A Joint Model for Extended Semantic Role Labeling

Vivek Srikumar and Dan Roth
University of Illinois, Urbana-Champaign
Urbana, IL 61801
{vsrikum2, danr}@illinois.edu

Abstract

This paper presents a model that extends semantic role labeling. Existing approaches independently analyze relations expressed by verb predicates or those expressed as nominalizations. However, sentences express relations via other linguistic phenomena as well. Furthermore, these phenomena interact with each other, thus restricting the structures they articulate. In this paper, we use this intuition to define a joint inference model that captures the inter-dependencies between verb semantic role labeling and relations expressed using prepositions. The scarcity of jointly labeled data presents a crucial technical challenge for learning a joint model. The key strength of our model is that we use existing structure predictors as black boxes. By enforcing consistency constraints between their predictions, we show improvements in the performance of both tasks without retraining the individual models.

1 Introduction

The identification of semantic relations between sentence constituents has been an important task in NLP research. It finds applications in various natural language understanding tasks that require complex inference going beyond the surface representation. In the literature, semantic role extraction has been studied mostly in the context of verb predicates, using the Propbank annotation of Palmer et al. (2005), and also for nominal predicates, using the Nombank corpus of Meyers et al. (2004).

However, sentences express semantic relations through other linguistic phenomena. For example, consider the following sentence:

(1) The field goal by Brien changed the game in the fourth quarter.

Verb centered semantic role labeling would identify the arguments of the predicate *change* as (a) *The field goal by Brien* (A0, the causer of the change), (b) *the game* (A1, the thing changing), and (c) *in the fourth quarter* (temporal modifier). However, this does not tell us that the scorer of the field goal was Brien, which is expressed by the preposition *by*. Also, note that the *in* indicates a temporal relation, which overlaps with the verb's analysis.

In this paper, we propose an extension of the standard semantic role labeling task to include relations expressed by lexical items other than verbs and nominalizations. Further, we argue that there are interactions between the different phenomena which suggest that there is a benefit in studying them together. However, one key challenge is that large jointly labeled corpora do not exist. This motivates the need for novel learning and inference schemes that address the data problem and can still benefit from the interactions among the phenomena.

This paper has two main contributions.

1. From the machine learning standpoint, we propose a joint inference scheme to combine *existing* structure predictors for multiple linguistic phenomena. We do so using hard constraints that involve only the labels of the phenomena. The strength of our model is that it is easily

Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 129–139,
Edinburgh, Scotland, UK, July 27–31, 2011. ©2011 Association for Computational Linguistics

extensible, since adding new phenomena does not require fully retraining the joint model from scratch. Furthermore, our approach minimizes the need for extensive jointly labeled corpora and, instead, uses existing predictors as black boxes.

2. From an NLP perspective, we motivate the extension of semantic role labeling beyond verbs and nominalizations. We instantiate our joint model for the case of extracting preposition and verb relations together. Our model uses existing systems that identify verb semantic roles and preposition object roles and jointly predicts the output of the two systems in the presence of linguistic constraints that enforce coherence between the predictions. We show that using constraints to combine models improves the performance on both tasks. Furthermore, since the constraints depend only on the labels of the two tasks and not on any specific dataset, our experiments also demonstrate that enforcing them allows for better domain adaptation.

The rest of the paper is organized as follows: We motivate the need for extending semantic role labeling and the necessity for joint inference in Section 2. In Section 3, we describe the component verb SRL and preposition role systems. The global model is defined in Section 4. Section 5 provides details on the coherence constraints we use and demonstrates the effectiveness of the joint model through experiments. Section 6 discusses our approach in comparison to existing work and Section 7 provides concluding remarks.

2 Problem Definition and Motivation

Semantic Role Labeling has been extensively studied in the context of verbs and nominalizations. While this analysis is crucial to understanding a sentence, it is clear that in many natural language sentences, information is conveyed via other lexical items. Consider, for example, the following sentences:

(2) Einstein's theory of relativity changed physics.

(3) The plays of Shakespeare are widely read.

(4) The bus, which was heading for Nairobi in Kenya, crashed in the Kabale district of Uganda.

The examples contain information that cannot be captured by analyzing the verbs and the nominalizations. In sentence (2), the possessive form tells us that the *theory of relativity* was discovered by *Einstein*. Furthermore, the *theory* is on the subject of *relativity*. The usage of the preposition *of* is different in sentence (3), where it indicates a creator-creation relationship. In the last sentence, the same preposition tells us that the Kabale district is located in Uganda. Prepositions, compound nouns, possessives, adjectival forms and punctuation marks often express relations, the identification of which is crucial for text understanding tasks like recognizing textual entailment, paraphrasing and question answering.

The relations expressed by different linguistic phenomena often overlap. For example, consider the following sentence:

(5) Construction of the library began in 1968.

The relation expressed by the nominalization *construction* recognizes *the library* as the argument of the predicate *construct*. However, the same analysis can also be obtained by identifying the sense of the preposition *of*, which tells us that the subject of the preposition is a nominalization of the underlying verb. A similar redundancy can be observed with analyses of the verb *began* and the preposition *in*. The above example motivates the following key intuition: *The correct interpretation of a sentence is the one that gives a consistent analysis across all the linguistic phenomena expressed in it.*

An inference mechanism that simultaneously predicts the structure for different phenomena should account for consistency between the phenomena. A model designed to address this has the following desiderata:

1. It should account for the dependencies between phenomena.

2. It should be extensible to allow easy addition of new linguistic phenomena.

3. It should be able to leverage existing state-of-the-art models with minimal use of jointly labeled data, which is expensive to obtain.

Systems that are trained on each task independently do not account for the interplay between them. One approach for tackling this is to define pipelines, where the predictions for one of the tasks acts as the input for another. However, a pipeline does not capture the two-way dependency between the tasks. Training a fully joint model from scratch is also unrealistic because it requires text that is annotated with all the tasks, thus making joint training implausible from a learning theoretic perspective (See Punyakanok et al. (2005) for a discussion about the learning theoretic requirements of joint training.)

3 Tasks and Individual Systems

Before defining our proposed model that captures the requirements listed in the previous section, we introduce the tasks we consider and their independently trained systems that we improve using the joint system. Though the model proposed here is general and can be extended to several linguistic phenomena, in this paper, we focus on relations expressed by verbs and prepositions. This section describes the tasks, the data sets we used for our experiments and the current state-of-the-art systems for these tasks.

We use the following sentence as our running example to illustrate the phenomena: *The company calculated the price trends on the major stock markets on Monday.*

3.1 Preposition Relations

Prepositions indicate a relation between the attachment point of the preposition and its object. As we have seen, the same preposition can indicate different types of relations. In the literature, the polysemy of prepositions is addressed by The Preposition Project[1] of Litkowski and Hargraves (2005), which is a large lexical resource for English that labels prepositions with their sense. This sense inventory formed the basis of the SemEval-2007 task of preposition word sense disambiguation of Litkowski and Hargraves (2007). In our example, the first *on* would be labeled with the sense **8(3)** which identifies the object of the preposition as the topic, while the second instance would be labeled as **17(8)**, which indicates that argument is the day of the occurrence.

The preposition sense inventory, while useful to identify the fine grained distinctions between preposition usage, defines a unique sense label for each preposition by indexing the definitions of the prepositions in the Oxford Dictionary of English. For example, in the phrase *at noon*, the at would be labeled with the sense **2(2)**, while the preposition in *I will see you in an hour* will be labeled **4(3)**. Note that both these (and also the second *on* in our running example) indicate a temporal relation, but are assigned different labels based on the preposition. To counter this problem we collapsed preposition senses that are semantically similar to define a new label space, which we refer to as *Preposition Roles*.

We retrained classifiers for preposition sense for the new label space. Before describing the preposition role dataset, we briefly describe the datasets and the features for the sense problem. The best performing system at the SemEval-2007 shared task of preposition sense disambiguation (Ye and Baldwin (2007)) achieves a mean precision of 69.3% for predicting the fine grained senses. Tratz and Hovy (2009) and Hovy et al. (2010) attained significant improvements in performance using features derived from the preposition's neighbors in the parse tree. We extended the feature set defined in the former for our independent system. Table 1 summarizes the rules for identifying the syntactically related words for each preposition. We used dependencies from the easy-first dependency parser of Goldberg and Elhadad (2010).

For each word extracted from these rules, the features include the word itself, its lemma, the POS tag, synonyms and hypernyms of the first WordNet sense and an indicator for capitalization. These features improved the accuracy of sense identification to 75.1% on the SemEval test set. In addition, we also added the following new features for each word:

1. Indicators for gerunds and nominalizations of verbs.

2. The named entity tag (Person, Location or Organization) associated with a word, if any. We

[1] http://www.clres.com/prepositions.html

Id.	Feature
1.	Head noun/verb that dominates the preposition along with its modifiers
2.	Head noun/verb that is dominated by the preposition along with its modifiers
3.	Subject, negator and object(s) of the immediately dominating verb
4.	Heads of sibling prepositions
5.	Words withing a window of 5 centered at the preposition

Table 1: Features for preposition relation from Tratz and Hovy (2009). These rules were used to identify syntactically related words for each preposition.

used the state-of-the-art named entity tagger of Ratinov and Roth (2009) to label the text.

3. Gazetteer features, which are active if a word is a part of a phrase that belongs to a gazetteer list. We used the gazetteer lists which were used by the NER system. We also used the CBC word clusters of Pantel and Lin (2002) as additional gazetteers and Brown cluster features as used by Ratinov and Roth (2009) and Koo et al. (2008).

Dahlmeier et al. (2009) annotated senses for the prepositions *at, for, in, of, on, to* and *with* in the sections 2-4 and 23 of the Wall Street Journal portion of the Penn Treebank[2]. We trained sense classifiers on both datasets using the Averaged Perceptron algorithm with the one-vs-all scheme using the Learning Based Java framework of Rizzolo and Roth (2010)[3]. Table 2 reports the performance of our sense disambiguation systems for the Treebank prepositions.

As mentioned earlier, we collapsed the sense labels onto the newly defined preposition role labels. Table 3 shows this label set along with frequencies of the labels in the Treebank dataset. According to this labeling scheme, the first *on* in our running example will be labeled TOPIC and the second one will be labeled TEMPORAL[4]. We re-trained the sense disambiguation system to predict preposition roles. When trained on the Treebank data, our system attains an accuracy of 67.82% on Section 23 of the Treebank. We use this system as our independent baseline for preposition role identification.

Train	Test set	
	Treebank Sec. 23	SemEval
Penn Treebank	61.41	38.22
SemEval	47.00	78.25

Table 2: Preposition sense performance. This table reports accuracy of sense prediction on the prepositions that have been annotated for the Penn Treebank dataset.

Role	Train	Test
ACTIVITY	57	23
ATTRIBUTE	119	51
BENEFICIARY	78	17
CAUSE	255	116
CONCOMITANT	156	74
ENDCONDITION	88	66
EXPERIENCER	88	42
INSTRUMENT	37	19
LOCATION	1141	414
MEDIUMOFCOMMUNICATION	39	30
NUMERIC/LEVEL	301	174
OBJECTOFVERB	365	112
OTHER	65	49
PARTWHOLE	485	133
PARTICIPANT/ACCOMPANIER	122	58
PHYSICALSUPPORT	32	18
POSSESSOR	195	56
PROFESSIONALASPECT	24	10
RECIPIENT	150	70
SPECIES	240	58
TEMPORAL	582	270
TOPIC	148	54

Table 3: Preposition role data statistics for the Penn Treebank preposition dataset.

3.2 Verb SRL

The goal of verb Semantic Role Labeling (SRL) is to identify the predicate-argument structure defined by verbs in sentences. The CoNLL Shared Tasks of 2004 and 2005 (See Carreras and Màrquez

[2]This dataset does not annotate all prepositions and restricts itself mainly to prepositions that start a Propbank argument. The data is available at `http://nlp.comp.nus.edu.sg/corpora`

[3]Learning Based Java can be downloaded from `http://cogcomp.cs.illinois.edu`.

[4]The mapping from the preposition senses to the roles defines a new dataset and is available for download at `http://cogcomp.cs.illinois.edu/`.

(2004), Carreras and Màrquez (2005)) studied the identification of the predicate-argument structure of verbs using the PropBank corpus of Palmer et al. (2005). Punyakanok et al. (2008) and Toutanova et al. (2008) used global inference to ensure that the predictions across all arguments of the same predicate are coherent. We re-implemented the system of Punyakanok et al. (2008), which we briefly describe here, to serve as our baseline verb semantic role labeler [5]. We refer the reader to the original paper for further details.

The verb SRL system of Punyakanok et al. (2008) consists of four stages – candidate generation, argument identification, argument classification and inference. The candidate generation stage involves using the heuristic of Xue and Palmer (2004) to generate an over-complete set of argument candidates for each predicate. The identification stage uses a classifier to prune the candidates. In the argument classification step, the candidates that remain after the identification step are assigned scores for the SRL arguments using a multiclass classifier. One of the labels of the classifier is $\emptyset$, which indicates that the candidate is, in fact, not an argument. The inference step produces a combined prediction for all argument candidates of a verb proposition by enforcing global constraints.

The inference enforces the following structural and linguistic constraints: (1) Each candidate can have at most one label. (2) No duplicate core arguments. (3) No overlapping or embedding arguments. (4) Given the predicate, some argument classes are illegal. (5) If a candidate is labeled as an R-arg, then there should be one labeled as arg. (6) If a candidate is labeled as a C-arg, there should be one labeled arg that occurs *before* the C-arg.

Instead of using the identifier to filter candidates for the classifier, in our SRL system, we added the identifier to the global inference and enforced consistency constraints between the identifier and the argument classifier predictions – the identifier should predict that a candidate is an argument if, and only if, the argument classifier does not predict the label $\emptyset$. This change is in keeping with the idea of using joint inference to combine independently learned systems, in this case, the argument identifier and the role classifier. Furthermore, we do not need to explicitly tune the identifier for high recall.

We phrase the inference task as an integer linear program (ILP) following the approach developed in Roth and Yih (2004). Integer linear programs were used by Roth and Yih (2005) to add general constraints for inference with conditional random fields. ILPs have since been used successfully in many NLP applications involving complex structures – Punyakanok et al. (2008) for semantic role labeling, Riedel and Clarke (2006) and Martins et al. (2009) for dependency parsing and several others[6].

Let $v^C_{i,a}$ be the Boolean indicator variable that denotes that the i^{th} argument candidate for a predicate is assigned a label a and let $\Theta^C_{i,a}$ represent the score assigned by the argument classifier for this decision. Similarly, let v^I_i denote the identifier decision for the i^{th} argument candidate of the predicate and Θ^I_i denote its identifier score. Then, the objective of inference is to maximize the total score of the assignment

$$\max_{\mathbf{v}^C,\mathbf{v}^I} \sum_{i,a} \Theta^C_{i,a} v^C_{i,a} + \sum_i \Theta^I_i v^I_i \qquad (1)$$

Here, $\mathbf{v}^C$ and $\mathbf{v}^I$ denote all the argument classifier and identifier variables respectively. This maximization is subject to the constraints described above, which can be transformed to linear (in)equalities. We denote these constraints as $\mathcal{C}^{SRL}$. In addition to $\mathcal{C}^{SRL}$ which were defined by Punyakanok et al. (2008), we also have the constraints linking the predictions of the identifier and classifier:

$$v^C_{v,i,\emptyset} + v^I_{v,i} = 1; \;\; \forall v, i. \qquad (2)$$

Inference in our baseline SRL system is, thus, the maximization of the objective defined in (1) subject to constraints $\mathcal{C}^{SRL}$, the identifier-classifier constraints defined in (2) and the restriction of the variables to take values in $\{0, 1\}$.

To train the classifiers, we used parse trees from the Charniak and Johnson (2005) parser with the

[5]The verb SRL system be downloaded from `http://cogcomp.cs.illinois.edu/page/software`

[6]The primary advantage of using ILP for inference is that this representation enables us to add arbitrary coherence constraints between the phenomena. If the underlying optimization problem itself is tractable, then so is the corresponding integer program. However, other approaches to solve the constrained maximization problem can also be used for inference.

same feature representation as in the original system. We trained the classifiers on the standard Propbank training set using the one-vs-all extension of the average Perceptron algorithm. As with the preposition roles, we implemented our system using Learning Based Java of Rizzolo and Roth (2010). We normalized all classifier scores using the softmax function. Compared to the 76.29% F1 score reported by Punyakanok et al. (2008) using single parse tree predictions from the parser, our system obtained 76.22% F1 score on section 23 of the Penn Treebank.

4 A Joint Model for Verbs and Prepositions

We now introduce our model that captures the needs identified in Section 2. The approach we develop in this paper follows the one proposed by Roth and Yih (2004) of training individual models and combining them at inference time. Our joint model is a Constrained Conditional Model (See Chang et al. (2011)), which allows us to build upon existing learned models using declarative constraints.

We represent our component inference problems as integer linear program instances. As we saw in Section 3.2, the inference for SRL is instantiated as an ILP problem. The problem of predicting preposition roles can be easily transformed into an ILP instance. Let $v^R_{p,r}$ denote the decision variable that encodes the prediction that the preposition p is assigned a role r and let $\Theta^R_{p,r}$ denote its score. Let $\mathbf{v}^R$ denote all the role variables for a sentence. Then role prediction is equivalent to the following maximization problem:

$$\max_{\mathbf{v}^R} \quad \sum_{p,r} \Theta^R_{p,r} \cdot v^R_{p,r} \tag{3}$$

$$\text{subj. to} \quad \sum_{r} v^R_{p,r} = 1, \quad \forall p \tag{4}$$

$$v^R_{p,r} \in \{0,1\}, \quad \forall p, r. \tag{5}$$

In general, let p denote a linguistic structure prediction task of interest and let $\mathcal{P}$ denote all such tasks. Let Z^p denote the set of labels that the parts of the structure associated with phenomenon p can take. For example, for the SRL argument classification component, the parts of the structure are all the candidates that need to be labeled for a given sentence and the set Z^p is the set of all argument labels. For each phenomenon $p \in \mathcal{P}$, we use $\mathbf{v}^p$ to denote its set of inference variables for a given sentence. Each inference variable $v^p_{Z,y} \in \mathbf{v}^p$ corresponds to the prediction that the part y has the label Z in the final structure. Each variable is associated with a score $\Theta^p_{Z,y}$ that is obtained from a learned score predictor. Let $\mathcal{C}^p$ denote the structural constraints that are "local" to the phenomenon. Thus, for verb SRL, these would be the constraints defined in the previous section, and for preposition role, the only local constraint would be the constraint (4) defined above.

The independent inference problem for the phenomenon p is the following integer program:

$$\max_{\mathbf{v}^p} \quad \sum_{Z \in Z^p} \sum_{v^p} v^p_{Z,y} \cdot \Theta^p_{Z,y}, \tag{6}$$

$$\text{subj. to} \quad \mathcal{C}^p(\mathbf{v}^p), \tag{7}$$

$$v^p_{Z,y} \in \{0,1\}, \qquad \forall v^p_{Z,y}. \tag{8}$$

As a technical point, this defines one inference problem per sentence, rather than per predicate as in the verb SRL system of Punyakanok et al. (2008). This simple extension enabled Surdeanu et al. (2007) to study the impact of incorporating cross-predicate constraints for verb SRL. In this work, this extension allows us to incorporate cross-phenomena inference.

4.1 Joint inference

We consider the problem of jointly predicting several phenomena incorporating linguistic knowledge that enforce consistency between the output labels. Suppose p_1 and p_2 are two phenomena. If $z^{p_1}_1$ is a label associated with the former and $z^{p_2}_1, z^{p_2}_2, \cdots$ are labels associated with the latter, we consider constraints of the form

$$z^{p_1}_1 \rightarrow z^{p_2}_1 \vee z^{p_2}_2 \vee \cdots \vee z^{p_2}_n \tag{9}$$

We expand this language of constraints by allowing the specification of pre-conditions for a constraint to apply. This allows us to enforce constraints of the form *"If an argument that starts with the preposition 'at' is labeled* AM-TMP*, then the preposition can be labeled either* NUMERIC/LEVEL *or* TEMPORAL." This constraint is universally quantified for

all arguments that satisfy the precondition of starting with the preposition *at*.

Given a first-order constraint in this form and an input sentence, suppose the inference variable $v_1^{p_1}$ is a grounding of $z_1^{p_1}$ and $v_1^{p_2}, v_2^{p_2}, \cdots$ are groundings of the right hand labels such that the preconditions are satisfied, then the constraint can be phrased as the following linear inequality.

$$-v_1^{p_1} + \sum_i v_i^{p_2} \geq 0$$

In the context of the preposition role and verb SRL, we consider constraints between labels for a preposition and SRL argument candidates that begin with that preposition. This restriction forms the precondition for all the joint constraints considered in this paper. Since the joint constraints involve only the labels, they can be derived either manually from the definition of the tasks or using statistical relation learning techniques. In addition to mining constraints of the form (9), we also use manually specified joint constraints. The constraints used in our experiments are described further in Section 5.

In general, let J denote a set of pairwise joint constraints. The joint inference problem can be phrased as that of maximizing the score of the assignment subject to the structural constraints of each phenomenon ($\mathcal{C}^p$) *and* the joint linguistic constraints (J). However, since, the individual tasks were not trained on the same datasets, the scoring functions need not be in the same numeric scale. In our model, each label Z for a phenomenon p is associated with a scoring function $\Theta^p_{Z,y}$ for a part y. To scale the scoring functions, we associate each label with a parameter λ^p_Z. This gives us the following integer linear program for joint inference:

$$\max_{\mathbf{v}} \quad \sum_{p \in \mathcal{P}} \sum_{Z \in Z^p} \lambda^p_Z \left(\sum_{y^p} v^p_{Z,y} \cdot \Theta^p_{Z,y} \right), \quad (10)$$

$$\text{subj. to} \quad \mathcal{C}^p(\mathbf{v}^p), \quad \forall p \in \mathcal{P} \quad (11)$$

$$J(\mathbf{v}), \quad (12)$$

$$v^p_{Z,y} \in \{0,1\}, \quad \forall v^p_{Z,y}. \quad (13)$$

Here, $\mathbf{v}$ is the vector of inference variables which is obtained by stacking all the inference variables of each phenomena.

For our experiments, we use a cutting plane solver to solve the integer linear program as in Riedel (2009). This allows us to solve the inference problem without explicitly having to instantiate all the joint constraints.

4.2 Learning to rescale the individual systems

Given the individual models and the constraints, we only need to learn the scaling parameters λ^p_Z. Note that the number of scaling parameters is the total number of labels. When we jointly predict verb SRL and preposition role, we have 22 preposition roles (from table 3), one SRL identifier label and 54 SRL argument classifier labels. Thus we learn only 77 parameters for our joint model. This means that we only need a very small dataset that is jointly annotated with all the phenomena.

We use the Structure Perceptron of Collins (2002) to learn the scaling weights. Note that for learning the scaling weights, we need each label to be associated with a real-valued feature. Given an assignment of the inference variables $\mathbf{v}$, the value of the feature corresponding to the label Z of task p is given by the sum of scores of all parts in the structure for p that have been assigned this label, i.e. $\sum_{y^p} v^p_{Z,y} \cdot \Theta^p_{Z,y}$. This feature is computed for the gold and the predicted structures and is used for updating the weights.

5 Experiments

In this section, we describe our experimental setup and evaluate the performance of our approach. The research question addressed by the experiments is the following: *Given independently trained systems for verb SRL and preposition roles, can their performance be improved using joint inference between the two tasks?* To address this, we report the results of the following two experiments:

1. First, we compare the joint system against the baseline systems and with pipelines in both directions. In this setting, both base systems are trained on the Penn Treebank data.

2. Second, we show that using joint inference can provide strong a performance gain even when the underlying systems are trained on different domains.

In all experiments, we report the F1 measure for the verb SRL performance using the CoNLL 2005

evaluation metric and the accuracy for the preposition role labeling task.

5.1 Data and Constraints

For both the verb SRL and preposition roles, we used the first 500 sentences of section 2 of the Penn Treebank corpus to train our scaling parameters. For the first set of experiments, we trained our underlying systems on the rest of the available Penn Treebank training data for each task. For the adaptation experiment, we train the role classifier on the SemEval data (restricted to the same Treebank prepositions). In both cases, we report performance on section 23 of the Treebank.

We mined consistency constraints from the sections 2, 3 and 4 of the Treebank data. As mentioned in Section 4.1, we considered joint constraints relating preposition roles to verb argument candidates that start with the preposition. We identified the following types of constraints: (1) For each preposition, the set of invalid verb arguments and preposition roles. (2) For each preposition role, the set of allowed verb argument labels if the role occurred more than ten times in the data, and (3) For each verb argument, the set of allowed preposition roles, similarly with a support of ten. Note that, while the constraints were obtained from jointly labeled data, the constraints could be written down because they encode linguistic intuition about the labels.

The following is a constraint extracted from the data, which applies to the preposition *with*:

$$
\begin{aligned}
\text{srlarg(A2)} \quad &\rightarrow \quad \text{prep-role(ATTRIBUTE)} \\
&\vee \quad \text{prep-role(CAUSE)} \\
&\vee \quad \text{prep-role(INSTRUMENT)} \\
&\vee \quad \text{prep-role(OBJECTOFVERB)} \\
&\vee \quad \text{prep-role(PARTWHOLE)} \\
&\vee \quad \text{prep-role(PARTICIPANT/ACCOMPAINER)} \\
&\vee \quad \text{prep-role(PROFESSIONALASPECT).}
\end{aligned}
$$

This constraint says that if any candidate that starts with *with* is labeled as an A2, then the preposition can be labeled only with one of the roles on the right hand side.

Some of the mined constraints have negated variables to enforce that a role or an argument label should not be allowed. These can be similarly converted to linear inequalities. See Rizzolo and Roth (2010) for a further discussion about converting logical expressions into linear constraints.

In addition to these constraints that were mined from data, we also enforce the following handwritten constraints: (1) If the role of a verb attached preposition is labeled TEMPORAL, then there should be a verb predicate for which this prepositional phrase is labeled AM-TMP. (2) For verb attached prepositions, if the preposition is labeled with one of ACTIVITY, ENDCONDITION, INSTRUMENT or PROFESSIONALASPECT, there should be at least one predicate for which the corresponding prepositional phrase is not labeled $\emptyset$.

The conversion of the first constraint to a linear inequality is similar to the earlier cases. For each of the roles in the second constraint, let r denote a role variable that assigns the label to some preposition. Suppose there are n SRL candidates across all verb predicates begin with that preposition, and let $s_1, s_2, \cdots, s_n$ denote the SRL variables that assign these candidates to the label $\emptyset$. Then the second constraint corresponds to the following inequality:

$$r + \sum_{i=1}^{n} s_i \leq n$$

5.2 Results of joint learning

First, we compare our approach to the performance of the baseline independent systems and to pipelines in both directions in Table 4. For one pipeline, we added the prediction of the baseline preposition role system as an additional feature to both the identifier and the argument classifier for argument candidates that start with a preposition. Similarly, for the second pipeline, we added the SRL predictions as features for prepositions that were the first word of an SRL argument. In all cases, we performed five-fold cross validation to train the classifiers.

The results show that both pipelines improve performance. This justifies the need for a joint system because the pipeline can improve only one of the tasks. The last line of the table shows that the joint inference system improves upon both the baselines. We achieve this improvement without retraining the underlying models, as done in the case of the pipelines.

On analyzing the output of the systems, we found that the SRL precision improved by 2.75% but the

Setting	SRL (F1)	Preposition Role (Accuracy)
Baseline SRL	76.22	–
Baseline Prep.	–	67.82
Prep. → SRL	**76.84**	–
SRL → Prep.	–	**68.55**
Joint inference	**77.07**	**68.39**

Table 4: Performance of the joint system, compared to the individual systems and the pipelines. All performance measures are reported on Section 23 of the Penn Treebank. The verb SRL systems were trained on sections 2-21, while the preposition role classifiers were trained on sections 2-4. For the joint inference system, the scaling parameters were trained on the first 500 sentences of section 2, which were held out. All the improvements in this table are statistically significant at the 0.05 level.

recall decreased by 0.98%, contributing to the overall F1 improvement. The decrease in recall is due to the joint hard constraints that prohibit certain assignments to the variables which would have otherwise been possible. Note that, for a given sentence, even if the joint constraints affect only a few argument candidates directly, they can alter the labels of the other candidates via the "local" SRL constraints.

Consider the following example of the system output which highlights the effect of the constraints.

(6) Weatherford said market conditions led to the cancellation of the planned exchange.

The independent preposition role system incorrectly identifies the *to* as a LOCATION. The semantic role labeling component identifies the phrase *to the cancellation of the planned exchange* as the A2 of the verb *led*. One of the constraints mined from the data prohibits the label LOCATION for the preposition *to* if the argument it starts is labeled A2. This forces the system to change the preposition label to the correct one, namely ENDCONDITION. Both the independent and the joint systems also label the preposition *of* as OBJECTOFVERB, which indicates that the phrase *the planned exchange* is the object of the deverbal noun *cancellation*.

5.3 Effect of constraints on adaptation

Our second experiment compares the performance of the preposition role classifier that has been trained on the SemEval dataset with and without joint constraints. Note that Table 2 in Section 3, shows the drop in performance when applying the preposition sense classifier. We see that the SemEval-trained preposition role classifier (baseline in the table) achieves an accuracy of 53.29% when tested on the Treebank dataset. Using this classifier jointly with the verb SRL classifier via joint constraints gets an improvement of almost 3 percent in accuracy.

Setting	Preposition Role (Accuracy)
Baseline	53.29
Joint inference	**56.22**

Table 5: Performance of the SemEval-trained preposition role classifier, when tested on the Treebank dataset with and without joint inference with the verb SRL system. The improvement, in this case is statistically significant at the 0.01 level using the sign test.

The primary reason for this improvement, even without re-training the classifier, is that the constraints are defined using only the labels of the systems. This avoids the standard adaptation problems of differing vocabularies and unseen features.

6 Discussion and Related work

Roth and Yih (2004) formulated the problem of extracting entities and relations as an integer linear program, allowing them to use global structural constraints at inference time even though the component classifiers were trained independently. In this paper, we use this idea to combine classifiers that were trained for two different tasks on different datasets using constraints to encode linguistic knowledge.

In the recent years, we have seen several joint models that combine two or more NLP tasks . Andrew et al. (2004) studied verb subcategorization and sense disambiguation of verbs by treating it as a problem of learning with partially labeled structures and proposed to use EM to train the joint model. Finkel and Manning (2009) modeled the task of named entity recognition together with parsing. Meza-Ruiz and Riedel (2009) modeled verb SRL, predicate identification and predicate sense recognition jointly using Markov Logic. Henderson et al. (2008) was designed for jointly learning to predict syntactic and semantic dependencies. Dahlmeier et

al. (2009) addressed the problem of jointly learning verb SRL and preposition sense using the Penn Treebank annotation that was introduced in that work. The key difference between these and the model presented in this paper lies in the simplicity of our model and its easy extensibility because it leverages existing trained systems. Moreover, our model has the advantage that the complexity of the joint parameters is small, hence does not require a large jointly labeled dataset to train the scaling parameters.

Our approach is conceptually similar to that of Rush et al. (2010), which combined separately trained models by enforcing agreement using global inference and solving its linear programming relaxation. They applied this idea to jointly predict dependency and phrase structure parse trees and on the task of predicting full parses together with part-of-speech tags. The main difference in our approach is that we treat the scaling problem as a separate learning problem in itself and train a joint model specifically for re-scaling the output of the trained systems.

The SRL combination system of Surdeanu et al. (2007) studied the combination of three different SRL systems using constraints and also by training secondary scoring functions over the individual systems. Their approach is similar to the one presented in this paper in that, unlike standard reranking, as in Collins (2000), we entertain all possible solutions during inference, while reranking approaches train a discriminative scorer for the top-K solutions of an underlying system. Unlike the SRL combination system, however, our approach spans multiple phenomena. Moreover, in contrast to their re-scoring approaches, we do not define joint features drawn from the predictions of the underlying components to define our global model.

We consider the tasks verb SRL and preposition roles and combine their predictions to provide a richer semantic annotation of text. This approach can be easily extended to include systems that predict structures for other linguistic phenomena because we do not retrain the underlying systems. The semantic relations can be enriched by incorporating more linguistic phenomena such as nominal SRL, defined by the Nombank annotation scheme of Meyers et al. (2004), the preposition function analysis of O'Hara and Wiebe (2009) and noun compound analysis as defined by Girju (2007) and Girju et al. (2009) and others. This presents an exciting direction for future work.

7 Conclusion

This paper presents a strategy for extending semantic role labeling without the need for extensive retraining or data annotation. While standard semantic role labeling focuses on verb and nominal relations, sentences can express relations using other lexical items also. Moreover, the different relations interact with each other and constrain the possible structures that they can take. We use this intuition to define a joint model for inference. We instantiate our model using verb semantic role labeling and preposition role labeling and show that, using linguistic constraints between the tasks and minimal joint learning, we can improve the performance of both tasks. The main advantage of our approach is that we can use existing trained models without re-training them, thus making it easy to extend this work to include other linguistic phenomena.

Acknowledgments

The authors thank the members of the Cognitive Computation Group at the University of Illinois for insightful discussions and the anonymous reviewers for valuable feedback.

This research is supported by the Defense Advanced Research Projects Agency (DARPA) Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181. Any opinions, ndings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reect the view of the DARPA, AFRL, or the US government.

References

G. Andrew, T. Grenager, and C. D. Manning. 2004. Verb sense and subcategorization: Using joint inference to improve performance on complementary tasks. In *Proceedings of EMNLP*.

X. Carreras and L. Màrquez. 2004. Introduction to the CoNLL-2004 shared tasks: Semantic role labeling. In *Proceedings of CoNLL-2004*.

X. Carreras and L. Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of CoNLL-2005*.

M. Chang, L. Ratinov, and D. Roth. 2011. Structured learning with constrained conditional models. *Machine Learning (To appear)*.

E. Charniak and M. Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *ACL*.

M. Collins. 2000. Discriminative reranking for natural language parsing. In *ICML*.

M. Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *EMNLP*.

D. Dahlmeier, H. T. Ng, and T. Schultz. 2009. Joint learning of preposition senses and semantic roles of prepositional phrases. In *EMNLP*.

J. R. Finkel and C. D. Manning. 2009. Joint parsing and named entity recognition. In *NAACL*.

R. Girju, P. Nakov, V. Nastase, S. Szpakowicz, P. Turney, and D. Yuret. 2009. Classification of semantic relations between nominals. *Language Resources and Evaluation*.

R. Girju. 2007. Improving the interpretation of noun phrases with cross-linguistic information. In *ACL*.

Y. Goldberg and M. Elhadad. 2010. An efficient algorithm for easy-first non-directional dependency parsing. In *NAACL*.

J. Henderson, P. Merlo, G. Musillo, and I. Titov. 2008. A latent variable model of synchronous parsing for syntactic and semantic dependencies. In *CoNLL*.

D. Hovy, S. Tratz, and E. Hovy. 2010. What's in a preposition? dimensions of sense disambiguation for an interesting word class. In *Coling 2010: Posters*.

T. Koo, X. Carreras, and M. Collins. 2008. Simple semi-supervised dependency parsing. In *ACL*.

K. Litkowski and O. Hargraves. 2005. The preposition project. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*.

K. Litkowski and O. Hargraves. 2007. Semeval-2007 task 06: Word-sense disambiguation of prepositions. In *SemEval-2007: 4th International Workshop on Semantic Evaluations*.

A. Martins, N. A. Smith, and E. Xing. 2009. Concise integer linear programming formulations for dependency parsing. In *ACL*.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The NomBank project: An interim report. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*.

I Meza-Ruiz and S. Riedel. 2009. Jointly identifying predicates, arguments and senses using markov logic. In *NAACL*.

T. O'Hara and J. Wiebe. 2009. Exploiting semantic role resources for preposition disambiguation. *Computational Linguistics*, 35(2), June.

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

P. Pantel and D. Lin. 2002. Discovering word senses from text. In *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

V. Punyakanok, D. Roth, W. Yih, and D. Zimak. 2005. Learning and inference over constrained output. In *IJCAI*.

V. Punyakanok, D. Roth, and W. Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*.

L. Ratinov and D. Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*.

S. Riedel and J. Clarke. 2006. Incremental integer linear programming for non-projective dependency parsing. In *EMNLP*.

S. Riedel. 2009. Cutting plane map inference for markov logic. In *SRL 2009*.

N. Rizzolo and D. Roth. 2010. Learning based java for rapid development of nlp systems. In *Language Resources and Evaluation*.

D. Roth and W. Yih. 2004. A linear programming formulation for global inference in natural language tasks. In *CoNLL*.

D. Roth and W. Yih. 2005. Integer linear programming inference for conditional random fields. In *ICML*.

A.M. Rush, D. Sontag, M. Collins, and T. Jaakkola. 2010. On dual decomposition and linear programming relaxations for natural language processing. In *EMNLP*. Association for Computational Linguistics.

M. Surdeanu, L. Màrquez, X. Carreras, and P. R. Comas. 2007. Combination strategies for semantic role labeling. *J. Artif. Int. Res.*, 29:105–151, June.

K. Toutanova, A. Haghighi, and C. D. Manning. 2008. A global joint model for semantic role labeling. *Computational Linguistics*, 34(2).

S. Tratz and D. Hovy. 2009. Disambiguation of preposition sense using linguistically motivated features. In *NAACL: Student Research Workshop and Doctoral Consortium*.

N. Xue and M. Palmer. 2004. Calibrating features for semantic role labeling. In *EMNLP*.

P. Ye and T. Baldwin. 2007. MELB-YB: Preposition Sense Disambiguation Using Rich Semantic Features. In *SemEval-2007*.