

# Segmentation of References with Skip-Chain Conditional Random Fields for Consistent Label Transitions

Martin Toepfer and Peter Kluegl and Andreas Hotho and Frank Puppe

University of Würzburg,

Department of Computer Science VI

Am Hubland, 97074 Würzburg, Germany

{toepfer, pkluegl, hotho, puppe}@informatik.uni-wuerzburg.de

## Abstract

The accurate segmentation of references in scientific publications is a crucial task for many important applications like academic search engines or analysis of citation graphs. One of the most successful and popular techniques for this labeling problem are Conditional Random Fields (CRF) and especially their chain structured variants. However, the structural consistencies introduced by the applied citation style guide in the natural source of references, namely the reference sections of published papers, are till now neglected by these methods. We propose a variant of skip-chain CRFs for collective information extraction and exploit these long range dependencies by adding potentials for consistent label transitions. The experimental results indicate that our approach outperforms the common approaches on the data set by achieving an average error reduction of 46%.

## 1 Introduction

The automatic extraction of references from research papers provides access to structured data that allows for several useful applications. For instance, academic search engines aid researchers in their daily work and the analysis of citation graphs yields valuable insights about research communities, topics and trends. However, this pipeline processing often induces cascading errors and consequently the accuracy of the base extraction components is crucial for the overall quality of the systems. This is a widely known problem of natural language processing in general. Hence, the improvement of reference extraction systems is still attractive even if state-of-the-art methods achieve average accuracies better than 90%.

One of the most popular techniques for this task are Conditional Random Fields (CRF) and their chain structured variant linear chain CRFs. They model conditional probabilities with undirected graphs and are trained in a supervised fashion to discriminate label sequences. Although CRFs and related methods achieve remarkable results, their accuracy in real world applications is often not sufficient. Many possibilities to improve their accuracy remain open. One aspect of improvement of CRFs in general has been the relaxation of the assumption that the instances are independent and identically distributed. Long-range dependencies of instances or interesting entities have been in the focus of collective information extraction in order to improve the accuracy. One example for named entity recognition are

skip-chain CRFs [Sutton and McCallum, 2004] that add additional factors between related tokens. They achieved significant improvements with the assumption that similar tokens should be labeled with consistent entities.

If semi-structured text contains several instances or entities, it may often lead to structural consistencies between its instances. This is especially true for the domain we are addressing in this work: the segmentation of references in research papers. The bibliographic section of a scientific publication applies a single style guide and its instances, namely the references, share a very similar structure and composition. While these references are locally homogeneously composed in one context, the data set is globally still heterogeneous and the structure of information is possibly contradictory.

We propose a variant of skip-chain CRFs for exploiting these structural consistencies. The common skip-chain approach introduces long range dependencies based on the observation, that is, between labels whose associated tokens are identical or share a predefined similarity. Instead, our work focuses on patterns and properties of the label sequence. Additional dependencies are added to the assigned label transitions within one reference section to represent the assumed homogeneity. We evaluate the proposed approach with reassigned instances of freely available data sets for the segmentation of references. In a five fold cross evaluation our methods outperforms the common approaches and achieves an average error reduction of 46%.

The paper is structured as follows: Section 2 introduces CRFs and the fundamentals of linear chain CRFs and the initially formulated skip-chain CRFs which are applied in the evaluation. Subsequently, we describe our variant of skip-chain CRFs. The evaluation setting and experimental results are presented and discussed in Section 3. Section 4 gives a short overview of the related work. Finally, Section 5 concludes with a summary of the presented work.

## 2 Method

We interpret reference extraction as a structured prediction problem between sequences of random variables  $\mathbf{x} = (x_t)_{t=1, \dots, n}$ , denoting observed input tokens, and corresponding output random variables  $\mathbf{y} = (y_t)_{t=1, \dots, n}$  for the labellings of  $\mathbf{x}$ . It is important for our approach that  $\mathbf{x}$  represents multiple instances that were drawn in the same context and exhibit structural consistencies, e.g., a sequence  $\mathbf{x}$  contains the tokens of all references of a reference section of a research paper,  $\mathbf{y}$  contains the corresponding BibTex types and  $\mathbf{x}$  was created from  $\mathbf{y}$  following a style guide.

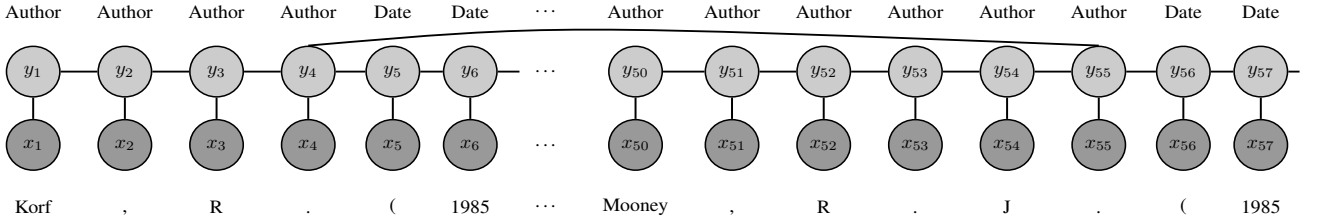


Figure 1: Excerpt of a skip-chain CRF for consistent label transitions. The output variables  $y_4$  and  $y_{55}$  both satisfy  $\kappa_{AE}$ , i.e., they both have Author values, followed by variables with non-Author values. Moreover, we assume that they are part of the same context but not part of the same reference, i.e.,  $r(y_4) \neq r(y_{55})$ . Further skip edges emanating from  $y_4$  and  $y_{55}$  are not shown for simplicity.  $y_{53}$  is not linked to  $y_4, y_{55}$  although all three have the same word identity.

## 2.1 Conditional Random Fields

Conditional Random Fields (CRFs) [Lafferty *et al.*, 2001] are undirected graphical models which model conditional distributions. Given exponential potential functions  $\Phi(\mathbf{y}_c, \mathbf{x}_c) = \exp(\sum_k \lambda_k f_k(\mathbf{y}_c, \mathbf{x}_c))$  a CRF assigns

$$P_\lambda(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\mathbf{x}} \prod_{c \in \mathcal{C}} \Phi(\mathbf{y}_c, \mathbf{x}_c) \quad (1)$$

to a graph with cliques  $\mathcal{C}$  under model parameters  $\hat{\lambda} = (\lambda_1, \dots, \lambda_K) \in \mathbb{R}^K$ . The partition function  $Z_\mathbf{x} = \sum_{\mathbf{y}'} \prod_{c \in \mathcal{C}} \Phi(\mathbf{y}_c, \mathbf{x}_c)$  is a normalization factor to assert  $\sum_{\mathbf{y}} P_\lambda(\mathbf{y}|\mathbf{x}) = 1$ . The feature functions  $f_k$  can be real valued in general, however, we assume binary feature functions.

For the following it is essential that CRFs allow to incorporate properties of arbitrary parts of both the input sequence and the output sequence into the model through the choice of the cliques  $c \in \mathcal{C}$  of the graph and the feature functions  $f_k$ .

## 2.2 Linear Chain CRFs

Linear chain CRFs restrict the underlying graph structures to be linear sequences, typically with a first order Markov assumption. The assignment of  $y_t$  given  $\mathbf{x}$  and  $\mathbf{y} - y_t = (y_t)_{t=1, \dots, t-1, t+1, \dots, n}$  is then only dependent on  $y_{t-1}, y_{t+1}$  and  $\mathbf{x}$ . The arguments of the feature functions in the conditional probability assignment

$$P_\lambda(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\mathbf{x}} \exp\left(\sum_{t=1}^T \sum_{k=1}^K \lambda_k \cdot f_k(y_{t-1}, y_t, \mathbf{x}, t)\right)$$

of linear chain CRFs show the characteristic independence assumptions.

## 2.3 Skip-Chain CRFs

Skip-chain CRFs [Sutton and McCallum, 2004] break the first order Markov assumption of standard linear chain CRFs by adding potentials to the graph that address dependencies between distant labels and tokens. A set  $I = \{(u, v)\} \subset \{1, \dots, n\} \times \{1, \dots, n\}$  defines positions  $u, v$  for which  $y_u, y_v$  are connected by skip edges. To reduce the computational cost,  $I$  has to be kept small. Hence,  $I = I_\mathbf{x}$  is determined on certain conditions of the input sequence  $\mathbf{x}$  to create a sparse graph structure, e.g., skip-chain CRFs for named entity recognition (NER) unroll skip edges based on the word identity of the tokens in  $\mathbf{x}$  [Sutton and McCallum, 2004]. The new potentials correspond to feature functions  $g_s, s = 1, \dots, S$  that have the form

$$g_s(y_u, y_v, \mathbf{x}, u, v) = q_{1,s}(y_u, y_v, u, v) \cdot q_{2,s}(\mathbf{x}, u, v).$$

The first factor  $q_{1,s}$  is introduced to learn different parameters for the various label assignments  $y_u, y_v$  can take. The second factor  $q_{2,s}$  allows  $g_s$  to share observed information between the positions  $u$  and  $v$  and their neighborhoods, for instance, by indicating if a certain local property occurred either in the neighborhood of  $x_u$  or in the neighborhood of  $x_v$ .

With the new potentials formula (1) for a skip-chain CRF becomes

$$P_\lambda(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\mathbf{x}} \exp\left(\sum_{t=1}^T \sum_{k=1}^K \lambda_k \cdot f_k(y_{t-1}, y_t, \mathbf{x}, t) + \sum_{s=1}^S \sum_{(u,v) \in I} \lambda'_s \cdot g_s(y_u, y_v, \mathbf{x}, u, v)\right). \quad (2)$$

## 2.4 Skip-Chain CRFs for Consistent Label Transitions

Our work builds upon the skip-chain approach of [Sutton and McCallum, 2004] and we investigate its applicability to exploit structural consistencies. Thus, we do not assign the skip edges only based on the observations in  $\mathbf{x}$  and instead we bind the skip edge assignments on properties of the label sequence  $\mathbf{y}$ . Afterwards, we check if the tokens that are connected by the skip edges satisfy certain conditions and model these properties with feature assignments. Put in different words, our approach applies the skip edge assignment of NER skip-chain CRFs upside down,  $I = I_\mathbf{y}$ . Instead of connecting the labels of all identical words, we link labels that have identical or related properties, respectively, labels that occur in similar label patterns. Hence, we also modify the form of the functions  $g_s$  slightly to incorporate wider ranges in  $\mathbf{y}$ . In this paper, we assume  $g_s(y_u, y_{u+1}, y_v, y_{v+1}, x_u, x_v)$  but  $g_s$  could address an arbitrary pattern in  $\mathbf{y}$  with properties of  $\mathbf{x}$ . We focus on structural consistency conditions of contexts, in particular word identity consistencies in label transitions which arise e.g. from the use of style guides in the creation of reference sections.

Let us given an example of a skip feature that models the dependencies between ends of author fields. For convenience, we first define an indicator function for ends of author fields

$$\kappa_{AE}(y_t, y_{t+1}) = \begin{cases} 1 & \text{iff } y_t = \text{Author} \wedge y_{t+1} \neq \text{Author}, \\ 0 & \text{else,} \end{cases}$$

and a function  $r(\cdot)$  that maps a given token to the id of its reference in the given context, e.g., if the token  $x_{16} =$

‘NIPS’ occurred in the 5th reference of a given context, then  $r(x_{16}) = 5$ .

Now we choose  $I_y$  to create skip edges for all author end label transitions by

$$I_y = \{(u, v) : \kappa_{AE}(y_u, y_{u+1}) \cdot \kappa_{AE}(y_v, y_{v+1}) = 1 \wedge u < v\}$$

and introduce a feature function

$$g_{AE}(y_u, y_{u+1}, y_v, y_{v+1}, x_u, x_v) = \begin{cases} 1 & \text{iff } \kappa_{AE}(y_u, y_{u+1}) \cdot \kappa_{AE}(y_v, y_{v+1}) = 1 \\ & \text{and } x_u = x_v \\ & \text{and } r(x_u) \neq r(x_v), \\ 0 & \text{else,} \end{cases} \quad (3)$$

to support consistent assignments, i.e., assignments which agree in the form of author field separation. The skip edges propagate evidence information of different parts of the context which can help to solve ambiguities. Figure 1 depicts an example of a skip edge assignment. In this sequence,  $f_{AE}(y_4, y_5, y_{55}, y_{56}, x_4, x_{55}) = 1$ . Note that  $y_4$  and  $y_{55}$  are connected as a result of their similar occurrence in author end patterns and that  $y_{53}$  is not skip edge connected although  $x_{53}$  has the same word identity as the tokens of  $y_4$  and  $y_{55}$ .

Generally,  $I_y$  and the functions  $g_s$  can take arbitrary meaningful patterns in  $y$  into account and  $g_s$  is not restricted to be binary, e.g., it is possible to apply similarity metrics and return a real as a measure of compatibility.

## 2.5 Inference and Parameter Estimation

Our work uses the same algorithms as [Sutton and McCallum, 2004] for inference, i.e., compute  $P_\lambda(\mathbf{y}|\mathbf{x})$  and decide which labels are most likely for the observed input, and to estimate the parameters  $\hat{\lambda}$  of the model. We apply tree based reparameterization (TRP) [Wainwright *et al.*, 2001] for inference. TRP is related to belief propagation and computes approximate marginals for loopy graphs. To obtain the parameters  $\hat{\lambda} = \{\lambda_k, \} \cup \{\lambda'_s\}$  of the model our approach uses training data  $D$  and maximum a posteriori estimation. The log likelihood  $\mathcal{L}(\hat{\lambda}|D)$  of the model parameters given the training examples is optimized with the quasi-Newton method L-BFGS and a Gaussian prior on the parameters.

## 3 Experimental Results

The presented approach is evaluated in the domain of reference segmentation, a popular task for the evaluation of information extraction techniques. Common approaches separately process the instances, namely the references. Within these references, the interesting entities, like the author, title or date, need to be identified. Since all tokens of a reference are part of exactly one entity, one speaks of a segmentation task. The presented work relies on additional factors between labels of different references. Therefore the complete reference section is addressed as the processed instance and the border of each reference need to be identified. In this section, we introduce the data sets that are used for the evaluation and the overall setting of the experiments. Afterwards, the results are presented and discussed.

### 3.1 Settings

Besides their differences with respect to skip edges, all evaluated models rely on the same set of feature functions

commonly used in the domain of reference segmentation. To these features belong the text and class of the token, some limited dictionaries, character n-grams and the text of neighboring tokens within a window of three tokens. The dictionaries cover first names, locations, keywords (e.g., “eds.”) and some well known journals and publishers. Overall, the applied features are comparable to previously published approaches, e.g., by Peng and McCallum [Peng and McCallum, 2004].

Available data sets for the segmentation of references are not applicable for an evaluation of the presented approach. They provide only a listing of labeled instances missing their creation context and the affiliation to a certain reference section respectively. Therefore, a new data set was created that consists only of references of the freely available data sets CORA, CITESEERX and FLUX-C1M<sup>1</sup>. A simple script was applied to assign references with similar style guides to documents up to the size of 20 instances. The resulting data set contains 28 documents and overall 452 references. Although this data set was automatically generated, it strongly resembles a set of natural reference sections, especially since not all entities in a document are consistently structured. The created data set is annotated with the label set of Peng and McCallum [Peng and McCallum, 2004] and can be downloaded in a format containing all used features<sup>2</sup>.

We utilized the GRaphical Models in Mallet (GRMM) package [Sutton, 2006] for an implementation of the CRFs. Exclusively default parameters were used and all models were trained with 50 iterations given the algorithms mentioned in section 2.5. The model of your approach applied only templates for the end of the *author*, *booktitle*, *date*, *pages* and *title*.

### 3.2 Results

The presented approach is evaluated in a five fold cross evaluation against a base line model, namely a linear chain CRF, and a previously published and comparable approach, a skip-chain CRF. The documents of the data set were randomly distributed over the folds. For comparability, we measure the token accuracy:

$$\text{accuracy} = \frac{\text{\#correctly labeled tokens}}{\text{\#all tokens}}.$$

The results of each fold and the average accuracy are depicted in Table 1. Our approach outperforms the base line CRF in each fold and is able to achieve an average error reduction of 46%. For a comparable model, the original skip-chain approach of [Sutton and McCallum, 2004] that introduces factors between identical capitalized words was applied on the data set. Its accuracy was not able to match the results of the base line CRF probably because of the limited amount of iterations. However, the training of our model was equally parameterized. Additionally to the original approach, we also evaluated different skip-chain CRF models including constraints for identical punctuation marks and all tokens in general, but none of those approaches achieved reasonable or noteworthy results.

## 4 Related Work

In this section we give a brief overview of related work which basically can be divided into two groups. On the

<sup>1</sup><http://wing.comp.nus.edu.sg/parsCit/>

<sup>2</sup><http://ki.informatik.uni-wuerzburg.de/~pkluegl/LWA2011>

Table 1: Results for the segmentation of references

|         | Linear-Chain | Skip-Chain | Our Approach  |
|---------|--------------|------------|---------------|
| Fold 1  | 96.83%       | 95.59%     | <b>97.96%</b> |
| Fold 2  | 97.02%       | 96.99%     | <b>98.04%</b> |
| Fold 3  | 98.09%       | 97.93%     | <b>98.94%</b> |
| Fold 4  | 94.86%       | 95.28%     | <b>97.96%</b> |
| Fold 5  | 98.32%       | 97.93%     | <b>99.10%</b> |
| Average | 97.02%       | 96.75%     | <b>98.40%</b> |

one hand, there is work on reference extraction which has not incorporated long distance dependencies. On the other hand, there are named entity recognition (NER) approaches incorporating context consistencies and non-linear structure.

Especially for NER modelling long-distance dependencies is crucial. The labeling of an entity is quite consistent within a given document, however, conclusive discriminating features are sparsely spread across the document. As a consequence, leveraging predictions of one instance to disambiguate others is essential. Besides the already mentioned skip-chain CRF approach of [Sutton and McCallum, 2004], [Bunescu and Mooney, 2004] also model dependencies between distant entities, however, they employ Relational Markov Networks. Their approach also claims that tokens with the same text are likely to have the same labels and create special repeat templates to support consistency. In contrast to these approaches, our study has investigated the applicability of models that assume similarity in the observed sequence under given similar patterns in the labels, namely label transitions.

Fundamental work on information extraction with CRFs can be found at [Peng and McCallum, 2004] who evaluate several design parameters of linear chain CRFs such as Markov order, features and different priors for regularization with application to the reference extraction domain and extracting meta-data from paper headers. A semi-supervised approach to reference extraction using database records has been contributed by [Bellare and McCallum, 2009]. Their approach uses a CRF for alignment of text sequences to database records and another CRF for information extraction. Both CRFs are bound to handcrafted expectation criteria, e.g. the extraction CRF has a criterion that states that a token with the word identity 'EMNLP' has always a booktitle label. Alternating Projections [Bellare *et al.*, 2009] is another possibility to incorporate domain knowledge through expectations and constraints that allows for semi-supervised reference extraction. For instance, one of these constraints assures that citations can only start with author or editor fields. Our work utilizes supervised learning and thus the results are not directly comparable. However, it should be possible to combine our approach with Alternating Projections and expectation criteria to fit it to semi-supervised learning as well.

## 5 Conclusions

We have presented a novel approach for segmenting references using a variation of a skip-chain CRF. The structural consistencies introduced by the applied style guide can be exploited with additional long range dependencies between the label transitions. Although our methods currently only models the end transition of five entities and only relies on token identity, it is able to outperform a normal CRF. The

results of the five fold cross evaluation on a generated data set indicate an average error reduction of 46%.

For future work, we plan experiments measuring the runtime and an extension of our approach with begin and end constraints for all interesting entities of bibliographic references. For this purpose the comparison of the token identity must be exchanged by a more sophisticated method that relies on the complete feature vector. An additional evaluation using a data set with original reference sections has been initiated and indicates similar results. Furthermore, experiments in other domains with structural consistencies in a given creation context will show the generalizability of the presented approach.

## References

- [Bellare and McCallum, 2009] Kedar Bellare and Andrew McCallum. Generalized expectation criteria for bootstrapping extractors using record-text alignment. In *EMNLP*, pages 131–140. ACL, 2009.
- [Bellare *et al.*, 2009] Kedar Bellare, Gregory Druck, and Andrew McCallum. Alternating projections for learning with expectation constraints. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 43–50, Arlington, Virginia, United States, 2009. AUAI Press.
- [Bunescu and Mooney, 2004] Razvan Bunescu and Raymond J. Mooney. Collective Information Extraction with Relational Markov Networks. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
- [Lafferty *et al.*, 2001] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proc. 18th International Conf. on Machine Learning*, pages 282–289, 2001.
- [Peng and McCallum, 2004] Fuchun Peng and Andrew McCallum. Accurate Information Extraction from Research Papers using Conditional Random Fields. In *HLT-NAACL*, pages 329–336, 2004.
- [Sutton and McCallum, 2004] Charles Sutton and Andrew McCallum. Collective Segmentation and Labeling of Distant Entities in Information Extraction. In *ICML Workshop on Statistical Relational Learning and Its Connections to Other Fields*, 2004.
- [Sutton, 2006] Charles Sutton. GRMM: GRaphical Models in Mallet, 2006. <http://mallet.cs.umass.edu/grmm/>.
- [Wainwright *et al.*, 2001] Martin J. Wainwright, Tommi Jaakkola, and Alan S. Willsky. Tree-based Reparameterization for Approximate Inference on Loopy Graphs. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *NIPS*, pages 1001–1008. MIT Press, 2001.