# Automatically Selected Skip Edges in Conditional Random Fields for Named Entity Recognition

**Roman Klinger**
Department of Bioinformatics
Fraunhofer Institute for Algorithms and Scientific Computing
Schloss Birlinghoven
53754 Sankt Augustin, Germany
roman.klinger@scai.fraunhofer.de

## Abstract

Incorporating distant information via manually selected skip chain templates has been shown to be beneficial for the performance of conditional random field models in contrast to a simple linear chain based structure (Sutton and McCallum, 2007; Galley, 2006; Liu et al., 2010). The set of properties to be captured by a template is typically manually chosen with respect to the application domain.

In this paper, a search strategy to find meaningful skip chains independent from the application domain is proposed. From a huge set of potentially beneficial templates, some can be shown to have a positive impact on the performance. The search for a meaningful graphical structure demonstrates the usefulness of the approach with an increase of nearly 2 % $F_1$ measure on a publicly available data set (Klinger et al., 2008).

## 1 Introduction

Many applications in the field of text segmentation, especially named entity recognition, have been addressed with linear chain conditional random fields. Using a linear chain of variables to represent the labeling of text is straight forward, as processing text in a sequential manner suggests itself due to the way it is written and firstly perceived.

While language suggests this linear structure to represent written text, it does not necessarily model all dependencies: Co-referencing a prior entity is an example (while it could be seen as higher order linearity typically pointing back but not forward). Especially in non-scientific texts, information may as well be left out and filled in later to keep a story exciting. Another example is the use of filler words as a trivial case where the meaning of words can be determined by distant tokens.

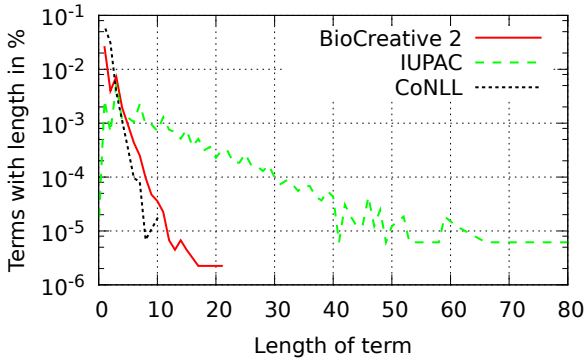


Figure 1: Distribution of the length of three entity classes.

In named entity recognition, typically linear chain structures of conditional random fields are used. The capabilities of a linear chain such structure may be limited in at least two cases: Firstly, relations between distant tokens can have an impact on their meaning. This is a motivation which lead to the previous work presented in the following Section 2. Secondly, long entity classes cannot be captured as a whole, which is especially interesting because a characteristic of named entities in biology and chemistry is their high length with inter-dependencies between tokens of an entity. The distribution of the length of terms in the classes of gene names (BioCreative 2, Smith et al. (2008)), IUPAC names (Klinger et al., 2008) and person names, organizations and places (CoNLL 2005, Sang and De Meulder (2003)) is shown in Figure 1. Gene names and especially IUPAC names are much longer than entities like names, organizations and places. This is the motivation to investigate if a linear-chain structure can be supported by other structures to capture this complexity. The work presented in this paper aims towards an automatic detection of beneficial struc-



*Proceedings of Recent Advances in Natural Language Processing*, pages 580–585,
Hissar, Bulgaria, 12-14 September 2011.

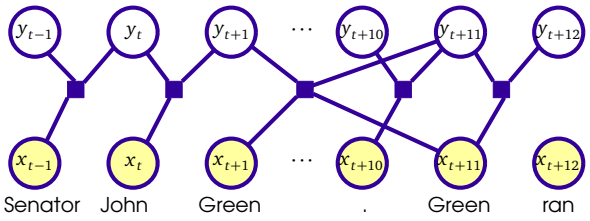


Figure 2: Example of a skip chain CRF structure as used by Sutton and McCallum (2007) (as factor graph depiction). Subsequent labels are connected as well as tokens representing the same string.

tures. The IUPAC domain with its notable long entities is used as an evaluation domain in this paper, presuming that a pure linear chain structure has specific difficulties here (using the training corpus presented by Klinger et al. (2008)).

In the following, the challenge is approached as a search for meaningful skip chain templates (Sutton and McCallum, 2007).

## 2 Previous Work

The class of CRFs including skip chain edges (unrolled from skip chain templates) has been described by Sutton and McCallum (2007) and Galley (2006) in a named entity recognition scenario. In addition to the linear chain, a template is used to measure the dependencies between same capitalized tokens. This is motivated by the assumption that same words in a sentence or document are likely to have the same label, despite their token distance. An example for such skip chain CRF is shown in Figure 2. As stated by Sutton and McCallum (2007), each pair of nodes can be connected by a skip chain which the developer believes to be similar. They point out that the number of edges unrolled from a template may not be too high as the runtime and memory consumption increases prohibitively. Connecting only capitalized words allows to match most proper names (which is an entity class of interest in their test domain) while they are sparsely distributed.

The work by Liu et al. (2010) enhances that approach by different classes of variables (as special keywords) to be connected. To adapt Sutton's and McCallum's approach to gene and protein names, they introduce three skip chain templates: Firstly, connecting the main parts of gene names (referred to as "keyword" in their work) defined by regular expressions, secondly connecting only similar keywords, only differing to a certain extent, and thirdly connecting typed dependencies like prepositional modifiers or noun compound modifiers. On the BioCreative 2 NER data set (Smith et al., 2008) they show an increase in $F_1$ measure from $71.73\,\%$ with a linear chain to $73.14\,\%$ with the best skip chain configuration for a strict evaluation not using the allowed alternatives in the gold data test set. Using the official evaluation with alternatives, they show an increase from $83.29\,\%$ to $84.67\,\%$ $F_1$. They argue that the quality of the skip chains is essential for the improvement of the result compared to simple linear chain structures.

In contrast to previous work, this paper addresses the question how to select meaningful skip edges automatically from a set of possibilities. This does not make the domain specific development unnecessary (as the application of feature selection still needs the development of features for a domain), but helps to find templates to improve the results. It can select a specific subset of automatically generated clique templates. This task can be understood as a combinatorial optimization problem: Finding a factor graph with a structure maximizing the performance of the model on a test set.

Several approaches have been published about optimizing the structure of Markov networks or more specifically conditional random fields. They can be divided into methods searching for such structure with a measure to judge the quality of a structure and filtering approaches to decide about the quality of an edge. Beside those, regularization is another way to find a good structure during training.

The work by Schmidt et al. (2008) states to be the first dealing with the structure learning task in discrimatively trained, undirected graphical models. Similarly to Lee et al. (2006) (which is dealing with general Markov networks), $L_1$ regularization is the incorporated method. While this approach is very elegant due to the joint structure and parameter estimation, it has limitations to deal with large, dynamically generated factor graphs with a lot of features on each factor.

As long as the candidates for the optimal structure have a tractable size, a search in the space of graph structures is feasible. This approach, together with an approximation for the quality measure of each graph is adopted by Parise and Welling (2006). The advantage is that all depen-

dencies in the graph are taken into account, the disadvantage is the complexity of the performed search.

A complementary and fast approach is to measure the quality of an edge with independence tests, as described by Bromberg et al. (2009). The main contribution in their work is to minimize the needed independence tests to find the optimal graph structure.

## 3 Methods

### 3.1 Problem Definition

The work described in Section 2 is focusing on graph structures of limited size or on non-conditional Markov graphs. In the following, the problem of finding skip edges is discussed in detail.

A graph structure $G = (V, E)$ is defined via vertexes $V$ and edges $E = V \times V$. Optimizing the structure corresponds to selecting a subset of edges which leads to the maximal performance. A factor graph (Kschischang et al., 2001) is a bipartite graph $G$ between variables and factors defining a probability distribution of a set of output variables $\vec{y}$ conditioned on input variables $\vec{x}$. Each factor $\Psi_j$ computes the so-called score of variables which are neighbors in the graph. It is typically formulated as an exponential function of the weighted sum of features:

$$\Psi_j(\vec{x}, \vec{y}) = \exp\left(\sum_{i=0}^{m} \lambda_i f_i(\vec{x}_j, \vec{y}_j)\right) .$$

A set of factor templates $\Theta = \{\theta_1, \ldots, \theta_n\}$ consists of templates $\theta_k$ describing a set of tuples $\{(\vec{x}_k, \vec{y}_k)\}$ on which factors are instantiated for which the property $p_k(\vec{x}_k, \vec{y}_k)$ holds and shares $\vec{\lambda}_k$ and $\vec{f}_k(\cdot)$ between all instantiated factors on the tuples. $K_j$ is the number of parameters of the $j$th template. The probability distribution on a factor graph with templates $\Theta$ becomes

$$P(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})} \prod_{\theta_j \in \Theta} \prod_{(\vec{x}_i, \vec{y}_i) \in \theta_j} \exp\left(\sum_{k=1}^{K_j} \lambda_{jk} f_{jk}(\vec{x}_i, \vec{y}_i)\right).$$

The task of finding meaningful skip chains corresponds to finding a set of templates $\tilde{\Theta}$ describing tuples $(y_u, y_v, \vec{x})$ with a property $p(x_u, x_v)$. A linear chain template $\theta_{\text{lc}}$ with features $\vec{g}^{\text{lc}}(\vec{x}, j)$ for all possible combinations of label variables is assumed to be present in all configurations. In the following, the set of templates to select from is defined by properties $p_k(x_u, x_v) :=$ holds iff $g_k^{\text{lc}}(\vec{x}, u) = 1$ and $g_k^{\text{lc}}(\vec{x}, v) = 1$ with $k \in \{1, \ldots, |\vec{g}^{\text{lc}}|\}$. Each template holds parameters for features $g_k^{\text{lc}}(\vec{x}, u) \vee g_k^{\text{lc}}(\vec{x}, v)$.[1] In other words, a skip chain factor is added to connect two labeling variables of two tokens if a specified property holds (where we take every occurring feature specified for the linear chain into account, like bag-of-words, prefixes, suffixes of tokens as well as several regular expressions, the full set is given by Klinger et al. (2008)). Each of the skip chain factors has the disjunction of the feature values in the linear chain of the connected tokens.

That definition of the templates to choose from is only for simplicity throughout this paper. A more general formulation does not limit the methods described, though a small set decreases runtime. Especially dependency properties as described by Liu et al. (2010) may be included.

An example of different skip chain factors to choose from is shown in Figure 3. The red property of matching `[.*ine]` seems to be a reasonable skip chain as it connects similar chemical names such that their class can influence the class of the others. The green matching stop words is an example how the size of the factor graph can prohibitively increase what should be avoided. The orange matching of `[,]` could be able to capture enumerations because features which take preceding and succeeding tokens into account may have some importance.

### 3.2 Best First Search

The most complete approach to find a suitable structure of the graph is a search through the space of all combinations of the skip chain templates. As the complexity is prohibitive even for a small set of templates, the dependencies between possible templates are proposed to be measured via best-first-search (BFS, Russell and Norvig (2003)) on all templates remaining from a heuristic filtering step together with the linear chain factors. Best-first-search is chosen as an exemplary search strategy as it follows only the best

[1] This definition of features on the skip chain factors has been chosen to capture not only shared properties but to measure characteristics of only one participant as well.

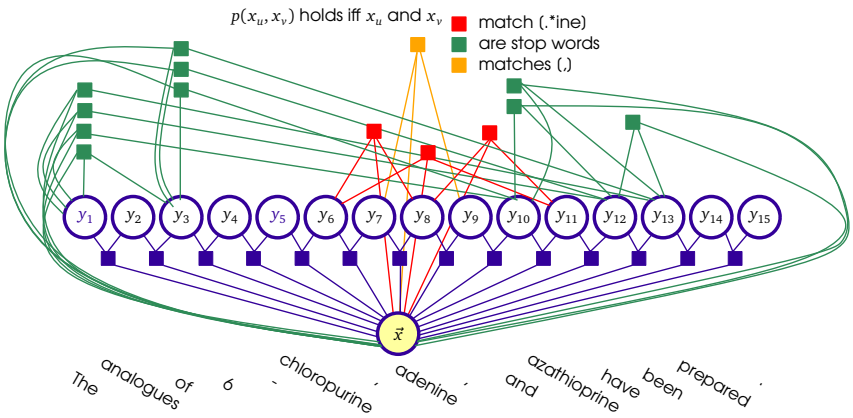


Figure 3: Different skip chain factor templates to choose additionally to the linear chain (example shortened from the abstract by Hasan and Srivastava (1992)).

alternative which limits the performance evaluations needed.

Starting with the linear chain, each template is added respectively, the model is trained and evaluated on a hold-out set. The best template is kept and the prior step is repeated. This process ends if none of the templates can improve the result.

Here, the same inference algorithm is used as for the final model: Loopy belief propagation with tree-based reparameterization for approximate inference (Wainwright et al., 2001). During the steps of BFS, the weights for each template to be kept can be adopted for the next search iteration retraining all parameters. Thereby retraining the model can be performed in a much smaller number of iterations than training from scratch.

## 4 Experiments and Evaluation

In the following, the feasibility of the basic idea of the proposed approach is evaluated on the IUPAC corpus (Klinger et al. (2008)[2], split into 90 % training and 10 % validation randomly). All templates which occur at least 10 times in the training data were taken into account. These are 5096 templates (from 16710 altogether). Training with and without the skip factors (via maximizing $\log P(\vec{y}|\vec{x})$ with a Gaussian regularization) leads to a ranked list of templates which forms the first layer of the BFS.

This leads to the results depicted in Figure 4 showing the impact of each template. The inference algorithm did not converge for 474 templates. Most of the templates have no or negative impact on the result (3656 templates); 966 have positive impact. The top 10 templates are depicted in Table 1, together with their contribution. The most important template is to add a skip chain between tokens "alpha" with nearly 2 % improvement in comparison to the linear chain only. This term occurs 319 times in the training data and is frequently part of IUPAC names (115) as well as outside of them (204). In a local, linear chain-based setting a feature based on this token can hardly contribute to a decision, but in a distant labeling setting it can. The second best feature to built skip chain factors is "PREFIX2=tr": It occurs 698 times, 284 times in IUPAC names and 414 times outside of them.

Most of the features forming the basis for templates with a positive impact are words occurring close to or in chemical names or are typical chemical pre- or suffixes. These features measure ambiguous characteristics of tokens where the probability of correctly identifying the surrounding terms can be increased by measuring the distant information of them. The context is taken into account by templates based on features of offset conjunction (like `W=alpha@2` measuring the token `alpha` two tokens left of the skip chain connection). The reason is presumably that their common occurrence in a sentence is not labeled differently.

For instance, the term `alpha` is occurring in `alpha-ribofuranosyl` (which is labeled as IUPAC) and in `alpha1-adrenergic` (not labeled as IUPAC). In both examples, `alpha` is occurring multiple times in the text, but not with different labels. Similarly, `tr` can be a prefix of `tributylstannyl` (as IUPAC) or

[2]Statistics of this data set can be found in the original publication and are not cited here due to page limitation.

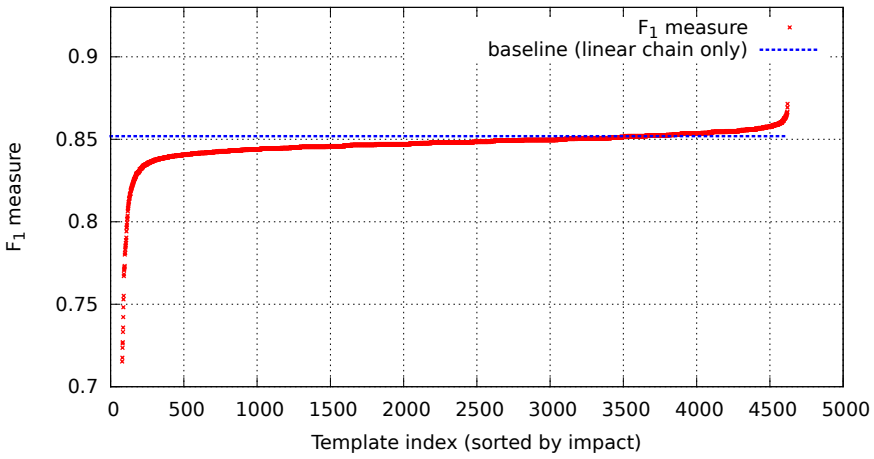


Figure 4: Impact for each proposed templates measured empirically.

`treatment` (as non-IUPAC), but it is not probable that different labels occur in the same text. The feature `W=group@1` is a slightly different case as it does not occur as part of an IUPAC name itself but can occur in the context of chemical names which are difficult to distinguish between IUPAC and not. As an example, `formamidino group` would not be labeled as IUPAC, but `p-methoxybenyl group` is labeled as such. Annotation is quite difficult here, but it is likely that the annotator produced consistent data in one text instance.

Notable is the occurrence of very strict features like `SUFFIX2=31`—it is surprising that obviously some numbers are occurring frequently in IUPAC names and outside of them such that a differentiation with distant information is beneficial.

This discussion illustrates that the found templates are meaningful in the context of the IUPAC example taken for evaluation here. The impact of the automated approach for the first layer of a search shows similar or better possible improvements than the manual approaches by Sutton and McCallum (2007, *cf.* Table 1.2) (0.4 %) or Liu et al. (2010) (1.2 % on BioCreative II data) (this is remarkable although they tested on different data sets).

## 5 Conclusion and Future Work

This paper presented the principle idea of building skip chain edges to capture distant information for named entity recognition in a similar manner as features to represent tokens are generated. Instead of hand-crafting domain- and problem-specific features, they are generated from the training data; analogously, potentially beneficial skip chain templates are taken into account. It has been shown that this approach is feasible and leads to an improved performance on an example domain.

To be able to apply this methodology in practice, the search complexity for meaningful structures needs to be reduced. Nevertheless, the analysis shows that the idea of automatically selecting distant tokens as a basis for additional factors makes sense. The presented analysis can help in further work and be used as a training set for novel template filtering methods.

Future work includes the analysis how different templates work together for named entity recognition: What is the relation between the linear chain and a skip chain? What characteristics does the linear chain have where skip chains help?

Additionally due to the high complexity of empirically testing all templates, the interaction between different skip chains has not been analyzed.

Another interesting topic is to investigate the impact of specific factors on different evaluation measures. It can be assumed that some support accuracy, whereas some have a special impact on precision or recall.

## References

[Bromberg et al.2009] Facundo Bromberg, Dimitris Margaritis, and Vasant Honavar. 2009. Efficient markov network structure discovery using independence tests. *Journal of Artificial Intelligence Research*, 35(1):449–484, May.

[Galley2006] Michel Galley. 2006. A skip-chain conditional random field for ranking meeting utterances

| | Template | $F_1$ measure | to base |
|---|---|---|---|
| 1 | W=alpha@2 | 87.15 | 1.96 |
| 2 | PREFIX2=tr | 86.93 | 1.74 |
| 3 | W=liver@1 | 86.66 | 1.47 |
| 4 | W=group@1 | 86.63 | 1.44 |
| 5 | WC=AAA@1 | 86.59 | 1.40 |
| 6 | REFIX2=ox | 86.59 | 1.40 |
| 7 | PREFIX2=fu@1 | 86.54 | 1.35 |
| 8 | W=rac@1 | 86.51 | 1.32 |
| 9 | W=cis@-1 | 86.47 | 1.28 |
| 10 | SUFFIX2=nt | 86.44 | 1.25 |
| 11 | SUFFIX2=ro@-2 | 86.39 | 1.20 |
| 12 | W=was@-2 | 86.39 | 1.20 |
| 13 | W=rac@-1 | 86.39 | 1.20 |
| 14 | PREFIX2=hy@1 | 86.39 | 1.20 |
| 15 | PREFIX2=am@1 | 86.36 | 1.17 |
| 16 | SUFFIX2=31 | 86.32 | 1.13 |
| 17 | SUFFIX2=,4 | 86.32 | 1.13 |
| 18 | SUFFIX2=in@2 | 86.28 | 1.09 |
| 19 | SUFFIX2=,4@1 | 86.28 | 1.09 |
| 20 | W=3H@-1 | 86.28 | 1.09 |

Table 1: Top 20 skip chain templates from all proposed templates occurring at least 10 times. ("To base" denotes the difference to the baseline, a linear chain CRF. W= denotes exact term match features, PREFIX2= the use of the string as prefix of length 2, analogously for SUFFIX. @n denotes the use of the feature characterizing tokens with an offset of n in the token sequence.)


by importance. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 364–372, Sydney, Australia, July. Association for Computational Linguistics.

[Hasan and Srivastava1992] A. Hasan and P. C. Srivastava. 1992. Synthesis and biological studies of unsaturated acyclonucleoside analogues of s-adenosyl-l-homocysteine hydrolase inhibitors. *J Med Chem*, 35(8):1435–1439, Apr.

[Klinger et al.2008] Roman Klinger, Corinna Kolářik, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M. Friedrich. 2008. Detection of IUPAC and IUPAC-like Chemical Names. *Bioinformatics*, 24(13):i268–i276. Proceedings of the International Conference Intelligent Systems for Molecular Biology (ISMB).

[Kschischang et al.2001] Frank Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. 2001. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519.

[Lee et al.2006] Su-In Lee, Varun Ganapathi, and Daphne Koller. 2006. Efficient structure learning of markov networks using l1-regularization. In *Advances in Neural Information Processing Systems*.

[Liu et al.2010] Jingchen Liu, Minlie Huang, and Xiaozan Zhu. 2010. Recognizing biomedical named entitities using skip-chain conditional random fields. In *Proceedings of the Workshop on Biomedical Natural Language Processing*.

[Parise and Welling2006] Sridevi Parise and Max Welling. 2006. Structure learning in markov random fields. In *Advances in Neural Information Processing Systems*.

[Russell and Norvig2003] Stuart Russell and Peter Norvig. 2003. *Artificial Intelligence – A Modern Approach*. Prentice Hall.

[Sang and De Meulder2003] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of Computational Natural Language Learning (CoNLL)*, pages 142–147. Edmonton, Canada.

[Schmidt et al.2008] Mark Schmidt, Kevin Murphy, Glenn Fung, and R'omer Rosales. 2008. Structure learning in random fields for heart motion abnormality detection. In *Computer Vision and Pattern Recognition*, Anchorage, AK, USA, June. IEEE.

[Smith et al.2008] Larry Smith, Lorraine K. Tanabe, Rie Johnson nee Ando, Cheng-Ju Juo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M. Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A. Struble, Richard J. Povinelli, Andreas Vlachos, William A. Baumgartner Jr., Lawrence Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafel Torres Perez, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Mana, Jacinto Mata-Vazquez, and W. John Wilbur. 2008. Overview of biocreative ii gene mention recognition. *Genome Biology*, 9(Suppl 2):S2.2–S2.18, September.

[Sutton and McCallum2007] Charles Sutton and Andrew McCallum. 2007. An introduction to conditional random fields for relational learning. In Lise Getoor and Benjamin Taskar, editors, *Introduction to Statistical Relational Learning*, chapter 4, pages 93–127. MIT Press, November.

[Wainwright et al.2001] Martin J. Wainwright, Tommi Jaakkola, and Alan S. Willsky. 2001. Tree-based reparameterization for approximate inference on loopy graphs. In *Proccedings of the Conference on Neural Information Processing Systems*.