Dynamic Conditional Random Fields for Semantic Role Labeling

Joseph Frazee
Department of Linguistics
University of Texas at Austin
1 University Station B5100
Austin, TX 78712-0198 USA
jfrazee@mail.utexas.edu

Abstract

This paper presents a semantic role labeling (SRL) scheme using dynamic conditional random fields (DCRFs). The proposed system treats SRL as sequential tagging with dependencies between and among different SRL labeling decisions modeled as dependencies among labels in a factorized random field. This approach is motivated by previous research showing the usefulness of conditional random fields (CRFs) in NLP tasks and current interest in finding better ways to represent complex dependencies in SRL models. I show competitive results in argument identification and labeling on small numbers of training instances.

1 Introduction

Semantic role labeling (SRL), generally, is the NLP task of automatically assigning semantic roles to predicate arguments according to labeling schemes like PropBank (Kingsbury and Palmer, 2002, Palmer et al., 2005) and FrameNet (Baker et al., 1998). For example, in (1) a PropBank-oriented SRL system is expected to assign *they* the A0 subject label, *analysts* the A1 object label, and *announcing openings* the A2 oblique label for the predicate *pleased*. For the predicate *announcing* in (2), though, the system should assign the A0 label to *they*, the A1 label to *openings* and no labels to all remaining tokens.

(1) [**A0** They] **pleased** [**A1** analysts] by [**A2** announcing openings].

(2) [**A0** They] pleased analysts by **announcing** [**A1** openings].

In addition to patterns between predicates and the role labels they assign, it is also evident that more complex interactions constrain semantic role labeling decisions. Haghighi et al. (2005), Punyakanok et al. (2005), and others (Roth and Yih, 2005, Toutanova et al., 2005) have observed that role labeling decisions must respect complex constraints and dependencies such as the tendency for subjects to come before objects or the patterning of distinct predicates with distinct kinds and numbers of arguments. In the CoNLL-2005 shared task[1] (Màrquez and Carreras, 2005), systems that accounted for this sort of complexity were among the state of the art for SRL. The approaches demonstrated, however, have captured these constraints in joint models as discriminative re-ranking (Haghighi et al., 2005, Toutanova et al., 2005) or optimization (Punyakanok et al., 2005, Roth and Yih, 2005) problems because of the exponential space of joint labelings.

I apply a dynamic condition random field (DCRF) to capture the complex constraints of SRL; because whereas performing exact inference over an exponential space of all joint labelings is intractable, approximate inference over a large subset of joint labelings is not (Sutton et al., 2004, 2007). Further, this approach is more in the spirit of the machine learning paradigm because it relies on inferred rather than explicitly stated label dependencies, opening up the possibility for the model to encode hidden relationships between joint decisions.

In section 2, I introduce conditional random fields (CRFs) and motivate the use of dynamic conditional random fields in the extant task. Section 3 provides an overview of the proposed system architecture. I

[1] http://www.lsi.upc.edu/ srlconll/

describe experiments and results in section 4 and conclude in section 5.

2 Conditional Random Fields

2.1 Background and Motivations

Conditional random fields (Lafferty et al., 2001, Sutton and McCallum, 2006) are discriminative models with an undirected graphical structure belonging to the general class of graphical models (Jordan, 2004). CRFs are aimed at structured learning problems such as sequence, graph, and tree labeling which makes them apropos for labeling or segmenting natural language data. CRFs have been successfully applied to a wide-range of NLP tasks (Lafferty et al., 2001) including SRL (Roth and Yih, 2005, Cohn and Blunsom, 2005) often with state of the art or near state of the art results.

CRFs, being discriminative in nature, allow conditioning on arbitrarily large feature sets (Lafferty et al., 2001, Wallach, 2004). Accommodation of a large yet appropriate feature space has been shown to be a critical factor in the construction of SRL systems (Gildea and Jurafsky, 2002, Pradhan et al., 2004). Any model unable to cope with numerous feature functions would be severely hampered in the SRL domain.

Previous work on SRL has also indicated that the task is best represented as a series or cascade of multiple subtasks — feature extraction, tree pruning or relation selection, argument identification, argument labeling — and that each of these tasks can contribute positively or negatively to the overall performance of labeling decisions (Gildea and Jurafsky, 2002, Xue and Palmer, 2004, Pradhan et al., 2004). Sutton et al. (2004) show that a single DCRF model outperforms multiple labeling tasks joined via cascades in part-of-speech tagging and noun-phrase chunking. This paper considers whether a DCRF is viable as a model of simultaneous argument identification and argument labeling.

And because DCRFs forgo the first-order Markov assumption among labels such as is made by linear-chain and tree CRFs and encode Markov dependence between distinct factors in the random field, DCRFs can represent complex interactions among observations and predictions. Existing CRF proposals for SRL have either provided no representation (Cohn and Blunsom, 2005) or only an explicit representation (Roth and Yih, 2005) of labeling constraints. This paper addresses the insights of Haghighi et al. (2005), Punyakanok et al. (2005), and others (Roth and Yih, 2005, Toutanova et al., 2005) that SRL decisions must respect complex constraints and dependencies, but it does so by offering a model that can infer the constraints and dependencies of SRL and represent them in defeasible manner.

2.2 Model Representation

A typical linear-chain CRF defines the conditional probability of a label sequence y given an observation sequence x as in equation (3):

$$p(\mathsf{y}|\mathsf{x},\Lambda) = \frac{1}{Z(\mathsf{x})}\prod_{c\in C} exp(\sum_k \lambda_k f_k(\mathsf{y}_c,\mathsf{x}_c)) \quad (3)$$

where Λ is a set of parameters, $Z(\cdot)$ is a partition function, $f_k(\cdot)$ are feature functions defined on input, and the set C is the set of cliques or transitions defined over the graph.

A dynamic conditional random field, though, is a CRF with its parameters and variables tied or repeated over a sequence (Sutton et al., 2004, 2007).

$$p(\mathsf{y}|\mathsf{x},\Lambda) = \frac{1}{Z(\mathsf{x})}\prod_{t\in T}\prod_{c\in C} exp(\sum_k \lambda_k f_k(\mathsf{y}_{c,t},\mathsf{x})) \quad (4)$$

This is reflected in the probability distribution given in (4) by the label sequences $\mathsf{y}_{c,t}$ being tied across both cliques, C, and time-steps, T. For example, in figure (1) an abstract DCRF is shown with 2 linear chains for labels. Figures (2) and (3) show the model applied to examples (1) and (2).

3 System Architecture

The system consists of a supervised machine learning pipeline, using automatically tagged, tokenized, and dependency parsed English language data that: (1) extracts features, (2) generates a set of of relations under consideration for each predicate, and (3) trains and tests SRL labeling over the sequence of suggested relations. These steps are described in sections 3.1, 3.2, 3.3, and 3.4.

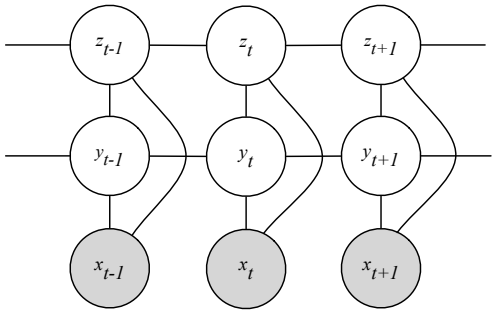

Figure 1: A graphical representation of a dynamic CRF with 2 linear chains for labels. The hidden variables are represented by clear circles and the observed variables are represented by shaded circles. In a labeling task z and y correspond to label predictions and x correspond to features of observed data such as words, parts of speech, etc.

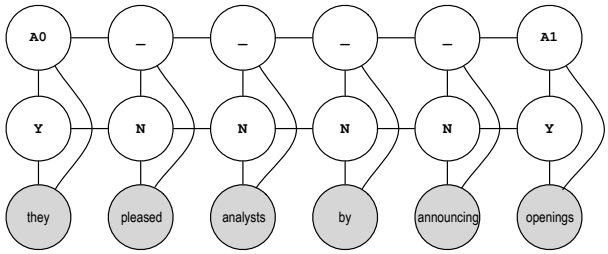

Figure 2: A dynamic CRF applied to a sequence of text for semantic role labeling for the predicate *announcing*. The top factor corresponds to role labeling decisions and the bottom factor corresponds to argument identification decisions. The shaded circles are observed data.

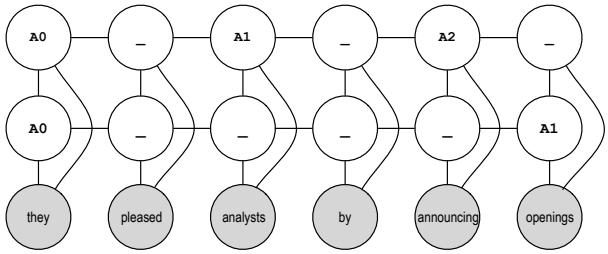

Figure 3: A dynamic CRF applied to a sequence of text for semantic role labeling for adjacent predicates. The top factor corresponds to role labeling decisions for the predicate *pleased*. The bottom factor corresponds to role labeling decisions for the predicate *announcing*. The shaded circles are observed data.

3.1 Feature Representation

For all tokens in the training and test instances, a set of features is extracted and then converted into binary features by combining the feature values and labels. The features are described as belonging to one of the following four categories.

Predicate features: *head word, lemma.*

Argument features: *dependent word, dependent POS, lemma.*

Argument context: *preceding and following words, preceding and following words' POS, is before predicate, is after predicate.*

Predicate-argument relational features: *relation to the predicate, path to the predicate*

This feature set is rather shallow as compared to previous approaches (Gildea and Jurafsky, 2002, Xue and Palmer, 2004, Pradhan et al., 2004) and excludes features such as predicate voice, subcategorization, temporal word flags, and word and path distances.

3.2 Relation Selection

Hacioglu (2004) defines a relation selection or dependency pruning algorithm that is able to reduce the data for SRL several-fold with a small percentage miss of target semantic role labels. Relation selection is critical in this model because it will make label sequences shorter so that adjacency is more predictive and soften the impact of the complexity of the inference algorithm used to decode the label sequences. Reducing sequence length without reducing the amount of substantive training material will make computation more feasible.

The relation selection algorithm reduces the set of dependency relation nodes under consideration by skipping all but the following while linearizing the dependency tree:

- *predicate*
- *predicate's parent*
- *predicate's children*
- *predicate's grand-children*
- *predicate's siblings*
- *predicate's siblings' children*
- *predicate's siblings' grand-children*

In addition, the set is further pruned by removing leaf nodes that never receive semantic role labels during training, e.g. nodes bearing determiner,

punctuation, and gap relations. Table 1 reports the impact of applying this procedure to the CoNLL-2008 shared task development data. A 3-fold decrease was achieved with little more than a 5% loss of predicate arguments.

3.3 Inference and Parameter Estimation

Because DCRFs have connections between and among labels and factors and between time-steps, the graph they define is "loopy" or cyclic. As such exact inference is intractable; however, Sutton et al. (2004, 2007) show that tree reparameterization (TRP), a form of loopy belief propagation, and L-BFGS parameter estimation can be used for approximate inference and parameter estimation, respectively, with negligible loss in accuracy.

Following Sutton et al. (2004), I use the belief propagation and parameter estimation schemes provided in the abstract CRF implementation of the GRMM (Sutton, 2006, McCallum, 2002) toolkit[2]. Using GRMM, a DCRF like that depicted in figure 1 can be constructed with the bulk of the effort focused on feature extraction and relation selection.

3.4 Sequential Tagging

CRFs have a graphical structure making them a natural choice for sequential labeling tasks. In a CRF for SRL, hidden variables correspond to labeling decisions and observed variables correspond to sets of features defined over input. Hidden nodes are connected, minimally, over time-steps.

A graphical representation of a DCRF model for SRL is given in figure 1. In figure 1, x_i represent features defined over input observations, z_i represent primary SRL labeling decisions such as the labels `A0` or `A1`, and y_i represent secondary SRL labeling decisions such as argument identification or adjacent predicate role label predictions. Figures 2 and 3 further illustrate the model showing multiple label predictions applied to the sequence of text in examples (1) and (2).

4 Experiments and Results

The model was trained and evaluated using the CoNLL-2008 shared task[3] (Surdeanu et al., 2008) training and development data, respectively. Gold standard dependency parses and predicate identification were used throughout. Three experiment setups were considered.

1. A DCRF was trained on the first 5,000 training instances using gold standard argument identification. Belief propagation was set to converge at a threshold of 0.001 or 1,000 iterations. A guassian prior ($\sigma^2 = 10.0$) was used to prevent overfitting. The model converged after 170 iterations, taking 100 hours of training time. This serves as a competitive benchmark for the remaining experiments which do not assume gold standard argument identification.
2. A DCRF was trained on the first 1,000 training instances using a simple relation selection algorithm that is known to over-prune. Arguments were predicted using a dedicated linear-chain in the DCRF. Belief propagation was set to converge at a threshold of 0.05 or 15 iterations. A gaussian prior ($\sigma^2 = 10.0$) was used to prevent overfitting. The model converged after 192 iterations, taking 80 hours of training time.
3. A DCRF was trained on the first 1,000 training instances using a relation selection algorithm with a small margin of error. Arguments were predicted using a dedicated linear-chain of the DCRF. Belief propagation was set to converge at a threshold of 0.05 or 15 iterations. A gaussian prior ($\sigma^2 = 10.0$) was used to prevent overfitting. Due to time, the model was evaluated after 40 iterations, taking 24 hours of training time.

Results for each experimental configuration are given in tables 2, 3, and 4. Experiment 1 results, $F_1 = 79.38$, indicate a sort of upper bound for this approach to semantic role labeling. Experiments 2 and 3 fall below this value at $F_1 = 67.48$ and $F_1 = 58.18$, respectively; however, in both cases, some experiment parameters — convergence threshold and belief propagation iterations — and the volume of training data were adjusted dramatically to favor smaller run times. Despite such handicaps, experiments 2 and 3 produce results above those of a comparable Maximum Entropy model trained on 5,000 training instances.

[2] http://mallet.cs.umass.edu

[3] http://www.yr-bcn.es/conll2008/

	Retained Relations	% Retained	Arg. Relations	Missed Args.	% Missed
ALL	202,813	100%	14,878	0	0%
SIMPLE	52,938	26.20%	13,333	1,545	10.38%
COMPLEX	66,567	30.82%	14,093	785	5.28%

Table 1: The size of the CoNLL-2008 shared task development set without relation selection (ALL), with relation section without the pruning of leaf nodes such as determiners (SIMPLE), and with relation selection with the pruning of leaf nodes such as determiners (COMPLEX).

Labels	P	R	F1
OVERALL	0.7907	0.7970	0.7938
CORE (A0–A5)	0.7875	0.8085	0.7979
ADJUNCT (AM-*)	0.8010	0.7585	0.7792
REFERENCE (R-*)	0.8209	0.7466	0.7820
CONTINUED (C-*)	0.8291	0.6736	0.7433

Table 2: Results for experiment 1 trained on 5,000 sentences using gold standard argument identification

Labels	P	R	F1
ARG. IDENT.	0.9247	0.9247	0.9247
OVERALL	0.7073	0.6452	0.6748
CORE (A0–A5)	0.6969	0.6598	0.6778
ADJUNCT (AM-*)	0.7511	0.6107	0.6107
REFERENCE (R-*)	0.8077	0.5701	0.6684
CONTINUED (C-*)	0.7627	0.3125	0.4434

Table 3: Results for experiment 2 trained on 1,000 sentences using SIMPLE relation selection and a dedicated factor for argument prediction

What is most striking in the experiments, though, is the accuracy of the argument identification predictions. Argument identification accuracy of 92% is close to the best results given in Xue and Palmer (2004) and Pradhan et al. (2004), but here the volume of training material is almost an order of magnitude lower. In view of the results of experiment 1 where argument identification was given, using more reasonable convergence thresholds and belief propagation iterations may indeed allow models like those used in experiments 2 and 3 to perform like that of experiment 1 due to the scant difference in argument identification accuracy between the models.

5 Conclusion

Dynamic conditional random fields are discriminative graphical models that dispense with first-order

Labels	P	R	F1
ARG. IDENT.	0.9193	0.9193	0.9193
OVERALL	0.6693	0.5146	0.5818
CORE (A0–A5)	0.6678	0.5333	0.5930
ADJUNCT (AM-*)	0.6704	0.4656	0.5495
REFERENCE (R-*)	0.7423	0.4449	0.5563
CONTINUED (C-*)	0.5000	0.0145	0.0282

Table 4: Results for experiment 3 trained on 1,000 sentences using COMPLEX relation selection and a dedicated factor for argument prediction

Markov independence assumptions and allow for simultaneous or joint predictions. I have proposed a DCRF semantic role labeling scheme that addresses the issue of complex dependencies and constraints in SRL labeling decisions. The model was evaluated against the CoNLL-2008 shared task development data showing promising results for simultaneous semantic role labeling and argument identification. The system best score for argument labeling and argument identification is $F_1 = 67.48$ and $F_1 = 92.47$, respectively.

Acknowledgments

I would like to thank my colleagues in the Spring 2008 Automatic Syntactic and Semantic Analysis seminar at the University of Texas at Austin for thoughtful feedback and discussion.

References

Collin Baker, Charles Fillmore, and John Lowe. The Berkeley FrameNet project. In *Proceedings of ACL/COLING-1998*, 1998.

Trevor Cohn and Philip Blunsom. Semantic role labeling with tree conditional random fields. In *Proceedings of CoNLL-2005*, 2005.

Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. In *Proceedings of ACL-2002*, 2002.

Kadri Hacioglu. Semantic role labeling using dependency trees. In *Proceedings of COLING-2004*, 2004.

Aria Haghighi, Kristina Toutanova, and Christopher Manning. A joint model for semantic role labeling. In *Proceedings of CoNLL-2005*, 2005.

Michael I. Jordan. Graphical models. *Statistical Science (Special Issue on Bayesian Statistics)*, 19: 140–155, 2004.

Paul Kingsbury and Martha Palmer. From treebank to propbank. In *Proceedings of LREC-2002*, 2002.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings ICML-2001*, 2001.

Lluís Màrquez and Xavier Carreras. Introduction to the conll-2005 shared task. In *Proceedings of CoNLL-2005*, 2005. http://www.lsi.upc.edu/ srlconll/.

Andrew McCallum. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1), 2005.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James Martin, and Dan Jurafsky. Shallow semantic parsing using support vector machines. In *Proceedings of HLT/NAACL-2004*, 2004.

Vasin Punyakanok, Peter Koomen, Dan Roth, and Wen-tau Yih. Generalized inference with multiple semantic role labeling systems. In *Proceedings of CoNLL-2005*, 2005.

Dan Roth and Wen-tau Yih. Integer linear programming inference for conditional random fields. In *Proceedings of ICML-2005*, 2005.

Mihai Surdeanu, Richard Johansson, Lluís Màrquez, Adam Meyers, and Joakim Nivre. Conll-2008 shared task. http://www.yr-bcn.es/conll2008/, 2008.

Charles Sutton. GRMM: A graphical models toolkit. http://mallet.cs.umass.edu, 2006.

Charles Sutton and Andrew McCallum. *An Introduction to Conditional Random Fields for Relational Learning*. MIT Press, 2006.

Charles Sutton, Khashayar Rohanimanesh, and Andrew McCallum. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. In *Proceedings of ICML-2004*, 2004.

Charles Sutton, Andrew McCallum, and Khashayar Rohanimanesh. *Journal of Machine Learning Research*, 8:693–723, 2007.

Kristina Toutanova, Aria Haghighi, , and Christopher D. Manning. Joint learning improves semantic role labeling. In *Proceedings of ACL-2005*, 2005.

Hanna M. Wallach. Conditional random fields: An introduction. Technical Report MS-CIS-04-21, Department of Computer and Information Science, University of Pennsylvania, 2004.

Nianwen Xue and Martha Palmer. Calibrating features for semantic role labeling. In *Proceedings of EMNLP-2004*, 2004.