Local Adaptive Extraction of References

Peter Kluegl, Andreas Hotho, and Frank Puppe

University of Würzburg,
Department of Computer Science VI
Am Hubland, 97074 Würzburg, Germany
{pkluegl, hotho, puppe}@informatik.uni-wuerzburg.de

Abstract. The accurate extraction of scholarly reference information from scientific publications is essential for many useful applications like BIBTEX management systems or citation analysis. Automatic extraction methods suffer from the heterogeneity of reference notation, no matter wether the extraction model was handcrafted or learnt from labeled data. However, references of the same paper or journal are usually homogeneous. We exploit this local consistency with a novel approach. Given some initial information from such a reference section, we try to derived generalized patterns. These patterns are used to create a local model of the current document. The local model helps to identify errors and to improve the extracted information incrementally during the extraction process. Our approach is implemented with handcrafted transformation rules working on a meta-level being able to correct the information independent of the applied layout style. The experimental results compete very well with the state of the art methods and show an extremely high performance on consistent reference sections.

1 Introduction

Extracting references from scientific publication is an active area of research and enables many interesting applications. One example is the analysis of cited publications and the resulting citation graph, another is the automatic creation of an internet portal [1]. In general, the extracted fields of a reference are equal to the fields of the well-known BIBTEX format which are used in many applications, e.g. in the social publication sharing systems Bibsonomy[1]. In order to identify duplicate or new references in these applications hash keys could be applied [2] on a restricted number of the BIBTEX fields, i.e. *Author*, *Title*, *Editor* and *Date*, moving these fields into the focus of this paper. Machine learning and sequence labeling approaches [3, 4] are often used for such an extraction task. These methods learn a statistical model using training sets with labeled data and apply these models on newly and unseen documents. No information of the unlabeled documents is used to adapt such models in its application phase which is a strong limitation of these approaches. The heterogeneous styles of the references make a suitable generalization difficult and decrease the accuracy of the extraction task.

This paper introduces a local adaptive information extraction approach that improves the extracted information online by using information of unlabeled documents directly during the extraction task. The proposed approach utilizes the homogeneity of

[1] http://www.bibsonomy.org/

references in one document which are defined by the author and by the prescribed style guides of the journal or conference. Our proposed solution to handle heterogeneous documents is based on a short term memory and on the analysis of the occurring patterns in the document. Handcrafted rules are then applied on these local patterns and provide an automatic adaption on the internal previously unknown consistency of the documents. This approach considerably improves the accuracy of the extracted references and can be applied in all domains with documents containing several expected information created by the same creation process (c.f. [5]).

The rest of the paper is structured as follows: Section 2 describes the transformation-based extraction technique, the basic idea and the application of the proposed approach. Then, the evaluation setting and experimental results are presented and discussed in section 3. Section 4 gives a short overview of the related work and section 5 concludes with a summary of the presented work.

2 Method

2.1 Transformation-based Information Extraction

Brill [6] introduced one of the first transformation-based natural language processing approaches for the part-of-speech tagging task. The transformation rules were learnt with a gold standard and corrected the errors of a simple, initial tagger. Nahm [7] utilized this technique to increase the recall of precision-driven matching rules in an information extraction task. The TEXTMARKER system [8] provides a language and development environment for the transformation-based approach. The handcrafted rules match on a combination of annotations and create new annotations or manipulate existing annotations. However, in contrast to other rule engines for information extraction, the rules are executed similar to an imperative programming paradigm in the order they are listed. Each rule element of a TEXTMARKER rule consists of a matching condition and of an optional quantifier, condition part and action part. The following rule with two rule elements tries to match on an *Author* (first rule element) and on a following *EditorIndicator* annotation (second rule element). If the match was successful, then the *Author* annotation is removed and a new *Editor* annotation is created[2].

```
Author{-> UNMARK(Author)} EditorIndicator{-> MARK(Editor,1,2)};
```

This transformation-based approach is an elegant way to create information extraction applications and enables the knowledge engineer to reuse existing information rather efficiently. Additionally to its comprehensible and extensible rule language, the TEXTMARKER system alleviates the knowledge acquisition bottleneck by supporting the knowledge engineer with a test-driven development and rule learning framework.

2.2 Basic Idea of Local Adaptivity

A text document with references like scientific publication is often created in a single creation process, e.g., an author writes an article by using LaTeX. These documents contain similar styles, e.g., an author uses in a single document always the same layout for

[2] A detailed description of the syntax and semantic of the TEXTMARKER language can be found at http://tmwiki.informatik.uni-wuerzburg.de/

headlines or the used BIBTEX style determines the appearance of the references in a paper. Patterns describing these similarities and regularities can be detected and used to improve the extracted information. However, patterns can vary strongly between different documents in a domain. A common problem with extracting references is the different style resulting in conflicting patterns for global models, e.g., different separators between interesting fields or different sequences of the fields. Based on prior work [5], the approach presented in this paper tries to apply a short term memory to exploit these intra-document similarities. A simple information extraction component extracts instances and fields of information similar to transformation-based, error-driven techniques [6, 7]. The next phase however does not simply try to correct errors using a gold standard, but rather applies a local adaption process (AIE). An analysis of the occurring patterns returns a set of conflicting information that can be corrected using our transformation-based method. This approach works best if the assumption holds that the document contains several instances of information and was created in the same process. The whole process is depicted in figure 1.

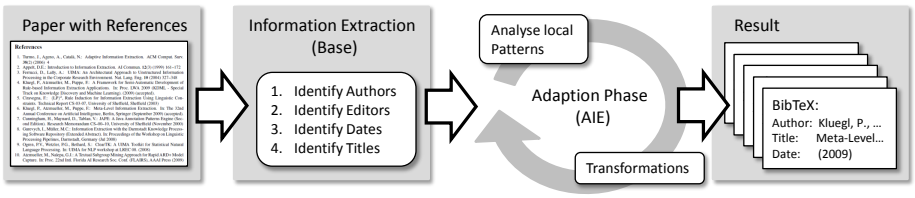

Fig. 1. Overview of the adaptive process with initial information extraction, analysis step and transformation phase.

2.3 Implementation

Both components the BASE and the meta approach were handcrafted using the TEXTMARKER system. The BASE component consists of a simple set of rules for the identification of the *Author*, *Title*, *Editor* and *Date*. For a quick start 10 references of the CORA dataset (cf. sec. 3.2) are used to handcraft these initial rules. The given features are token classes and word lists for months, first names, stop words and keywords, e.g., indicators for an editor. Additional features are created during the transformation process which makes it impossible provide a complete list of them in this paper. After some reasonable rules were handcrafted, 11 randomly selected papers of the workplace of the authors are used to develop the adaptive component AIE and to refine the rules of the BASE component (cf. dataset D_{Dev} in sec. 3.2). The separators located at the beginning and the end of the BIBTEX fields and the sequence of the fields are sufficient for a description of the applied creation process. Therefore, handcrafted rules aggregate the separators and sequences of the initialy extracted information to the local model by remembering the most occurring separator and neighbors of each field. This local model is compared with all initially extracted information for an identification of possible conflicts. After this analysis phase, transformation rules are applied that utilize the

local model and the information about conflicts in order to restore the consistency of the extracted information and therefore to correct the errors of the BASE component. Processing a reference sections with several conflicting style guides requires the adaptation of our rules to improve the robustness of the approach. Therefore, the rules also rely on normal patterns of the domain, e.g., the *Author* normally contains no numbers. This analysis and transformation step is iterated if necessary in order to incrementally improve the extracted information. The complete set of used rules is available on the web.[3]

2.4 Example

The behavior of our proposed approach is illustrated along the following example which is a typical text fragment representing a single reference of a reference section:

> Kluegl, P., Atzmueller, M., Puppe, F., TextMarker: A Tool for Rule-Based Information Extraction, Biennial GSCL Conference 2009, 2nd UIMA@GSCL Workshop, 2009

A reasonable result of the BASE component contains four errors: "TextMarker:" is part of the *Author* and is missing in the *Title*. The *Title* and *Date* contain additional tokens of the conference. Inline annotations for the *Author* (**<A>**), *Title* (**<T>**) and *Date* (**<D>**) are used as a simple description:

> **<A>**Kluegl, P., Atzmueller, M., Puppe, F., TextMarker:**</A> <T>**A Tool for Rule-Based Information Extraction, Biennial GSCL Conference**</T> <D>**2009**</D>**, 2nd UIMA@GSCL Workshop, **<D>**2009**</D>**

The analysis of the document results in the following patterns describing the internal consistency of the references: Most of the identified *Author* and *Title* annotations end with a comma. The *Title* follows directly after the *Author* and the *Date* is located near the end of the reference. With this information at hand, conflicts for the first *Date*, the end of the *Author* and *Title* are identified. Several handcrafted transformation rules are applied to solve these conflicts. For our example, the *Author* and *Title* are both reduced to the last comma and only the last *Date* is retained. The following exemplary rule shifts the end of the *Author* by a maximum of four tokens (*ANY*) to the next separator listed in the local model (*EndOfAuthor*) if a conflict was identified at the end of the *Author* annotation (*ConflictAtEnd*):

```
ANY+{PARTOF(Author) -> UNMARK(Author), MARK(Author,1,2)}
  EndOfAuthor ANY[0,4]?{PARTOF(Author)} ConflictAtEnd;
```

However, after these changes the *Title* does not directly follow the currently detected *Author* field. Therefore the *Title* is expanded which results in the following correct annotation of our example:

> **<A>**Kluegl, P., Atzmueller, M., Puppe, F.,**</A> <T>**TextMarker: A Tool for Rule-Based Information Extraction,**</T>** Biennial GSCL Conference 2009, 2nd UIMA@GSCL Workshop, **<D>**2009**</D>**

[3] http://ki.informatik.uni-wuerzburg.de/~pkluegl/KI2010

The transformation rules used to correct the errors of the BASE level approach match on a meta-level. They are completely independent of the specific local patterns detected for the current document. This is an immense advantage of our approach, because the same rules will work with different and unknown separators or field sequences.

3 Experimental Study

3.1 Performance Measures

Following the evaluation methodology of other reference extraction publications (cf. [3]) we apply the word level F_1 measure, defined as follows:

$$Precision = \frac{tp}{tp+fp}, \quad Recall = \frac{tp}{tp+fn}, \quad F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

The true positives (tp) refer to the amount of all correctly identified alpha-numeric tokens within a labeled field, the false positives (fp) to the number of tokens erroneously assigned to a field and the false negative (fn) to the amount of erroneously missing tokens. As a second measure we will use the $Instance$ accuracy which is the percentage of references in which all fields are correctly extracted.

3.2 Datasets

For a comparable evaluation of the presented approach we refer to the commonly used CORA data set (D_{Cora}: 500 References, D_{Cora}^{All}: 489 References[4]) [1]. However, this data set is not directly applicable for the development and evaluation of the approach as the CORA data set does not contain information about the original document of every reference which we need to derive the local patterns. Therefore, a simple script was developed in order to reconstruct reference sections as they would occur in real publications using only references originated in the available data set. D_{Cora}^{Paper} contains 299 references of D_{Cora}^{All} in 21 documents and represents a selection of papers with a strict style guide applied. Due to the simplicity of the assignment script and the distribution of the reference styles in the dataset a considerable amount of references could not be assigned to a paper. The data set D_{Cora}^{Rest} contains the missing 190 references and the data set D_{Cora}^{All} is the union of D_{Cora}^{Paper} and D_{Cora}^{Rest}. The data set D_{Dev} for the development of the adaptive component was created using 11 randomly selected publications and contains 213 references. All additional data sets are available for download.[5]

3.3 Experimental Results

Table 1 contains the evaluation results of the simple rule-based component (BASE) and the combination with the adaption phase (AIE). The word level F_1 measure was applied on a single field and the instance accuracy was calculated for the complete

[4] D_{Cora} without 10 references (line 100-109) for the development of the BASE component and one reference with damaged markup.

[5] http://ki.informatik.uni-wuerzburg.de/~pkluegl/KI2010

BIBTEX instance. AIE was only applied on the data set D_{Cora}^{Paper} which are 61% of all references. BASE reached an average F_1 score of 95.3% and an instance accuracy of 85.4% on the development set D_{Dev}. We tuned AIE to achieve an average F_1 score of 100.0% and an instance accuracy of 100% on D_{Dev}.

Table 1. Experimental results of the normal rule set (BASE) alone and with the adaption phase (AIE) for the additional data sets. The results for the data set D_{Cora}^{Rest} always refer to the BASE component alone since no adaption was applicable.

	BASE			AIE			Peng[3]	ParsCit[4]
	D_{Cora}^{Paper}	D_{Cora}^{Rest}	D_{Cora}^{All}	D_{Cora}^{Paper}	D_{Cora}^{Rest}	D_{Cora}^{All}	D_{Cora}	D_{Cora}
Author	98.4	98.3	98.4	99.9	98.3	99.3	99.4	99.0
Title	96.4	95.9	96.2	99.2	95.9	97.9	98.3	97.2
Editor	100.0	92.9	94.9	100.0	92.9	94.9	87.7	86.2
Date	98.1	95.8	97.3	99.9	95.8	98.3	98.9	99.2
Average	98.3	95.7	96.7	99.7	95.7	97.6	96.1	95.4
Instance	89.0	81.6	86.1	98.7	81.6	92.0	-	-

The BASE component yields results for all data sets which are comparable with knowledge-based approaches. The results of the machine learning methods are considerably better. Merely the score of the *Editor* field outperforms the related results and implies that transformation rules seem to be very suitable for this task. The overall lower performance for the *Date* field is caused by the fact that the development data set contains no date information with a time span. Hence the BASE component missed several true positives of the test data sets.

The adaptive approach is able to improve the accuracy of the initial rule set for both data sets, D_{Cora}^{Paper} and D_{Cora}^{All}. The results of D_{Cora}^{Rest} refer always to the result of the BASE component since the context of the references is missing and the local adaption cannot be applied. The AIE approach achieves a remarkable result on the (D_{Cora}^{Paper}) data set with an average F_1 score of 99.7%. 98.7% of the references are extracted correctly without a single error. This is a reduction of the harmonic mean error rate by 88.2% for the complete BIBTEX instances. Errors in the extraction process can be observed for complicated references as well as for simpler ones. The presented approach is rather resilient to the difficulty of such references because the approach extracts difficult references correctly by learning from other references of the same document. There was no adaption applied for the *Editor* field. The development data set did not encourage any adaption of the *Editor*, because the BASE component already achieved a F_1 score of 100.0% for this part on the development set D_{Dev}, resulting in a limited amount and quality of the necessary rules.

Although the adaption phase of our approach was only applied to 61% of the references of the data set D_{Cora}^{All}, its evaluation results are able to compete with the results of Peng [3] and ParsCit [4], the state of the art approaches to the best knowledge of the authors. Our results are difficult to compare the results from other researchers. First of all,

the achieved outcomes are already very good and admit only marginally improvements for this data set. The results of all three approaches were accomplished with different training or development data sets: A defined amount of references [3], a 10 fold cross evaluation [4] and an external data set for the presented approach. The set of features applied, e.g., the content of the additional word lists, or even the tokenizer used to count the true positives may vary between different approaches. Besides that, it is difficult to compare a knowledge engineering approach with machine learning methods, because the knowledge engineer contributes an intangible amount of background knowledge to the rule set. A different approach to evaluate the extraction result might introduce a better comparability of the applied methods and their benefit in real applications. Matching the extraction results of an available large data set, e.g., the ACL Anthology Reference Corpus [9], against a defined database of references could provide this information.

4 Related Work

The extraction of references is an active area of research mainly dominated by machine learning methods. Techniques based on Hidden Markov Models, Maximum Entropy Models, Support Vector Machines, and several approaches using Conditional Random Fields (CRF) were published. Peng and McCallum [3] underlined the applicability of CRF with their results on the CORA data set and established CRF as the state of the art approach for the reference extraction task. 350 references of the CORA data sets were used for training and 150 references for the evaluation of the approach. Councill et al. [4] also applied CRF on the CORA data set and used two other data sets for their evaluation. The results of both approaches are depicted in table 1.

There are some knowledge engineering approaches. One is Cortez et al. [10] which proposed an unsupervised lexicon-based approach. After a chunking phase each text segment is matched against an automatically generated, domain-specific knowledge base in order to identify the fields. They evaluated their approach on data sets of two different domains from health science and computer science. Day et al. [11] used template matching to extract references from journal articles with a well-known style and achieved an average field-level F_1 score of 90.5% for the extraction of the BIBTEX fields *Author*, *Title* and *Date*.

5 Conclusions

We presented a novel adaptive information extraction approach which we successfully applied on the extraction of references from scientific publications. The information extraction component is extended by a short term memory and utilizes the homogenous information of e.g. one document on a meta-level. The approach is able to increase the accuracy of the initially extracted information and performs considerably well compared to state of the art methods. Applied on consistent references sections, an average F_1 score of 99.7% is achieved. Additionally, the results can easily be improved by extending the approach to other BIBTEX fields, because of its disjoint partition of the information: If, for example, a publisher or journal was identified, then other possible

labels like the title can easily be excluded. Furthermore, the nature of a knowledge engineering approach includes also the possibility to improve the extraction component by investing more time in extending and refining the rule set.

There is ongoing work to either combine or replace the presented knowledge engineering approach with machine learning methods. The straightforward combination with statistical methods is given by extending their feature extraction with the results of the adaption phase or by applying the described approach afterwards. The adaption phase is independent of the initial BASE component and can be combined with arbitrary information extraction components. Staying with the rule-based information extraction, well-known coverage algorithms like LP^2 [12] can be adapted to automatically acquire the knowledge necessary for this approach.

References

1. McCallum, A., Nigam, K., Rennie, J., Seymore, K.: Automating the Construction of Internet Portals with Machine Learning. Information Retrieval Journal **3** (2000) 127–163
2. Voss, J., Hotho, A., Jaeschke, R.: Mapping Bibliographic Records with Bibliographic Hash Keys. In Kuhlen, R., ed.: Information: Droge, Ware oder Commons? Proceedings of the ISI, Hochschulverband Informationswissenschaft, Verlag Werner Huelsbusch (2009)
3. Peng, F., McCallum, A.: Accurate Information Extraction from Research Papers using Conditional Random Fields. In: HLT-NAACL. (2004) 329–336
4. Isaac Councill, C.L.G., Kan, M.Y.: ParsCit: an Open-source CRF Reference String Parsing Package. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, ELRA (2008)
5. Kluegl, P., Atzmueller, M., Puppe, F.: Meta-Level Information Extraction. In: The 32nd Annual German Conference on Artificial Intelligence, Berlin, Springer (September 2009)
6. Brill, E.: Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging. Computational Linguistics **21**(4) (1995) 543–565
7. Nahm, U.Y.: Transformation-Based Information Extraction Using Learned Meta-rules. In: CICLing. (2005) 535–538
8. Kluegl, P., Atzmueller, M., Puppe, F.: TextMarker: A Tool for Rule-Based Information Extraction. In: Proceedings of the Biennial GSCL Conference 2009, 2nd UIMA@GSCL Workshop, Gunter Narr Verlag (2009) 233–240
9. Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.Y., Lee, D., Powley, B., Radev, D., Tan, Y.F.: The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, ELRA (2008)
10. Cortez, E., da Silva, A.S., Gonçalves, M.A., Mesquita, F., de Moura, E.S.: FLUXCiM: Flexible Unsupervised Extraction of Citation Metadata. In: JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, New York, USA (2007) 215–224
11. Day, M.Y., Tsai, R.T.H., Sung, C.L., Hsieh, C.C., Lee, C.W., Wu, S.H., Wu, K.P., Ong, C.S., Hsu, W.L.: Reference Metadata Extraction using a Hierarchical Knowledge Representation Framework. Decis. Support Syst. **43**(1) (2007) 152–167
12. Ciravegna, F.: $(LP)^2$, Rule Induction for Information Extraction Using Linguistic Constraints. Technical Report CS-03-07, University of Sheffield, Sheffield (2003)