Collective Information Extraction with Context-specific Consistencies

Peter Kluegl[1,2], Martin Toepfer[1], Florian Lemmerich[1], Andreas Hotho[1], and Frank Puppe[1]

[1] Department of Computer Science VI, University of Würzburg, Am Hubland, Würzburg, Germany
[2] Comprehensive Heart Failure Center, University of Würzburg, Straubmühlweg 2a, Würzburg, Germany
{pkluegl,toepfer,lemmerich,hotho,puppe}@informatik.uni-wuerzburg.de

Abstract. Conditional Random Fields (CRFs) have been widely used for information extraction from free texts as well as from semi-structured documents. Interesting entities in semi-structured domains are often consistently structured within a certain context or document. However, their actual compositions vary and are possibly inconsistent among different contexts. We present two collective information extraction approaches based on CRFs for exploiting these context-specific consistencies. The first approach extends linear-chain CRFs by additional factors specified by a classifier, which learns such consistencies during inference. In a second extended approach, we propose a variant of skip-chain CRFs, which enables the model to transfer long-range evidence about the consistency of the entities. The practical relevance of the presented work for real-world information extraction systems is highlighted in an empirical study. Both approaches achieve a considerable error reduction.

Keywords: information extraction, conditional random fields, collective, context-specific consistencies, long-range dependencies

1 Introduction

The accurate transformation of unstructured data into a structured representation for further processing is an active area of research with many interesting challenges. One central task for mining unstructured textual data is Information Extraction (IE), which tries to find well-defined entities and relations in textual data. Over the last decade, statistical sequence labeling models and especially Conditional Random Fields (CRFs) [10] became the dominant technique for IE tasks. CRFs are discriminative undirected probabilistic graphical models often trained in a supervised fashion. When applied on textual data, they are usually designed as a linear chain with the first order Markov assumption.

In many scenarios, the entities in textual data are not independent and identically distributed. Recently, much effort went in new approaches that can be

summarized under the term Collective IE [2, 4, 8, 9, 15]. They break the linear-chain assumption and model also long-range dependencies in order to label related entities or instances collectively. One example is Named Entity Recognition (NER), a task that aims at the extraction of persons or similar entities. Here, the accuracy can be improved by the assumption that similar tokens should have the same label or by providing contextual evidence of related tokens.

In semi-structured documents a different form of long-range dependency often occurs. Here, the context in which the textual data is created or written introduces a homogeneous composition of the entities. The reference section of this paper, for example, is generated using a style guide that defines the layout of the citation information. Thus, all author entities end with a colon. However, the reference sections of other publications follow different style guides in which the author possibly ends with a period. Another example for consistency introduced in a certain context is curricula vitae: Each author describes his or her employments homogeneously but possibly with an arrangement of the interesting entities different from other authors. If these long-range dependencies are not taken into account, then the IE system faces a heterogeneous and inconsistent composition of the entities in the complete dataset. However, by considering the similarities of entities within a context and processing those entities collectively, many labeling errors can be prevented. The accuracy for the detection of the author of a reference, for example, can be greatly increased when the model is encouraged that all authors in a reference section should end identically.

In this work, we present two collective IE approaches based on CRFs that are able to exploit such context-specific consistencies. Both approaches consult a classifier, which detects consistent boundaries of an entity within one context. This classifier is trained during inference on an intermediate label sequence predicted by an additional model. The generalization of the classifier's learning algorithm detects only equally shaped boundaries ignoring entities that break the consistency assumption. This evidence about the consistency is exploited in two different models. The first model extends linear-chain CRFs with additional unigram factors. The positions of the factors are given by the classification result of the classifier combined with the predicted label sequence. In a second approach, we investigate a variant of skip-chain CRFs [15]. Instead of adding dependencies for similar tokens, the boundaries of related entities are connected. These additional edges then transport evidence about the consistency of the entities' compositions at the positions indicated by the classifier. In an empirical study, we evaluate our approaches with real-word datasets, for the segmentation of references and for template extraction in curricula vitae. The results show the practical relevance of the presented work for real-world IE systems. Our approaches are able to achieve a substantial error reduction, up to 34%.

The rest of the paper is structured as follows: In Section 2, we recap different variants of CRFs for information extraction. The two novel approaches for exploiting context-specific consistencies are described in Section 3. Their results in an empirical study are presented in Section 4. Section 5 gives a short overview of the related work and Section 6 concludes with a summary.

2 Conditional Random Fields

Conditional Random Fields (CRFs) [10] are undirected graphical models which model conditional distributions over random variables $\mathbf{y}$ and $\mathbf{x}$. 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

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{c \in \mathcal{C}} \Phi(\mathbf{y_c}, \mathbf{x_c}) \tag{1}$$

to a graph with cliques $\mathcal{C}$ under model parameters $\theta = (\lambda_1, \ldots, \lambda_K) \in \mathbb{R}^K$. The partition function $Z(\mathbf{x}) = \sum_{y'} \prod_{c \in \mathcal{C}} \Phi(\mathbf{y_c}, \mathbf{x_c})$ is a normalization factor to assert $\sum_y p_\theta(\mathbf{y}|\mathbf{x}) = 1$. The feature functions f_k can be real valued in general, however, we assume binary feature functions if not mentioned differently.

When CRFs are applied for IE tasks, the model is adapted to the properties of sequential data or textual documents respectively. Therefore the graph structure is normally restricted to be a linear chain representing the sequence of labels that are assigned to a sequence of tokens. The entities of the IE tasks are identified by sequences of equal labels. If linear-chain CRFs also model long-range dependencies with additional edges between distant labels, then the models are called skip-chain CRFs [15]. Both models are shortly outlined in the following.

2.1 Linear-Chain CRFs

Linear chain CRFs [10] 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 $\mathbf{y} - y_t = (y_t)_{t=1,\ldots,t-1,t+1,\ldots,T}$ is then only dependent on y_{t-1},y_t, y_{t+1} and $\mathbf{x}$. The probability of a label sequence $\mathbf{y}$ given an token sequence $\mathbf{x}$ is modeled by

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \Phi_{\text{L}}(y_t, y_{t-1}, \mathbf{x}). \tag{2}$$

We are using Φ_{L} to describe the factors of the linear-chain edges that link adjacent labels:

$$\Phi_{\text{L}}(y_t, y_{t-1}, \mathbf{x}) = \exp\left\{ \sum_k \lambda_{\text{L}k} f_{\text{L}k}(y_t, y_{t-1}, \mathbf{x}, t) \right\}. \tag{3}$$

The discriminative impact of the feature functions $f_{\text{L}k}$ is weighted by the parameters $\theta = \theta_{\text{L}} = \{\lambda_{\text{L}k}\}_{k=1}^K$. The feature functions can typically be further factorized into indicator functions $p_{\text{L}k}$ and observation functions $q_{\text{L}k}$

$$f_{\text{L}k}(y_t, y_{t-1}, \mathbf{x}, t) = p_{\text{L}k}(y_t, y_{t-1}) \cdot q_{\text{L}k}(\mathbf{x}, t). \tag{4}$$

$p_{\text{L}k}$ returns 1 for a certain label configuration and $q_{\text{L}k}$ relies only on the input sequence $\mathbf{x}$. Thus, a feature function, e.g., that indicates capitalized tokens, can be separately weighted for each label transition. Figure 1 contains an example

of a linear-chain CRF in factor graph representation, which is applied for the reference segmentation task. We added the label and token sequence for better understanding. Dependencies of the factors to tokens are omitted for simplicity.

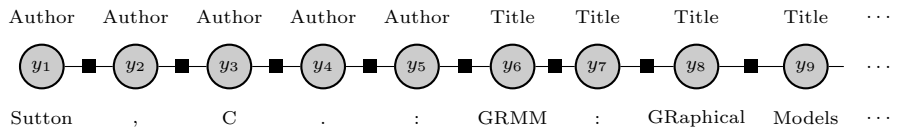

Fig. 1. A linear-chain CRF applied on the reference segmentation task, i.e., the 14th reference of this paper. The associated labels and tokens are depicted above and below the variables.

2.2 Skip-Chain CRFs

Skip-chain CRFs [15] break the first order Markov assumption of linear-chain CRFs by adding potentials to the graph that address dependencies between distant labels and tokens. A set $I_{\text{X}} = \{(u, v)\} \subset \{1, \ldots, T\} \times \{1, \ldots, T\}$ defines positions u, v for which y_u, y_v are connected by skip edges. We refer to components of skip-chain CRFs with the index X in order to point out their usage in previous publications, e.g., [15]. The set I_{X} unrolls skip edges based on token similarity and is therefore only dependent on the token sequence $\mathbf{x}$. In NER tasks, for example, the accuracy can often be increased when the model is encouraged to label similar tokens identically. For controlling the computational cost, I_{X} has to be kept small. An extension of Equation 2 with additional skip edges results in the conditional probability

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \Phi_{\text{L}}(y_t, y_{t-1}, \mathbf{x}) \prod_{(u,v)\in I_{\text{X}}} \Psi_{\text{X}}(y_u, y_v, \mathbf{x}). \tag{5}$$

The potentials Ψ_{X} for the skip edges are given by

$$\Psi_{\text{X}}(y_u, y_v, \mathbf{x}) = \exp\left\{\sum_k \lambda_{\text{X}k} f_{\text{X}k}(y_u, y_v, \mathbf{x}, u, v)\right\} \tag{6}$$

extending the complete set of parameters $\theta = \theta_{\text{L}} \cup \theta_{\text{X}}$. The feature functions factorize again in an indicator function $p_{\text{X}k}$ and an observation function $q_{\text{X}k}$:

$$f_{\text{X}k}(y_u, y_v, \mathbf{x}, u, v) = p_{\text{X}k}(y_u, y_v, u, v) \cdot q_{\text{X}k}(\mathbf{x}, u, v) \tag{7}$$

The observation function enables the model to share observed information between the positions u and v and their neighborhoods, e.g., for providing local evidence at a position where such information is missing.

3 CRFs with Context-specific Consistencies

This section introduces two different approaches to exploit context consistencies. Both methods are divided into two different parts. When unrolling the graph during inference, we first have to detect the patterns that describe the consistency of the context. Secondly, we need to incorporate the gained knowledge into the graph structure for a better prediction. In the following, we first describe challenges of dependencies on the label sequence and the applied method to learn context-specific consistencies. Then, we explain the differences of the two approaches which only concern the structure and complexity of the models that exploit the context-specific patterns.

3.1 Context-specific Consistencies

Context-specific consistencies refer to a special kind of long-range dependencies that are often found in semi-structured documents. The interesting entities within a specific context or document share a similar composition caused by the process the document is created or written in. Examples for this process are authors that arrange the entities homogeneously or templates that enforce a specific layout. We call these consistencies context-specific, because the actual composition is unknown at application time and can strongly vary between contexts. There are many different ways to describe the composition of entities. In this work we take a closer look at the entity boundaries, that is, the first and the last label of the entity[3]. Other possibilities include the labels within the boundaries of an entity. More generally, this can be extended to any kind of label transition. However, the boundaries alone are very suitable to classify an entity independently of the actual label transition and allow to restrict the long-range dependencies to a minimal amount.

3.2 Dependencies based on the label sequence

In this paper, we investigate how these consistencies can be exploited with the idea of skip-chain CRFs or in general CRFs with additional potentials for long-range dependencies. In contrast to skip-chain CRFs, where the potentials are only based on the token sequence (cf. Equation 6), our additional potentials are mainly dependent on the label sequence. Our approaches need a prediction of the assignment in order to be able to link or relate the boundaries of the entities. The label sequence (hidden variables) is of course not available during inference when we unroll the graph on an instance with all potentials since it is the result of the computation of $p_\theta(\mathbf{y}|\mathbf{x})$. However, there are many different ways to provide a prediction of the label sequence during inference. Our initial choice was to incrementally unroll the graph: We first unrolled the potentials of the linear-chain part, computed the currently most likely label sequence and used

[3] No additional encoding like IOB is applied in order to identify the entities in the label sequence.

this prediction to further unroll the additional potentials. However, we observed problems with the parameter estimation and inference mechanism applied in this work (cf. Section 3.6). While we sometimes achieved remarkable improvements, the approach frequently did not converge at all. Therefore, we utilize a separate static linear-chain model in order to provide a constant prediction of the label sequences, which corresponds to the approach of stacked graphical models [8, 9, 7]. Here, an initial model is used to compute new features for a stacked model. In our approach, however, the predicted assignments of the initial model lead to additional potentials. Normally, cross-fold training is applied for the initial model in order to prevent unrealistic predictions during training of the stacked model. We neglect this improvement in the belief that the advantages of the presented models prevail.

3.3 Learning Context-specific Consistencies

When we try to exploit the context-specific consistencies, it is very helpful to acquire a description or model for the consistencies in each context or document. Thereby, one can distinguish consistent and inconsistent boundaries of the entities. As in previous work [7], we train and apply a binary classifier on the boundaries of an entity within one context. The learning task of the classifier for a boundary of one type of entity is defined as following: Each token of the context is a training example and the features of the CRF become binary attributes, possibly with an additional windowing. The intermediate label sequence (cf. 3.2), respectively the predicted boundaries, specifies the learning target of the classifier. The generalization capacity of the classifier's learning algorithm is the key to gain knowledge about the context-specific consistency. We assume that the hypothesis space of the classifier is not sufficient to provide a perfectly accurate model and therefore only describes the dominant consistency.

A suitable classifier for the tasks presented in this work has to provide following properties:

- The classifier should be efficient with respect to its execution time since it is trained and applied on all emerging label sequences during inference.
- The classifier should not tend to overfit since it is trained and applied on possibly erroneous data. These errors should not be reproduced. In general, overfitting can also be restrained by limiting the amount of attributes.
- The classifier should not combine different hypotheses in order to solve the classification problem if only one consistency for the boundary exists in data as it is in our examples.
- The classifier should handle label bias correctly, even if there are only a few true positives and thousands of true negatives.

We decided to utilize a simple but efficient rule learner based on subgroup discovery [6], an exhaustive search for the best conjunctive pattern describing an target attribute, respectively the entity's boundary. This technique fulfills all requirements with minimal efforts of configuration and is fast enough if the set

of attributes is constrained. As an improvement to [7], a new quality function F_1^{exp} selects the best pattern:

$$F_1^{exp} = \frac{2 \cdot tp}{2 \cdot tp + fn + fp} \cdot \left(1 - \left(\frac{|tp + fp - E_y|}{\max(tp + fp, E_y)}\right)^2\right) \tag{8}$$

The left part of this measure describes the traditional F_1-Measure, that is how well the pattern reproduces the predicted boundaries. The right factor is a penalty term for the divergence of the amount of instances classified as boundaries to a given variable E_y, the expected amount of boundaries in a context. E_y can simply be estimated using the token sequence and the feature functions in the data set applied in this work. In the domain of reference segmentation, for example, we expect that each reference contains exactly one author. Although this is not true in general, it provides for a valuable weighting of the hypothesis space and further reduces overfitting.

3.4 Comb-Chain CRFs

In a first approach, we extend the variables of a linear chain model with additional (unigram) factors dependent on the classification result (cf. Figure 2). Hence, we chose the name comb-chain CRFs for this approach because of the layout of the graph.

Let $\mathcal{R}_b(y)$ and $\mathcal{R}_e(y)$ be the set of positions, which are identified by the classifier as the beginning and end of an entity with the label y. We can now define the positions of additional factors:

$$\begin{aligned} \mathcal{U}_b &= \left\{ u : y_{u-1} \neq y_u \vee u \in \bigcup_y \mathcal{R}_b(y) \right\} \\ \mathcal{U}_e &= \left\{ u : y_u \neq y_{u+1} \vee u \in \bigcup_y \mathcal{R}_e(y) \right\} \\ \mathcal{U} &= \mathcal{U}_b \cup \mathcal{U}_e \end{aligned} \tag{9}$$

$\mathcal{U}_b$ and $\mathcal{U}_e$ contain all positions that are either intermediately labeled by the external model or are identified by the classifier as the beginning, respectively end of an entity. The conditional probability is then defined as[4]

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \Phi_{\mathrm{L}}(y_t, y_{t-1}, \mathbf{x}) \prod_{u \in \mathcal{U}} \Psi_{\mathrm{C}}(\mathbf{y}, u) \tag{10}$$

and the potentials for the unigram factor are given by

[4] The different usage of $\mathbf{y}$ for the predicted sequence and the label configuration of the parameters deduces from the context.

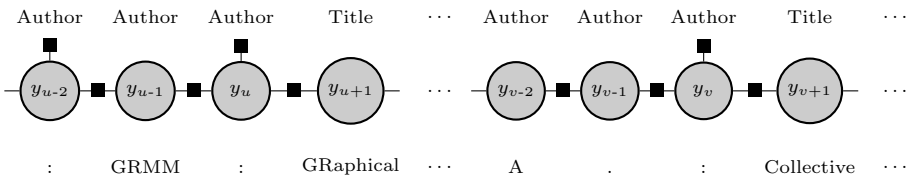

Fig. 2. An excerpt of a comb-chain graph with erroneous labeling whereas only additional factors for the end of the author are displayed. The output functions indicate a missing end at position $y_{u\text{-}2}$, a surplus end at y_u and a consistent end at y_v

$$\Psi_{\mathrm{C}}(\mathbf{y}, u) = \exp\left\{\sum_k \lambda_{\mathrm{C}k} f_{\mathrm{C}k}(\mathbf{y}, u)\right\} \tag{11}$$

whereas $\theta_{\mathrm{C}} = \{\lambda_{\mathrm{C}k}\}$ is the set of additional parameters for the classifier template. We let the feature function factorize into an indicator function $p_{\mathrm{C}k}$ and an output function $q_{\mathrm{C}k}$:

$$f_{\mathrm{C}k}(\mathbf{y}, u) = p_{\mathrm{C}k}(y_u) \cdot q_{\mathrm{C}k}(\mathbf{y}, u) \tag{12}$$

We introduce six different output functions:

$$\begin{aligned}
q_{\text{e-consistent}}(\mathbf{y}, u) &= \begin{cases} 1 & \text{iff } y_u \neq y_{u+1} \wedge u \in \mathcal{R}_e(y_u) \\ 0 & \text{else} \end{cases} \\
q_{\text{e-project}}(\mathbf{y}, u) &= \begin{cases} 1 & \text{iff } y_u \neq \tilde{y} \wedge u \in \mathcal{R}_e(\tilde{y}) \\ 0 & \text{else} \end{cases} \\
q_{\text{e-suppress}}(\mathbf{y}, u) &= \begin{cases} 1 & \text{iff } y_u \neq y_{u+1} \wedge u \notin \mathcal{R}_e(y_u) \\ 0 & \text{else} \end{cases}
\end{aligned} \tag{13}$$

The output functions $q_{\text{b-consistent}}$, $q_{\text{b-project}}$ and $q_{\text{b-suppress}}$ are defined equivalently for the beginning of an entity. This reflects the meaning, that is the result of the classification combined with the intermediate labeling: $q_{\text{e-consistent}}$ indicates a true positive, $q_{\text{e-project}}$ a false positive and $q_{\text{e-suppress}}$ a false negative classification compared to the label sequence. Together these feature functions supply evidence, which parts of the label sequence agree with the consistency and which parts should be altered in order to gain a higher likelihood. The resulting graph of the model contains no loops and provides therefore less challenges for an inference mechanism.

The idea of comb-chain CRFs is summarized with an example for the segmentation of references (cf. Figure 2). Let the reference section of this paper be the input sequence. When unrolling the graph, we ask the external model for

an intermediate labeling specifying the entities. A classifier is trained to detect the boundaries of the entities. The descriptive result of the classifier for the end of the author is, for example, a pattern like "A period followed by a colon". Now, the additional potentials with the output functions influence the model to assign a high likelihood to label sequences that confirm with the description of the classifier.

3.5 Skyp-Chain CRFs

Skyp-chain CRFs are a variant of skip-chain CRFs. But instead of creating additional edges between labels whose tokens are similar or identical, this approach adds long-range dependencies based on the patterns occurring in the predicted label sequence $\mathbf{y}$ and the classification result. Thus, the small modification of the name. When applying skyp-chain CRFs for exploiting context-specific consistencies, two additional differences to published approaches for skip-chain CRFs or similar collective IE models can be identified:

1. There is no need to transfer local evidence to distant labels since we already assume a homogeneous composition of the entities.
2. Useful observation functions for the skip edges cannot be specified, because the relevance of certain properties is unknown.

We first define the set of additional edges that specify the positions of the long-range dependencies using the positions $\mathcal{U}_b$ and $\mathcal{U}_e$ of Equation 9.

$$\begin{aligned} \mathcal{E}_b &= \{(u,v) : u \neq v \wedge y_u = y_v \wedge u \in \mathcal{U}_b \wedge v \in \mathcal{U}_b\} \\ \mathcal{E}_e &= \{(u,v) : u \neq v \wedge y_u = y_v \wedge u \in \mathcal{U}_e \wedge v \in \mathcal{U}_e\} \\ \mathcal{E} &= \mathcal{E}_b \cup \mathcal{E}_e \end{aligned} \tag{14}$$

The set $\mathcal{E}_b$ contains edges that connect the start label of an entity with all other start labels of entities with the same type. The set $\mathcal{E}_e$ refers accordingly to the links between the end labels of entities. Further, we introduce a parameter m_e for controlling the model complexity that restricts the maximal amount of additional long-range dependencies for each variable. E.g., for $m_e = 2$, a label is only connected to the closest previous and following boundary of the same entity type.

Our skyp-chain approach extends the linear-chain model with additional potentials for edges defined in Equation 14. The conditional probability for the assignment of the label sequence is given by

$$p_\theta(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \Phi_{\mathrm{L}}(y_t, y_{t-1}, \mathbf{x}) \prod_{(u,v)\in\mathcal{E}} \Psi_{\mathrm{Y}}(\mathbf{y}, u, v)\,. \tag{15}$$

An example of an unrolled graph of this model is depicted in Figure 3. Similar to Equation 6, the additional potentials factorize to

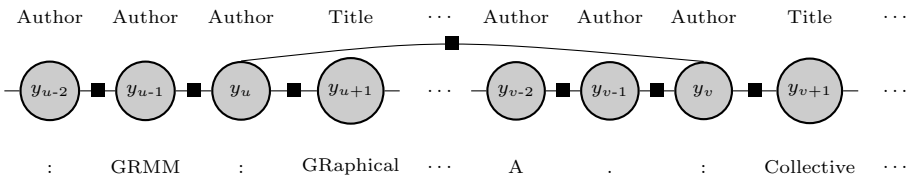

Fig. 3. An excerpt of a skyp-chain graph with erroneous labeling. Only one additional edge for the end of the author is displayed. The likelihood of the sequence is decreased because only position $u-2$ and v but not u were identified as a boundary by the classifier.

$$\Psi_{\mathrm{Y}}\left(\mathbf{y}, u, v\right) = \exp\left\{\sum_{k} \lambda_{\mathrm{Y}k} f_{\mathrm{Y}k}\left(\mathbf{y}, u, v\right)\right\}, \tag{16}$$

resulting in the complete parameter set $\theta = \theta_{\mathrm{L}} \cup \theta_{\mathrm{Y}}$ with $\theta = \theta_{\mathrm{Y}} = \{\lambda_{\mathrm{Y}k}\}$ to be estimated for this model. In contrast to the skip-chain model, our feature functions depend on the complete (predicted) label sequence $\mathbf{y}$. The feature functions consist again of an indicator function for the label configuration, but not of an observation function on the input sequence. Instead we apply the output functions of Equation 13 separately for the source and destination of the skip edge.

Let us illustrate the skyp-chain model in an example for reference segmentation (cf. Figure 3). Let the input sequence be the reference section of this paper. When the graph of the model is unrolled during inference, the most probable label assignments are calculated. During this process we consider long-range dependencies, e.g., for the end of the author entities (cf. labels y_u and y_v in Figure 3). Due to our additional potentials, label sequences with boundaries that are identified by the classifier as consistently structured become more likely. In Figure 3, the likelihood of the sequence is decreased in comparison to a graph with an additional edge between the labels y_{u-2} and y_v.

3.6 Parameter Estimation and Inference

We compute $p_\theta\left(\mathbf{y}|\mathbf{x}\right)$ to decide which label sequence $\mathbf{y}$ is most likely for the observed token sequence $\mathbf{x}$, and to estimate the parameters θ of the model. The applied inference technique, tree based reparameterization (TRP) [17], is related to belief propagation and computes approximate marginals for loopy graphs. TRP is also used in [15] for the original skip-chain models. Unfortunately, severe convergence problems could be oberserved when applied on complex graph structures. The parameters θ of our models are obtained using training data $D = \left\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right\}_{i=1}^{N}$ and maximum a-rrposteriori estimation. The log likelihood $\mathcal{L}(\theta|D)$ of the model parameters given the training examples is optimized with

the quasi-Newton method L-BFGS and a Gaussian prior on the parameters as in [15].

4 Experimental Results

We demonstrate the advantages of the presented approach in a five-fold cross evaluation in two different real-world applications: The segmentation of references and the template extraction in curricula vitae. First, both domains and the real-world datasets are described and then we specify the settings of the evaluation. Finally, we present and discuss the empirical results.

4.1 Datasets

Two datasets are utilized in the evaluation of this work. The dataset *References* origins in a domain that is very popular for the evaluation of novel IE techniques (cf. [1, 11–13]), whereas the dataset *Curricula Vitae* belongs to classical IE problems of template extraction.

References This dataset for the segmentation of references was introduced in previous work [7] and consists only of complete reference sections of real publications, mainly from the computer science domain. The application behind this dataset consists mainly in the identification of Bibtex fields in crawled publications, which can be used to improve scientific search engines or to analyze citation graphs. The dataset contains 566 references in 23 reference sections with overall 15 different labels and is comparable to datasets of previous publications with respect of size, label and feature set, e.g., Peng et al. [11]. For the evaluation in this paper, we reduced the label set for the identification of the entities Author, Date, Title and Venue, which are sufficient for the targeted application. The dataset can be freely downloaded[5]. We skip a detailed description of the features and refer to the archive because it contains all applied features.

Curricula Vitae The IE task of this dataset is to identify the time span and company for which the author of these documents worked in a stage of his or her life (employments). This information can be used to improve the search for suitable future employees for certain projects. The data set consists of 68 curricula vitae and is annotated with 896 companies or sectors[6] and 937 time spans in overall 921 stages of life. We use the label Date for the time span and the label Client for the companies or sectors. The feature set extends the feature set of the dataset *References* with additional domain-specific features like the number of the line, the position within a line and keywords for company prefixes/suffixes and date indicators. Unfortunately, we can not publish this dataset due to non-disclosure agreements.

[5] `http://www.is.informatik.uni-wuerzburg.de/staff/kluegl_peter/research/`

[6] The authors of the curricula vitae sometimes anonymize the actual name of a company and replace it with the sector in which the company is located.

4.2 Evaluation Measure

The performance of the presented models is measured with the F_1 score. Let tp be the number of true positive labeled tokens and fn and fp respectively the number of false negatives and false positives tokens. *Precision*, *recall* and F_1 are then defined as:

$$precision = \frac{tp}{tp + fp}, \;\; recall = \frac{tp}{tp + fn}, \;\; F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}.$$

For the dataset *References* we present the F_1 score combined for all labels, whereas we distinguish the labels DATE and CLIENT for *Curricula Vitae*.

4.3 Settings

All models are trained with identical settings and features. In order to minimize the model complexity of the skyp-chain approach, we set $m_e = 2$. We used 11 (for *References*) and 12 (for *Curricula Vitae*) manually selected feature functions in a window of five tokens as attributes for the rule learner. The learned descriptions had a maximum of three attributes and a minimum quality score of 0.01. For the dataset *References*, only the boundaries for the labels AUTHOR, DATE, TITLE are considered. Our implementation of the CRFs is based on the GRMM package [14].

4.4 Results

We compare the proposed models to a linear-chain CRF (base line). Additionally, we have applied the stacked approach with exact inference of [7] for a comparable model. However, its evaluated F_1 score was surprisingly lower than our base line. An analysis revealed that the different implementations of CRFs and the varying definition of an instance influenced the results. We have also considered different variants of skip-chain CRFs, but none of them returned noteworthy results. As a consequence, we compare our models only with the base line.

The results of the five-fold cross evaluation are depicted in Table 1 for the dataset *References* and in Table 2 for the dataset *Curricula Vitae*. The comb-chain models achieve overall an average error reduction of over 30% and increased the measured averaged F_1 score by at least 1%, 9% for the label CLIENT. The skyp-chain model provides more challenges for the inference technique and is only able surpass the comb-chain results for the label DATE of the dataset *Curricula Vitae*. In the evaluation of the remaining label, the average error reduction is 14%.

If the comb-chain model is compared to the skyp-chain model, then it becomes apparent that the skyp-chain model with the applied inferencing technique TRP has no advantages when exploiting consistencies even at the cost of a computationally more expensive inference. Table 3 and Table 4 contain the average evaluation time for one fold. In general, it takes longer to train models with the larger dataset *Curricula Vitae*.

Table 1. F_1 scores for the segmentation of references

References	
	All
Linear Chain	0.966
Comb Chain	0.976
Skyp Chain	0.972

Table 2. F_1 scores for template extraction in curricula vitae

Curricula Vitae		
	Date	Client
Linear Chain	0.944	0.725
Comb Chain	0.962	0.814
Skyp Chain	0.962	0.764

Table 3. Average time for one fold (*References*)

References	
Linear Chain	0.03h
Comb Chain	0.17h
Skyp Chain	0.53h

Table 4. Average time for one fold (*Curricula Vitae*)

Curricula Vitae	
Linear Chain	0.11h
Comb Chain	0.27h
Skyp Chain	0.97h

4.5 Discussion

The evaluated results of the presented IE models have a valuable influence on real-world applications. An error reduction of 30% considerably improves the quality of automatically extracted entities in the database and reduces the workload to correct possible IE errors. The reported increase of the accuracy and the corresponding error reduction of the presented models compete well with published approaches for Collective IE, joint inference in IE or other models that exploit long-range dependencies.

The performance time of the presented models is in our opinion fast enough for the planned applications, but can still be increased with further optimizations or faster inference and learning techniques.

5 Related Work

In this section, we give a short overview of the related work, which can be categorized into Information Extraction (IE) publications about:

- Collective IE for Named Entity Recognition (NER).
- Collective IE with respect to structured texts.
- Collective IE with context-specific consistencies.
- Improved IE models in general, evaluated for the segmentation of references.

Collective IE is an active and popular field of research and thus we can only discuss some representatives of each category.

Models of collective approaches for NER are often motivated by two assumptions: The labeling of similar tokens is quite consistent within a given context

or document since those mentions mostly refer to the same type of entity. The discriminative features to detect the entities are sparsely distributed over the document. Thus, the accuracy for different mentions of an entity can be improved by leveraging and transferring their local context to distant positions.

Bunescu et al.[2] use Relational Markov Networks and model dependencies between distant entities. They apply special templates in order to assign equal labels if the text of the tokens is identical. The skip-chain approach introduced by Sutton et al.[15] extends linear-chain CRFs with additional factors for long-range dependencies. They link the labels of similar tokens and provide feature functions that combine evidence of both positions by which missing context can be transferred. Finkel et al. [4] criticize the usage of believe propagation and apply Gibbs sampling for enforcing label consistency and extraction template consistency constraints. All of these approaches with higher order structures fight the exponential increase in model complexity and are forced to apply approximate inference techniques instead of exact algorithms. Kou et al. [8] and Krishnan et al. [9] have shown that stacked graphical models with exact inference can compete with the accuracy of those complex models. They reduce the computational cost by applying an ensemble for two linear-chain CRFs where they aggregate the output of the first models in order to provide information about related instances or entities to a stacked model.

Yang et al. [19] and Gulhane et al. [5] presented work about IE in webforums and websites. The first approach applies Markov Logic Networks to encode properties of a typical forum page like attribute similarities among different posts and sites. The second approach developed an Apriori-style algorithm and assumes that values of an attribute distributed over different pages are similar for equal entities and the pages of one website share a similar structure due to the creation template. In contrast to our models, both approaches are domain-dependent and rely on prior knowledge about the structure.

In previous work [7], we proposed stacked CRFs in combination with rule learning techniques to exploit context-specific consistencies. The output of the first CRF was utilized to learn the manifestation of feature functions for the stacked CRF. The approach was evaluated only for the segmentation of references and achieved a significant error reduction compared to a linear-chain CRF. The stacked CRFs with feature induction during inference is similar to the comb-chain model. However, we developed a novel quality function and utilize the classification result to add new potentials instead of only normal features for a label transition. The skyp-chain approach further increases the model complexity and adds edges for long range dependencies.

The segmentation of references is a widely used domain for the evaluation of novel machine learning and IE models. The work of Peng et al. [11] provides a deep analysis of different settings and established linear-chain CRFs as the state-of-the-art for the segmentation of references. Approaches for joint inference [12, 13] combine different tasks within a model. Here, the accuracy of the labeling can be increased when entity resolution and segmentation are jointly performed. Finally, Bellare et al. [1] present a semi-supervised approach for ref-

erence segmentation by encoding expectations in higher-order constraints that cover more expressive and structural dependencies than the underlying model.

6 Conclusions

Exploiting context-specific consistencies can substantially increase the accuracy of sequence labeling in semi-structured documents. We presented two approaches based on CRFs, which combine ideas of stacked graphical models and higher-order models like skip-chain CRFs. Both approaches outperform the common models and have a valuable impact for real-world IE applications. The comb-chain CRFs are able to achieve an average error reduction of about 30% in two datasets.

For future work, two interesting improvements can be identified: On a technical level, the usage of newer inference techniques for factor graphs like SampleRank [18] should be able to avoid some of the described problems. On a more conceptual level, a joint inference approach like [13] that combines labeling and consistency identification within a probabilistic graphical model has the potential to gain further advantages in the evaluated domains.

Acknowledgments This work was supported by the Competence Network Heart Failure, funded by the German Federal Ministry of Education and Research (BMBF01 EO1004).

References

1. Bellare, K., Druck, G., McCallum, A.: Alternating Projections for Learning with Expectation Constraints. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in AI. pp. 43–50. AUAI Press (2009)
2. Bunescu, R., Mooney, R.J.: Collective Information Extraction with Relational Markov Networks. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. ACL '04, Association for Computational Linguistics, Stroudsburg, PA, USA (2004)
3. Elidan, G., McGraw, I., Koller, D.: Residual Belief Propagation: Informed Scheduling for Asynchronous Message Passing. In: Proceedings of the Twenty-second Conference on Uncertainty in AI (UAI). Boston, Massachussetts (July 2006)
4. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. pp. 363–370. ACL '05, Association for Computational Linguistics, Stroudsburg, PA, USA (2005)
5. Gulhane, P., Rastogi, R., Sengamedu, S.H., Tengli, A.: Exploiting Content Redundancy for Web Information Extraction. Proc. VLDB Endow. 3, 578–587 (2010)
6. Klösgen, W.: Explora: A Multipattern and Multistrategy Discovery Assistant. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 249–271. AAAI Press (1996)

7. Kluegl, P., Toepfer, M., Lemmerich, F., Hotho, A., Puppe, F.: Stacked Conditional Random Fields Exploiting Structural Consistencies. In: Carmona, P.L., Sánchez, J.S., Fred, A. (eds.) Proceedings of 1st International Conference on Pattern Recognition Applications and Methods (ICPRAM). pp. 240–248. SciTePress, Vilamoura, Algarve, Portugal (February 2012)
8. Kou, Z., Cohen, W.W.: Stacked Graphical Models for Efficient Inference in Markov Random Fields. In: Proceedings of the 2007 SIAM Int. Conf. on Data Mining (2007)
9. Krishnan, V., Manning, C.D.: An Effective two-stage Model for Exploiting non-local Dependencies in Named Entity Recognition. In: Proc. of the 21st Int. Conf. on Computational Linguistics and the 44th Annual Meeting of the ACL. pp. 1121–1128. ACL-44, ACL, Stroudsburg, PA, USA (2006)
10. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proc. 18th International Conf. on Machine Learning pp. 282–289 (2001)
11. Peng, F., McCallum, A.: Accurate Information Extraction from Research Papers using Conditional Random Fields. In: HLT-NAACL. pp. 329–336 (2004)
12. Poon, H., Domingos, P.: Joint Inference in Information Extraction. In: AAAI'07: Proceedings of the 22nd National Conference on Artificial intelligence. pp. 913–918. AAAI Press (2007)
13. Singh, S., Schultz, K., McCallum, A.: Bi-directional Joint Inference for Entity Resolution and Segmentation Using Imperatively-Defined Factor Graphs. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II. pp. 414–429. ECML PKDD '09, Springer-Verlag (2009)
14. Sutton, C.: GRMM: GRaphical Models in Mallet (2006), `http://mallet.cs.umass.edu/grmm/`
15. Sutton, C., McCallum, A.: Collective Segmentation and Labeling of Distant Entities in Information Extraction. In: ICML Workshop on Statistical Relational Learning and Its Connections to Other Fields (2004)
16. Sutton, C.A., McCallum, A.: Improved Dynamic Schedules for Belief Propagation. In: Parr, R., van der Gaag, L.C. (eds.) UAI. pp. 376–383. AUAI Press (2007)
17. Wainwright, M.J., Jaakkola, T., Willsky, A.S.: Tree-based Reparameterization for Approximate Inference on Loopy Graphs. In: NIPS. pp. 1001–1008 (2001)
18. Wick, M.L., Rohanimanesh, K., Bellare, K., Culotta, A., McCallum, A.: SampleRank: Training Factor Graphs with Atomic Gradients. In: Getoor, L., Scheffer, T. (eds.) ICML. pp. 777–784. Omnipress (2011)
19. Yang, J.M., Cai, R., Wang, Y., Zhu, J., Zhang, L., Ma, W.Y.: Incorporating Site-level Knowledge to Extract Structured Data from Web Forums. In: Proceedings of the 18th International Conference on World Wide Web. pp. 181–190. ACM (2009)