Rafael, CA, USA: Morgan & Playpool.

Finkel, J. R. and Manning, C. D. (2010). Hierarchical Joint Learning: Improving Joint Parsing and Named Entity Recognition with Non-Jointly Labeled Data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, pp. 720–8.

Giles, C. L., Bollacker, K. D., and Lawrence, S. (1998). Citeseer: an automatic citation indexing system. In

Proceedings of the 3rd ACM Conference on Digital Libraries. Pittsburg, Pa, USA: ACM Press, pp. 89–98.

Grenager, T., Klein, D., and Manning, C. D. (2005). Unsupervised Learning of Field Segmentation Models for Information Extraction. In *Proceedings of the 43rd Annual Meeting of the ACL*. Ann Arbor, pp. 371–8.

Groza, T., Grimnes, A., and Handschuh, S. (2012). Reference Information Extraction and Processing Using Random Conditional Fields. *Information Technology and Libraries*, **31**(2): 6–20.

Hazai, G. and Kellner-Heinkele, B. (eds), (1975). *Turkologischer Anzeiger,* Universität Wien. Institut für Orientalistik and Universität Wien. Orientalisches Institut. http://orientalistik.univie.ac.at/forschung/publikationen/turkologischer-anzeiger (accessed 19 July 2013).

Jewell, M. (2000). *ParaCite: An Overview*. http://paracite.eprints.org/docs/overview.html. (accessed 19 July 2013).

Kim, Y.-M., Bellot, P., Faath, E., and Dacos, M. (2012a). Annotated Bibliographical Reference Corpora in Digital Humanities. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. Istanbul, Turkey, pp. 494–501.

Kim, Y.-M., Bellot, P., Faath, E., and Dacos, M. (2012b). Automatic annotation of incomplete and scattered bibliographical references in Digital Humanities papers. In *Proceedings of Conférence en Recherche d'Information, Applications (CORIA)*. Bordeaux, France, pp. 329–40.

Lawrence, S., Giles, C. L., and Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *Computer*, **32**(6): 67–71.

Li, Q., Anzaroot, S., Lin, W.-P., Li, X., and Ji, H. (2011). Joint inference for cross-document information extraction. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. Glasgow, United Kingdom: ACM, pp. 2225–8.

McCallum, A., Nigam, K., Rennie, J., and Seymore, K. (1999). A machine learning approach to building domain-specific search engines. In *Proceedings of IJCAI*, pp. 662–7.

Niu, F., Ré, Ch., Doan, A., and Shavlik, J. (2011). Tuffy: scaling up statistical inference in Markov logic networks using an RDBMS. *Proceedings of the VLDB Endowment*, **4**(6): 373–84.

Padó, S. (2006). *User's guide to sigf: Significance testing by approximate randomisation*. http://www.nlpado.de/∼sebastian/software/sigf.shtml (accessed 19 July 19 2013).

Park, S. H., Ehrich, R., and Fox, E. (2012). A hybrid two-stage approach for discipline-independent canonical representation extraction from references. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*. Washington, D.C., WA, USA, pp. 285–94.

Peng, F. and McCallum, A. (2004). Accurate information extraction from research papers using Conditional Random Fields. In *Proceedings of HLT-NAACL*. Boston, MA, USA, pp. 329–36.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*. Istanbul, Turkey, pp. 2089–96.

Poon, H. and Domingos, P. (2007). Joint Inference in information extraction. In *Proceedings of the Twenty-Second National Conference on Artificial Intelligence*. Vancouver, Canada: AAAI Press.

Poon, H. and Domingos, P. (2008). Joint unsupervised co-reference resolution with markov logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, pp. 650–9.

Poon, H. and Vanderwende, L. (2010). Joint inference for knowledge extraction from biomedical literature. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Los Angeles, California: Association for Computational Linguistics, pp. 813–21.

Richardson, M. and Domingos, P. (2006). Markov logic networks. *Machine Learning*, **62**(1): 107–36.

Riedel, S. and Meza-Ruiz, I. (2008). Collective semantic role labelling with markov logic. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Manchester, England, pp. 193–7.

Roth, D. and Yih, W. (2004). A linear programming formulation for global inference in natural language tasks. In *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*. Boston, Massachusetts, USA, pp. 1–8.

Seymore, K., McCallum, A., and Rosenfeld, R. (1999). Learning hidden Markov model structure for information extraction. In *Proceedings of the AAAI Conference*. Orlando, FL, USA, pp. 37–42.

Takasu, A. (2003). Bibliographic attribute extraction from erroneous references based on a statistical model. In *Proceedings of the 3rd ACM/IEEE-CS joint* *conference on Digital libraries*. JCDL'03. Washington, DC, USA: IEEE Computer Society, pp. 49–60.

Yeh, A. (2000). More accurate tests for the statistical significance of result differences. In *Proceedings of COLING 2000*. Saarbrücken, Germany, pp. 947–53.

Notes

[1] http://www.loc.gov/standards/mods

[2] See Hazai and Kellner-Heinkele (1975) and the digitized online resource created on the basis of the present work: http://kjc-fs2.kjc.uni-heidelberg.de:8000/en/.

[3] This number is based on estimations for the first 26 volumes as used in the experiments reported below.

[4] See for instance the *Index Islamicus* database that offers such functionalities, at http://search.proquest.com/indexislamicus.

[5] The project from which the present work arose was a joint effort by members from different institutions at Heidelberg University, namely the *Department of Languages and Cultures of the Near East* (Islamic Studies), the *Department of Computational Linguistics*, and the *Heidelberg Research Architecture* (HRA)—the Digital Humanities Unit at the Cluster of Excellence 'Asia and Europe in a Global Context' at Heidelberg University. Based on an agreement with the TA editorial board to provide open access to the new database, the project group was formed and received funding from the Cluster of Excellence. The project results including access to the database are available at http://www.asia-europe.uni-heidelberg.de/en/research/heidelberg-research-architecture/detail/m/turkology-annual-online.html.

[6] Volumes 27 and 28 had not been available at that time and have not been processed within the project.

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

[8] The CiteSeer data were created by Lawrence *et al.* (1999).

[9] Cf. McCallum *et al.* (1999) and http://www.cs.umass.edu/∼mccallum.

[10] For instance, Councill *et al.* (2008) report a 95.7% F_1-score for field assignment for *ParsCit* on the Cora data set (in a CV setting on 200 references). In Cora, all citations are non-singletons. Groza *et al.* (2012) further improve on *ParsCit* results on the same data set. In comparison, *ParsCit* trained on 200 items of our TA data set (and evaluated on held-out data, cf. Section 5.1) yields a field F_1-score of 76.7%.

[11] An example of related projects in *lexicography* is the German Loan Word Portal of the Institute for German Language (Institut für Deutsche Sprache, IDS) Mannheim at http://lwp.ids-mannheim.de. The project digitizes and connects multiple lexica of German loan words in foreign languages and makes them searchable cross-lingually in flexible ways.

Another example of a highly precious lexicon resource of large scale and with highly complex entry structure and access functions is the French Etymological Dictionary (Französisches Etymologisches Wörterbuch, FEW) first edited by Walter von Wartburg. Started in 1922, it was finished in 2002 and is considered the most important etymological lexicon for the gallo-roman language family. It comprises 25 volumes with more than 17,000 pages. The FEW has been digitized and is offered for electronic search at http://stella.atilf.fr/few.

[12] See for example Casinius and Sporleder (2007), Section 3.

[13] The terms *citation* and *reference* are often used interchangeably.

[14] In fact, novel developments such as the 'Universal Part-of-Speech Tagset' (Petrov *et al.*, 2012) could open ways for a PoS-based approach. However, truly multilingual (i.e. mixed-language) PoS sequence tagging by itself is a largely unexplored problem, and it is unclear whether the coarse nature of the tagset would be informative enough for our setting.

[15] See e.g. Kim *et al.* (2012a).

[16] MLNs have been applied to a variety of problems in NLP, for instance, coreference resolution (Poon and Domingos, 2008) or semantic role labeling (Riedel and Meza-Ruiz, 2008), and information extraction tasks (Poon and Vanderwende, 2010).

[17] The distribution is estimated from the analyses returned by one of our baseline systems, a manually developed regex-based entry parser.

[18] We concentrated on the most prominent entry types and ignore conference proceedings in what follows.

[19] Such rules are assigned infinite rule weights.

[20] Free (previously unmentioned) variables are universally quantified.

[21] Delimiters are not represented as tokens, and hence do not have indices of their own.

[22] We determined such tendencies by data inspection on the development set. Calculating the distribution on the evaluation set (after the fact), such tendencies were in general confirmed: for example, 90.3% of all occurrences of `year` were surrounded by commas.

[23] Cf. Table 2 for an overview of features used.

[24] The overlap could also be determined during inference. However, this would increase the size of the resulting MLN considerably.

[25] http://worldcat.org

[26] The rule below processes `lastName` and `firstName` tokens in a `CURRENT` entry and refers to a field `Fauthor` with matching `lastName` and `firstName` tokens in the clearly identified starting positions of a (disjoint) reference entry `bib`. If such a reference can be found, the rule confirms the corresponding `Fauthor` field in `CURRENT`.

[27] Rules of this type are expensive in inferencing if a large number of citations must be considered. This problem can be alleviated by preprocessing the set of citations to determine a subset of candidates to be considered for joint inference rules: those that show significant string overlap with the relevant field values of the citation under investigation.

[28] We identified 19 languages and assigned their special characters to one language code (Xhosa).

[29] This training set size is roughly equivalent to the number of training instances used in standardized evaluations. For evaluations on the Cora data set, for instance, Councill *et al.* (2008) employed a gold standard of 200 citations in a cross-validation setting for training and evaluating *ParsCit*. We train on 200 entries and evaluate against our held-out evaluation set.

[30] F_β-score with $\beta = 1$ computes the harmonic mean of precision and recall:

$$F_{(\beta=1)} = (1+\beta^2)\frac{PR}{\beta^2 P + R}$$

[31] In this setting, we do not employ additional citation entries from TA-internal repetitions or WorldCat, in order to measure individual performance differences for joint inference at the field level.

[32] We experimented with both JI-Cit-TA and JI-Cit-WCat. Since the results were very close, and given that JI-Cit-WCat depends on external resources, we only report the results for the slightly better performing JI-Cit-TA for comparison in Table 4 (see discussion below and footnote 35).

[33] We report micro- and macro-average over all field types. Due to the predominance of certain fields across the evaluation set (cf. Table 1), micro-average shows a bias towards large classes. For discussion, we therefore concentrate on macro-average. The relative performance shows similar tendencies across both measures.

[34] See footnote 9 above. Councill *et al.* (2008) report 95.7% F_1-score for field assignment for *ParsCit* on the Cora data set (in a CV setting on 200 references). In Cora, all citations are non-singletons. Groza *et al.* (2012) further improve on *ParsCit* results on the same data set. Our *ParsCit* model is trained on 200 citations (cf. Section 5.1) and evaluated on the held-out evaluation data set.

[35] The results for individual field categories are contrasted below.

	JI-Cit-TA			**JI-Cit-WCat**		
	P	**R**	**F_1**	**P**	**R**	**F_1**
Title	81.2	81.4	81.3	82.3	**82.3**	82.3
Author	88.9	90.1	89.5	89.5	**90.9**	90.2
Ref	94.4	**94.4**	**94.4**	94.4	**94.4**	**94.4**
Comm.	93.0	**91.6**	92.3	**93.4**	**91.6**	**92.5**
Pages	93.1	**90.8**	**91.9**	91.4	**90.8**	91.1
Location	84.1	85.7	84.9	87.1	**88.2**	**87.7**
Year	**98.1**	93.8	**95.9**	95.6	**95.0**	95.3
Editor	**66.7**	**66.7**	**66.7**	48.9	**63.9**	55.4
all	87.4	86.8	**87.1**	85.3	**87.1**	86.1

[36] Again, we concentrate on macro-average for field assignment measures.

[37] For full citation match, precision equals recall.

[38] Compared to all MLN formalizations, as displayed in Table 4, the performance of ParsCit400 is significantly lower (at $p = 0.00009$), the same holds in comparison to all baselines in Table 4.

[39] Comments are mostly chunks of text, hence the similarity to titles.

[40] For instance, relative order of fields; positive and negative constraints on field occurrence in particular types of entries, etc. (cf. Section 4.2).

[41] Cf. Section 5.1 for the definitions of error classes.

[42] The Material field was not modeled for TA-Regex, therefore it was not taken into account in previous comparisons.

[43] In fact, when replicating Councill *et al.* (2008)'s CV setting on 200 entries on our data, we obtain an overall F_1-score for field assignment of 69.3% on our data, compared to their 97.5% on the Cora data set, with 13 as opposed to our overall 10 field types.

[44] The *Turkology Annual Online* is frequently consulted, with about 510 visitors recorded per month (March to May 2013).