An Effective Two-Stage Model for Exploiting Non-Local Dependencies in Named Entity Recognition

Vijay Krishnan
Computer Science Department
Stanford University
Stanford, CA 94305
vijayk@cs.stanford.edu

Christopher D. Manning
Computer Science Department
Stanford University
Stanford, CA 94305
manning@cs.stanford.edu

Abstract

This paper shows that a simple two-stage approach to handle non-local dependencies in Named Entity Recognition (NER) can outperform existing approaches that handle non-local dependencies, while being much more computationally efficient. NER systems typically use sequence models for tractable inference, but this makes them unable to capture the long distance structure present in text. We use a Conditional Random Field (CRF) based NER system using local features to make predictions and then train another CRF which uses both local information and features extracted from the output of the first CRF. Using features capturing non-local dependencies from the same document, our approach yields a 12.6% relative error reduction on the F1 score, over state-of-the-art NER systems using local-information alone, when compared to the 9.3% relative error reduction offered by the best systems that exploit non-local information. Our approach also makes it easy to incorporate non-local information from other documents in the test corpus, and this gives us a 13.3% error reduction over NER systems using local-information alone. Additionally, our running time for inference is just the inference time of two sequential CRFs, which is much less than that of other more complicated approaches that directly model the dependencies and do approximate inference.

1 Introduction

Named entity recognition (NER) seeks to locate and classify atomic elements in unstructured text into predefined entities such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. A particular problem for Named Entity Recognition(NER) systems is to exploit the presence of useful information regarding labels assigned at a long distance from a given entity. An example is the label-consistency constraint that if our text has two occurrences of *New York* separated by other tokens, we would want our learner to encourage both these entities to get the same label.

Most statistical models currently used for Named Entity Recognition, use sequence models and thereby capture local structure. Hidden Markov Models (HMMs) (Leek, 1997; Freitag and McCallum, 1999), Conditional Markov Models (CMMs) (Borthwick, 1999; McCallum et al., 2000), and Conditional Random Fields (CRFs) (Lafferty et al., 2001) have been successfully employed in NER and other information extraction tasks. All these models encode the Markov property i.e. labels directly depend only on the labels assigned to a small window around them. These models exploit this property for tractable computation as this allows the Forward-Backward, Viterbi and Clique Calibration algorithms to become tractable. Although this constraint is essential to make exact inference tractable, it makes us unable to exploit the non-local structure present in natural language.

Label consistency is an example of a non-local dependency important in NER. Apart from label consistency between the same token sequences, we would also like to exploit richer sources of dependencies between similar token sequences. For example, as shown in Figure 1, we would want it to encourage *Einstein* to be labeled "Person" if there is strong evidence that *Albert Einstein* should be labeled "Person". Sequence models unfortu-

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 1121–1128,
Sydney, July 2006.

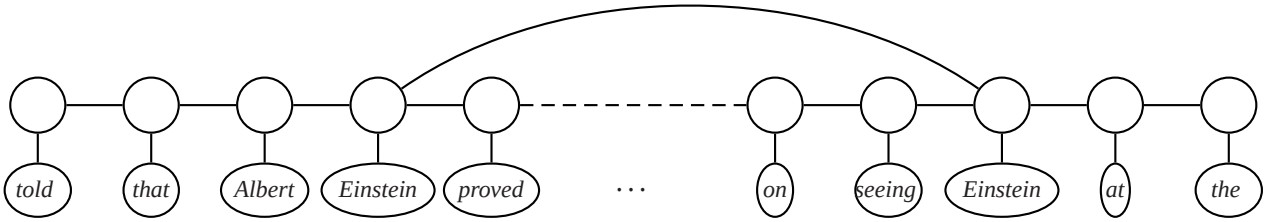

Figure 1: An example of the label consistency problem. Here we would like our model to encourage entities *Albert Einstein* and *Einstein* to get the same label, so as to improve the chance that both are labeled *PERSON*.

nately cannot model this due to their Markovian assumption.

Recent approaches attempting to capture non-local dependencies model the non-local dependencies directly, and use approximate inference algorithms, since exact inference is in general, not tractable for graphs with non-local structure.

Bunescu and Mooney (2004) define a *Relational Markov Network* (RMN) which explicitly models long-distance dependencies, and use it to represent relations between entities. Sutton and McCallum (2004) augment a sequential CRF with *skip-edges* i.e. edges between different occurrences of a token, in a document. Both these approaches use loopy belief propagation (Pearl, 1988; Yedidia et al., 2000) for approximate inference.

Finkel et al. (2005) hand-set penalties for inconsistency in entity labeling at different occurrences in the text, based on some statistics from training data. They then employ Gibbs sampling (Geman and Geman, 1984) for dealing with their local feature weights and their non-local penalties to do approximate inference.

We present a simple two-stage approach where our second CRF uses features derived from the output of the first CRF. This gives us the advantage of defining a rich set of features to model non-local dependencies, and also eliminates the need to do approximate inference, since we do not explicitly capture the non-local dependencies in a single model, like the more complex existing approaches. This also enables us to do inference efficiently since our inference time is merely the inference time of two sequential CRF's; in contrast Finkel et al. (2005) reported an increase in running time by a factor of 30 over the sequential CRF, with their Gibbs sampling approximate inference.

In all, our approach is simpler, yields higher F1 scores, and is also much more computationally efficient than existing approaches modeling non-local dependencies.

2 Conditional Random Fields

We use a Conditional Random Field (Lafferty et al., 2001; Sha and Pereira, 2003) since it represents the state of the art in sequence modeling and has also been very effective at Named Entity Recognition. It allows us both discriminative training that CMMs offer as well and the bi-directional flow of probabilistic information across the sequence that HMMs allow, thereby giving us the best of both worlds. Due to the bi-directional flow of information, CRFs guard against the myopic locally attractive decisions that CMMs make. It is customary to use the Viterbi algorithm, to find the most probably state sequence during inference. A large number of possibly redundant and correlated features can be supplied without fear of further reducing the accuracy of a high-dimensional distribution. These are well-documented benefits (Lafferty et al., 2001).

2.1 Our Baseline CRF for Named Entity Recognition

Our baseline CRF is a sequence model in which labels for tokens directly depend only on the labels corresponding to the previous and next tokens. We use features that have been shown to be effective in NER, namely the current, previous and next words, character n-grams of the current word, Part of Speech tag of the current word and surrounding words, the shallow parse chunk of the current word, shape of the current word, the surrounding word shape sequence, the presence of a word in a left window of size 5 around the current word and the presence of a word in a left window of size 5 around the current word. This gives us a competitive baseline CRF using local information alone, whose performance is close to the best published local CRF models, for Named Entity Recognition

3 Label Consistency

The intuition for modeling label consistency is that within a particular document, different occur-

	Document Level Statistics				Corpus Level Statistics			
	PER	LOC	ORG	MISC	PER	LOC	ORG	MISC
PER	3141	4	5	0	33830	113	153	0
LOC		6436	188	3		346966	6749	60
ORG			2975	0			43892	223
MISC				2030				66286

Table 1: Table showing the number of pairs of different occurrences of the same token sequence, where one occurrence is given a certain label and the other occurrence is given a certain label. We show these counts both within documents, as well as over the whole corpus. As we would expect, most pairs of the same entity sequence are labeled the same(i.e. the diagonal has most of the density) at both the document and corpus levels. These statistics are from the CoNLL 2003 English training set.

	Document Level Statistics				Corpus Level Statistics			
	PER	LOC	ORG	MISC	PER	LOC	ORG	MISC
PER	1941	5	2	3	9111	401	261	38
LOC	0	167	6	63	68	4560	580	1543
ORG	22	328	819	191	221	19683	5131	4752
MISC	14	224	7	365	50	12713	329	8768

Table 2: Table showing the number of (token sequence, token subsequence) pairs where the token sequence is assigned a certain entity label, and the token subsequence is assigned a certain entity label. We show these counts both within documents, as well as over the whole corpus. Rows correspond to sequences, and columns to subsequences. These statistics are from the CoNLL 2003 English training set.

rences of a particular token sequence (or similar token sequences) are unlikely to have different entity labels. While this constraint holds strongly at the level of a document, there exists additional value to be derived by enforcing this constraint less strongly across different documents. We want to model label consistency as a *soft* and not a *hard* constraint; while we want to encourage different occurrences of similar token sequences to get labeled as the same entity, we do not want to force this to always hold, since there do exist exceptions, as can be seen from the off-diagonal entries of tables 1 and 2.

A named entity recognition system modeling this structure would encourage all the occurrences of the token sequence to the same entity type, thereby sharing evidence among them. Thus, if the system has strong evidence about the label of a given token sequence, but is relatively unsure about the label to be assigned to another occurrence of a similar token sequence, the system can gain significantly by using the information about the label assigned to the former occurrence, to label the relatively ambiguous token sequence, leading to accuracy improvements.

The strength of the label consistency constraint, can be seen from statistics extracted from the CoNLL 2003 English training data. Table 1 shows the counts of entity labels pairs assigned for each pair of identical token sequences both within a document and across the whole corpus. As we would expect, inconsistent labelings are relatively rare and most pairs of the same entity sequence are labeled the same(i.e. the diagonal has most of the density) at both the document and corpus levels. A notable exception to this is the labeling of the same text as both organization and location within the same document and across documents. This is a due to the large amount of sports news in the CoNLL dataset due to which city and country names are often also team names. We will see that our approach is capable of exploiting this as well, i.e. we can learn a model which would not penalize an Organization-Location inconsistency as strongly as it penalizes other inconsistencies.

In addition, we also want to model subsequence constraints: having seen *Albert Einstein* earlier in a document as a person is a good indicator that a subsequent occurrence of *Einstein* should also be labeled as a person. Here, we would expect that a subsequence would gain much more by knowing the label of a supersequence, than the other way around.

However, as can be seen from table 2, we find that the consistency constraint does not hold nearly so strictly in this case. A very common case of this in the CoNLL dataset is that of documents containing references to both *The China Daily*, a newspaper, and *China*, the country (Finkel et al., 2005). The first should be labeled as an organization, and second as a location. The counts of subsequence labelings within a document and across documents listed in Table 2, show that there are many off-diagonal entries: the *China Daily* case is among the most common, occurring 328 times in the dataset. Just as we can model off-diagonal pat-

terns with exact token sequence matches, we can also model off-diagonal patterns for the token subsequence case.

In addition, we could also derive some value by enforcing some label consistency at the level of an individual token. Obviously, our model would learn much lower weights for these constraints, when compared to label consistency at the level of token sequences.

4 Our Approach to Handling non-local Dependencies

To handle the non-local dependencies between same and similar token sequences, we define three sets of feature pairs where one member of the feature pair corresponds to a function of aggregate statistics of the output of the first CRF at the document level, and the other member corresponds to a function of aggregate statistics of the output of the first CRF over the whole test corpus. Thus this gives us six additional feature types for the second round CRF, namely Document-level Token-majority features, Document-level Entity-majority features, Document-level Superentity-majority features, Corpus-level Token-majority features, Corpus-level Entity-majority features and Corpus-level Superentity-majority features. These feature types are described in detail below.

All these features are a function of the output labels of the first CRF, where predictions on the test set are obtained by training on all the data, and predictions on the train data are obtained by 10 fold cross-validation (details in the next section). Our features fired based on document and corpus level statistics are:

- **Token-majority features:** These refer to the majority label assigned to the particular token in the document/corpus. Eg: Suppose we have three occurrences of the token Australia, such that two are labeled *Location* and one is labeled *Organization*, our token-majority feature would take value *Location* for all three occurrences of the token. This feature can enable us to capture some dependence between token sequences corresponding to a single entity and having common tokens.

- **Entity-majority features:** These refer to the majority label assigned to the particular entity in the document/corpus. Eg: Suppose we have three occurrences of the *entity sequence* (we define it as a token sequence labeled as a single entity by the first stage CRF) *Bank of Australia*, such that two are labeled *Organization* and one is labeled *Location*, our entity-majority feature would take value *Organization* for all tokens in all three occurrences of the entity sequence. This feature enables us to capture the dependence between identical entity sequences. For token labeled as not a Named Entity by the first CRF, this feature returns the majority label assigned to that token when it occurs as a single token named entity.

- **Superentity-majority features:** These refer to the majority label assigned to supersequences of the particular entity in the document/corpus. By entity supersequences, we refer to entity sequences, that strictly contain within their span, another entity sequence. For example, if we have two occurrences of *Bank of Australia* labeled *Organization* and one occurrence of *Australia Cup* labeled *Miscellaneous*, then for all occurrences of the entity *Australia*, the superentity-majority feature would take value *Organization*. This feature enables us to take into account labels assigned to supersequences of a particular entity, while labeling it. For token labeled as not a Named Entity by the first CRF, this feature returns the majority label assigned to all entities containing the token within their span.

The last feature enables entity sequences to benefit from labels assigned to entities which are entity supersequences of it. We attempted to add subentity-majority features, analogous to the superentity-majority features to model dependence on entity subsequences, but got no benefit from it. This is intuitive, since the basic sequence model would usually be much more certain about labels assigned to the entity supersequences, since they are longer and have more contextual information. As a result of this, while there would be several cases in which the basic sequence model would be uncertain about labels of entity subsequences but relatively certain about labels of token supersequences, the converse is very unlikely. Thus, it is difficult to profit from labels of entity subsequences while labeling entity sequences. We also attempted using more fine

grained features corresponding to the majority label of supersequences that takes into account the position of the entity sequence in the entity supersequence(whether the entity sequence occurs in the start, middle or end of the supersequence), but could obtain no additional gains from this.

It is to be noted that while deciding if token sequences are equal or hold a subsequence-supersequence relation, we ignore case, which clearly performs better than being sensitive to case. This is because our dataset contains several entities in allCaps such as *AUSTRALIA*, especially in news headlines. Ignoring case enables us to model dependences with other occurrences with a different case such as *Australia*.

It may appear at first glance, that our framework can only learn to encourage entities to switch to the most popular label assigned to other occurrences of the entity sequence and similar entity sequences. However this framework is capable of learning interesting off-diagonal patterns as well. To understand this, let us consider the example of different occurrences of token sequences being labeled *Location* and *Organization*. Suppose, the majority label of the token sequence is *Location*. While this majority label would encourage the second CRF to switch the labels of all occurrences of the token sequence to *Location*, it would not strongly discourage the CRF from labeling these as *Organization*, since there would be several occurrences of token sequences in the training data labeled *Organization*, with the majority label of the token sequence being *Location*. However it would discourage the other labels strongly. The reasoning is analogous when the majority label is *Organization*.

In case of a tie (when computing the majority label), if the label assigned to a particular token sequence is one of the majority labels, we fire the feature corresponding to that particular label being the majority label, instead of breaking ties arbitrarily. This is done to encourage the second stage CRF to make its decision based on local information, in the absence of compelling non-local information to choose a different label.

5 Advantages of our approach

With our two-stage approach, we manage to get improvements on the F1 measure over existing approaches that model non-local dependencies. At the same time, the simplicity of our two-stage approach keeps inference time down to just the inference time of two sequential CRFs, when compared to approaches such as those of Finkel et al. (2005) who report that their inference time with Gibbs sampling goes up by a factor of about 30, compared to the Viterbi algorithm for the sequential CRF.

Below, we give some intuition about areas for improvement in existing work and explain how our approach incorporates the improvements.

- Most existing work to capture label-consistency, has attempted to create all $\binom{n}{2}$ pairwise dependencies between the different occurrences of an entity, (Finkel et al., 2005; Sutton and McCallum, 2004), where n is the number of occurrences of the given entity. This complicates the dependency graph making inference harder. It also leads to the penalty for deviation in labeling to grow linearly with n, since each entity would be connected to $\Theta(n)$ entities. When an entity occurs several times, these models would force all occurrences to take the same value. This is not what we want, since there exist several instances in real-life data where different entities like persons and organizations share the same name. Thus, our approach makes a certain entity's label depend on certain *aggregate information* of other labels assigned to the same entity, and does not enforce pairwise dependencies.

- We also exploit the fact that the predictions of a learner that takes non-local dependencies into account would have a good amount of overlap with a sequential CRF, since the sequence model is already quite competitive. We use this intuition to approximate the aggregate information about labels assigned to other occurrences of the entity by the non-local model, with the aggregate information about labels assigned to other occurrences of the entity by the sequence model. This intuition enables us to learn weights for non-local dependencies in two stages; we first get predictions from a regular sequential CRF and in turn use aggregate information about predictions made by the CRF as extra features to train a second CRF.

- Most work has looked to model non-local dependencies only within a document (Finkel

et al., 2005; Chieu and Ng, 2002; Sutton and McCallum, 2004; Bunescu and Mooney, 2004). Our model can capture the weaker but still important consistency constraints across the whole document collection, whereas previous work has not, for reasons of tractability. Capturing label-consistency at the level of the whole test corpus is particularly helpful for token sequences that appear only once in their documents, but occur a few times over the corpus, since they do not have strong non-local information from within the document.

- For training our second-stage CRF, we need to get predictions on our train data as well as test data. Suppose we were to use the same train data to train the first CRF, we would get unrealistically good predictions on our train data, which would not be reflective of its performance on the test data. One option is to partition the train data. This however, can lead to a drop in performance, since the second CRF would be trained on less data. To overcome this problem, we make predictions on our train data by doing a 10-fold cross validation on the train data. For predictions on the test data, we use all the training data to train the CRF. Intuitively, we would expect that the quality of predictions with 90% of the train data would be similar to the quality of predictions with all the training data. It turns out that this is indeed the case, as can be seen from our improved performance.

6 Experiments

6.1 Dataset and Evaluation

We test the effectiveness of our technique on the CoNLL 2003 English named entity recognition dataset downloadable from *http://cnts.uia.ac.be/conll2003/ner/*. The data comprises Reuters newswire articles annotated with four entity types: *person* (PER), *location* (LOC), *organization* (ORG), and *miscellaneous* (MISC). The data is separated into a training set, a development set (testa), and a test set (testb). The training set contains 945 documents, and approximately 203,000 tokens and the test set has 231 documents and approximately 46,000 tokens. Performance on this task is evaluated by measuring the precision and recall of annotated entities (and not tokens), combined into an F1 score. There is no partial credit for labeling part of an entity sequence correctly; an incorrect entity boundary is penalized as both a false positive and as a false negative.

6.2 Results and Discussion

It can be seen from table 3, that we achieve a 12.6% relative error reduction, by restricting ourselves to features approximating non-local dependency within a document, which is higher than other approaches modeling non-local dependencies within a document. Additionally, by incorporating non-local dependencies across documents in the test corpus, we manage a 13.3% relative error reduction, over an already competitive baseline. We can see that all three features approximating non-local dependencies within a document yield reasonable gains. As we would expect the additional gains from features approximating non-local dependencies across the whole test corpus are relatively small.

We use the approximate randomization test (Yeh, 2000) for statistical significance of the difference between the basic sequential CRF and our second round CRF, which has additional features derived from the output of the first CRF. With a 1000 iterations, our improvements were statistically significant with a p-value of 0.001. Since this value is less than the cutoff threshold of 0.05, we reject the null hypothesis.

The simplicity of our approach makes it easy to incorporate dependencies across the whole corpus, which would be relatively much harder to incorporate in approaches like (Bunescu and Mooney, 2004) and (Finkel et al., 2005). Additionally, our approach makes it possible to do inference in just about twice the inference time with a single sequential CRF; in contrast, approaches like Gibbs Sampling that model the dependencies directly can increase inference time by a factor of 30 (Finkel et al., 2005).

An analysis of errors by the first stage CRF revealed that most errors are that of single token entities being mislabeled or missed altogether followed by a much smaller percentage of multiple token entities mislabelled completely. All our features directly encode information that is useful to reducing these errors. The widely prevalent boundary detection error is that of missing a single-token entity (i.e. labeling it as *Other(O)*). Our approach helps correct many such errors based on occurrences of the token in other

F1 scores on the CoNLL Dataset						
Approach	LOC	ORG	MISC	PER	ALL	Relative Error reduction
Bunescu and Mooney (2004) (Relational Markov Networks)						
Only Local Templates	-	-	-	-	**80.09**	
Global and Local Templates	-	-	-	-	**82.30**	11.1%
Finkel et al. (2005)(Gibbs Sampling)						
Local+Viterbi	88.16	80.83	78.51	90.36	**85.51**	
Non Local+Gibbs	88.51	81.72	80.43	92.29	**86.86**	9.3%
Our Approach with the 2-stage CRF						
Baseline CRF	88.09	80.88	78.26	89.76	**85.29**	
+ Document token-majority features	89.17	80.15	78.73	91.60	86.50	
+ Document entity-majority features	89.50	81.98	79.38	91.74	86.75	
+ Document superentity-majority features	89.52	82.27	79.76	92.71	**87.15**	12.6%
+ Corpus token-majority features	89.48	82.36	79.59	92.65	87.13	
+ Corpus entity-majority features	89.72	82.40	79.71	92.65	87.23	
+ Corpus superentity-majority features (All features)	89.80	82.39	79.76	92.57	**87.24**	13.3%

Table 3: Table showing improvements obtained with our additional features, over the baseline CRF. We also compare our performance against (Bunescu and Mooney, 2004) and (Finkel et al., 2005) and find that we manage higher relative improvement than existing work despite starting from a very competitive baseline CRF.

named entities. Other kinds of boundary detection errors involving multiple tokens are very rare. Our approach can also handle these errors by encouraging certain tokens to take different labels. This together with the clique features encoding the markovian dependency among neighbours can correct some multiple-token boundary detection errors.

7 Related Work

Recent work looking to directly model non-local dependencies and do approximate inference are that of Bunescu and Mooney (2004), who use a Relational Markov Network (RMN) (Taskar et al., 2002) to explicitly model long-distance dependencies, Sutton and McCallum (2004), who introduce skip-chain CRFs, which add additional non-local edges to the underlying CRF sequence model (which Bunescu and Mooney (2004) lack) and Finkel et al. (2005) who hand-set penalties for inconsistency in labels based on the training data and then use Gibbs Sampling for doing approximate inference where the goal is to obtain the label sequence that maximizes the product of the CRF objective function and their penalty. Unfortunately, in the RMN model, the dependencies must be defined in the model structure before doing any inference, and so the authors use heuristic part-of-speech patterns, and then add dependencies between these text spans using clique templates. This generates an extremely large number of overlapping candidate entities, which renders necessary additional templates to enforce the constraint that text subsequences cannot both be different entities, something that is more naturally modeled by a CRF. Another disadvantage of this approach is that it uses loopy belief propagation and a voted perceptron for approximate learning and inference, which are inherently unstable algorithms leading to convergence problems, as noted by the authors. In the skip-chain CRFs model, the decision of which nodes to connect is also made heuristically, and because the authors focus on named entity recognition, they chose to connect all pairs of identical capitalized words. They also utilize loopy belief propagation for approximate learning and inference. It is hard to directly extend their approach to model dependencies richer than those at the token level.

The approach of Finkel et al. (2005) makes it possible a to model a broader class of long-distance dependencies than Sutton and McCallum (2004), because they do not need to make any initial assumptions about which nodes should be connected and they too model dependencies between whole token sequences representing entities and between entity token sequences and their token supersequences that are entities. The disadvantage of their approach is the relatively ad-hoc selection of penalties and the high computational cost of running Gibbs sampling.

Early work in discriminative NER employed two stage approaches that are broadly similar to ours, but the effectiveness of this approach appears to have been overlooked in more recent work. Mikheev et al. (1999) exploit label consistency information within a document using relatively ad hoc multi-stage labeling procedures. Borth-

wick (1999) used a two-stage approach similar to ours with CMM's where *Reference Resolution* features which encoded the frequency of occurrences of other entities similar to the current token sequence, were derived from the output of the first stage. Malouf (2002) and Curran and Clark (2003) condition the label of a token at a particular position on the label of the most recent previous instance of that same token in a previous sentence of the same document. This violates the Markov property and therefore instead of finding the maximum likelihood sequence over the entire document (exact inference), they label one sentence at a time, which allows them to condition on the maximum likelihood sequence of previous sentences. While this approach is quite effective for enforcing label consistency in many NLP tasks, it permits a forward flow of information only, which can result in loss of valuable information. Chieu and Ng (2002) propose a solution to this problem: for each token, they define additional features based on known information, taken from other occurrences of the same token in the document. This approach has the advantage of allowing the training procedure to automatically learn good weights for these "global" features relative to the local ones. However, it is hard to extend this to incorporate other types of non-local structure.

8 Conclusion

We presented a two stage approach to model non-local dependencies and saw that it outperformed existing approaches to modeling non-local dependencies. Our approach also made it easy to exploit various dependencies across documents in the test corpus, whereas incorporating this information in most existing approaches would make them intractable due to the complexity of the resultant graphical model. Our simple approach is also very computationally efficient since the inference time is just twice the inference time of the basic sequential CRF, while for approaches doing approximate inference, the inference time is often well over an order of magnitude over the basic sequential CRF. The simplicity of our approach makes it easy to understand, implement, and adapt to new applications.

Acknowledgments

We wish to Jenny R. Finkel for discussions on NER and her CRF code. Also, thanks to Trond Grenager for NER discussions and to William Morgan for help with statistical significance tests. Also, thanks to Vignesh Ganapathy for helpful discussions and Rohini Rajaraman for comments on the writeup.

This work was supported in part by a Scottish Enterprise Edinburgh-Stanford Link grant (R37588), as part of the EASIE project.

References

A. Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. thesis, New York University.

R. Bunescu and R. J. Mooney. 2004. Collective information extraction with relational Markov networks. In *Proceedings of the 42nd ACL*, pages 439–446.

H. L. Chieu and H. T. Ng. 2002. Named entity recognition: a maximum entropy approach using global information. In *Proceedings of the 19th Coling*, pages 190–196.

J. R. Curran and S. Clark. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the 7th CoNLL*, pages 164–167.

J. Finkel, T. Grenager, and C. D. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 42nd ACL*.

D. Freitag and A. McCallum. 1999. Information extraction with HMMs and shrinkage. In *Proceedings of the AAAI-99 Workshop on Machine Learning for Information Extraction*.

S. Geman and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transitions on Pattern Analysis and Machine Intelligence*, 6:721–741.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th ICML*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

T. R. Leek. 1997. Information extraction using hidden Markov models. Master's thesis, U.C. San Diego.

R. Malouf. 2002. Markov models for language-independent named entity recognition. In *Proceedings of the 6th CoNLL*, pages 187–190.

A. McCallum, D. Freitag, and F. Pereira. 2000. Maximum entropy Markov models for information extraction and segmentation. In *Proceedings of the 17th ICML*, pages 591–598. Morgan Kaufmann, San Francisco, CA.

A. Mikheev, M. Moens, and C. Grover. 1999. Named entity recognition without gazetteers. In *Proceedings of the 9th EACL*, pages 1–8.

J. Pearl. 1988. Probabilistic reasoning in intelligent systems: Networks of plausible inference. In *Morgan Kauffmann*.

F. Sha and F. Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of NAACL-2003*, pages 134–141.

C. Sutton and A. McCallum. 2004. Collective segmentation and labeling of distant entities in information extraction. In *ICML Workshop on Statistical Relational Learning and Its connections to Other Fields*.

B. Taskar, P. Abbeel, and D. Koller. 2002. Discriminative probabilistic models for relational data. In *Proceedings of UAI-02*.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. 2000. Generalized belief propagation. In *Proceedings of NIPS-2000*, pages 689–695.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of COLING 2000*.