Context-aware Frame-Semantic Role Labeling

Michael Roth and **Mirella Lapata**
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
{mroth,mlap}@inf.ed.ac.uk

Abstract

Frame semantic representations have been useful in several applications ranging from text-to-scene generation, to question answering and social network analysis. Predicting such representations from raw text is, however, a challenging task and corresponding models are typically only trained on a small set of sentence-level annotations. In this paper, we present a semantic role labeling system that takes into account sentence and discourse context. We introduce several new features which we motivate based on linguistic insights and experimentally demonstrate that they lead to significant improvements over the current state-of-the-art in FrameNet-based semantic role labeling.

1 Introduction

The goal of *semantic role labeling* (SRL) is to identify and label the arguments of semantic predicates in a sentence according to a set of predefined relations (e.g., "who" did "what" to "whom"). In addition to providing definitions and examples of role labeled text, resources like FrameNet (Ruppenhofer et al., 2010) group semantic predicates into so-called *frames*, i.e., conceptual structures describing the background knowledge necessary to understand a situation, event or entity as a whole as well as the roles participating in it. Accordingly, semantic roles are defined on a per-frame basis and are shared among predicates.

In recent years, frame representations have been successfully applied in a range of downstream tasks, including question answering (Shen and Lapata, 2007), text-to-scene generation (Coyne et al., 2012), stock price prediction (Xie et al., 2013), and social network extraction (Agarwal et al., 2014). Whereas some tasks directly utilize information encoded in the FrameNet resource, others make use of FrameNet indirectly through the output of SRL systems that are trained on data annotated with frame-semantic representations. While advances in machine learning have recently given rise to increasingly powerful SRL systems following the FrameNet paradigm (Hermann et al., 2014; Täckström et al., 2015), little effort has been devoted to improve such models from a linguistic perspective.

In this paper, we explore insights from the linguistic literature suggesting a connection between discourse and role labeling decisions and show how to incorporate these in an SRL system. Although early theoretical work (Fillmore, 1976) has recognized the importance of discourse context for the assignment of semantic roles, most computational approaches have shied away from such considerations. To see how context can be useful, consider as an example the DELIVERY frame, which states that a THEME can be handed off to either a RECIPIENT or "more indirectly" to a GOAL. While the distinction between the latter two roles might be clear for some fillers (e.g., people vs. locations), there are others where both roles are equally plausible and additional information is required to resolve the ambiguity (e.g., countries). If we hear about *a letter being delivered to Greece*, for instance, reliable cues might be whether the sender is a person or a country and

Transactions of the Association for Computational Linguistics, vol. 3, pp. 449–460, 2015. Action Editor: Diana McCarthy.
Submission batch: 5/2015; Revision batch 7/2015; Published 8/2015.

whether *Greece* refers to the geographic region or to the Greek government.

The example shows that context can generally influence the choice of correct role label. Accordingly, we assume that modeling contextual information, such as the meaning of a word in a given situation, can improve semantic role labeling performance. To validate this assumption, we explore different ways of incorporating contextual cues in a SRL model and provide experimental support that demonstrates the usefulness of such additional information.

The remainder of this paper is structured as follows. In Section 2, we present related work on semantic role labeling and the various features applied in traditional SRL systems. In Section 3, we provide additional background on the FrameNet resource. Sections 4 and 5 describe our baseline system and contextual extensions, respectively, and Section 6 presents our experimental results. We conclude the paper by discussing in more detail the output of our system and highlighting avenues for future work.

2 Related Work

Early work in SRL dates back to Gildea and Jurafsky (2002), who were the first to model role assignment to verb arguments based on FrameNet. Their model makes use of lexical and syntactic features, including binary indicators for the words involved, syntactic categories, dependency paths as well as position and voice in a given sentence. Most subsequent work in SRL builds on Gildea and Jurafsky's feature set, often with the addition of features that describe relevant syntactic structures in more detail, e.g., the argument's leftmost/rightmost dependent (Johansson and Nugues, 2008).

More sophisticated features include the use of convolution kernels (Moschitti, 2004; Croce et al., 2011) in order to represent predicate-argument structures and their lexical similarities more accurately. Beyond lexical and syntactic information, a few approaches employ additional semantic features based on annotated word senses (Che et al., 2010) and selectional preferences (Zapirain et al., 2013). Deschacht and Moens (2009) and Huang and Yates (2010) use sentence-internal sequence information, in the form of latent states in a hidden markov model. More recently, a few approaches (Roth and Woodsend, 2014; Lei et al., 2015; Foland and Martin, 2015) explore ways of using low-rank vector and tensor approximations to represent lexical and syntactic features as well as combinations thereof.

To the best of our knowledge, there exists no prior work where features based on discourse context are used to assign roles on the sentence level. Discourse-like features have been previously applied in models that deal with so-called implicit arguments, i.e., roles which are not locally realized but resolvable within the greater discourse context (Ruppenhofer et al., 2010; Gerber and Chai, 2012). Successful features for resolving implicit arguments include the distance between mentions and any discourse relations occurring between them (Gerber and Chai, 2012), roles assigned to mentions in the previous context, the discourse prominence of the denoted entity (Silberer and Frank, 2012), and its centering status (Laparra and Rigau, 2013). None of these features have been used in a standard SRL system to date (and trivially, not all of them will be helpful as, for example, the number of sentences between a predicate and an argument is always zero within a sentence). In this paper, we extend the contextual features used for resolving implicit arguments to the SRL task and show how a set of discourse-level enhancements can be added to a traditional sentence-level SRL model.

3 FrameNet

The Berkeley FrameNet project (Ruppenhofer et al., 2010) develops a semantic lexicon and an annotated example corpus based on Fillmore's (1976) theory of frame semantics. Annotations consist of *frame-evoking elements* (i.e., words in a sentence that are associated with a conceptual frame) and *frame elements* (i.e., instantiations of semantic roles, which are defined per frame and filled by words or word sequences in a given sentence). For example, the DELIVERY frame describes a scene or situation in which a DELIVERER hands off a THEME to a RECIPIENT or a GOAL.[1] In total, there are 1,019 frames and 8,886 frame elements defined in the lat-

[1] See `https://framenet2.icsi.berkeley.edu/` for a comprehensive list of frames and their definitions.

est publicly available version of FrameNet.[2] An average number of 11.6 different frame-evoking elements are provided for each frame (11,829 in total). Following previous work on FrameNet-based SRL, we use the full text annotation data set, which contains 23,087 frame instances.

Semantic annotations for frame instances and fillers of frame elements are generally provided at the level of word sequences, which can be single words, complete or incomplete phrases, and entire clauses (Ruppenhofer et al., 2010, Chapter 4). An instance of the DELIVERY frame, with annotations of the frame-evoking element (underlined) and instantiated frame elements (in brackets), is given in the example below:

(1) The Soviet Union agreed to speed up [oil]$_{\text{THEME}}$ deliveries$_{\text{DELIVERY}}$ [to Yugoslavia]$_{\text{RECIPIENT}}$.

Note that the *oil deliveries* here concern *Yugoslavia* as a geopolitical entity and hence the RECIPIENT role is assigned. If *Yugoslavia* was referred to as the location of a delivery, the GOAL role would be assigned instead. In general, roles can be restricted by so-called semantic types (e.g., every filler of the THEME element in the DELIVERY frame needs to be a *physical_object*). However, not all roles are typed and whether a specific phrase is a suitable filler largely depends on context.

4 Baseline Model

As a baseline for implementing contextual enhancements to an SRL model, we use the semantic role labeling components provided by the mate-tools (Björkelund et al., 2010). Given a frame-evoking element in a sentence and its associated frame (i.e., a predicate and its sense), the mate-tools form a pipeline of logistic regression classifiers that identify and label frame elements which are instantiated within the same sentence (i.e., a given predicate's arguments).

The adopted SRL system has been developed for PropBank/NomBank-style role labeling and we make several changes to adapt it to FrameNet. Specifically, we change the argument labeling procedure from predicate-specific to frame-specific roles and implement I/O methods to read and generate FrameNet XML files. For direct comparison with the previous state-of-the-art for FrameNet-based SRL, we further implement additional features used in the SEMAFOR system (Das et al., 2014) and combine the role labeling components of mate-tools with SEMAFOR's preprocessing toolchain.[3] All features used in our system are listed in Table 1.

The main differences between our adaptation of mate-tools and SEMAFOR are as follows: whereas the latter implements identification and labeling of role fillers in one step, mate-tools follow the insight that these two steps are conceptually different (Xue and Palmer, 2004) and should be modeled separately. Accordingly, mate-tools contain a global reranking component which takes into account identification and labeling decisions while SEMAFOR only uses reranking techniques to filter overlapping argument predictions and other constraints (see Das et al., 2014 for details). We discuss the advantage of a global reranker for our setting in Section 5.

5 Extensions based on Context

Context can be relevant for semantic role labeling in various different ways. In this section, we motivate and describe four extensions over previous approaches.

The first extension is a set of features that model document-specific aspects of word meaning using distributional semantics. The motivation for this feature class stems from the insight that the meaning of a word in context can influence correct role assignment. While concepts such as polysemy, homonymy and metonymy are all relevant here, the scarce training data available for FrameNet-based SRL calls for a light-weight model that can be applied without large amounts of labeled data. We therefore employ distributional word representations which we critically adapt based on document content. We describe our contribution in Section 5.1.

Entities that fill semantic roles are sometimes mentioned in discourse. Given a specific mention

[2] Version 1.5, released September 2010.

[3] We note that better results have been reported in Hermann et al. (2014) and Täckström et al. (2015). However, both of these more recent approaches rely on a custom frame identification component as well as proprietary tools and models for tagging and parsing which are not publicly available.

Argument identification and classification	
Lemma form of f	POS tag of f
Any syntactic dependents of f*	Subcat frame of f*
Voice of a*	Any lemma in a*
Number of words in a	
First word and POS tag in a	
Second word and POS tag in a	
Last word and POS tag in a	
Relation from first word in a to its parent	
Relation from second word in a to its parent	
Relation from last word in a to its parent	
Relative position of a with respect to p	
Voice of a and relative position with respect to p*	
Identification only	
Lemma form of the first word in a	
Lemma form of the syntactic head of a	
Lemma form of the last word in a	
POS tag of the first word in a	
POS tag of the syntactic head of a	
POS tag of the last word in a	
Relation from syntactic head of a to its parent	
Dependency path from a to f	
Length of dependency path from a to f	
Number of words between a and f	

Table 1: Features from Das et al. (2014) which we adopt in our model; a denotes the argument span under consideration, f refers to the corresponding frame evoking element. Identification features are instantiated as binary indicator features. Features marked with an asterisk are role specific. All other features apply to combinations of role and frame.

for which a role is to be predicted, we can also directly use previous role assignments as classification cues. We describe our implementation of this feature in Section 5.2.

The filler of a semantic role is often a word or phrase which occurs only once or a few times in a document. If neither syntax nor aspects of lexical meaning provide cues indicating a unique role, useful information can still be derived from the discourse salience of the denoted entity. Our model makes use of a simple salience indicator that can be reliably derived from automatically computed coreference chains. We describe the motivation and actual implementation of this feature in Section 5.3.

The aforementioned features will influence role labeling decisions directly, however, further improvements can be gained by considering interactions between labeling decisions. As discussed in Das et al. (2014), role annotations in FrameNet are unique with respect to a frame instance in more than 96% of cases. This means that even if a feature is not a positive indicator for a candidate role filler, knowing it would be a better cue for another candidate can also prevent a hypothetical model from assigning a frame element label incorrectly. While this kind of knowledge has been successfully implemented as constraints in recent FrameNet-based SRL models (Hermann et al., 2014; Täckström et al., 2015), earlier work on PropBank-based role labeling suggests that better performance can be achieved with a re-ranking component which has the potential to learn such constraints and other interactions implicitly (Toutanova et al., 2005; Björkelund et al., 2010). In our model, we adopt the latter method and extend it with additional frame-based features. We describe this approach in more detail in Section 5.4.

5.1 Modeling Word Meaning in Context

The underlying idea of distributional models of semantics is that meaning can be acquired based on distributional properties (typically represented by co-occurrence counts) of linguistic entities such as words and phrases (Sahlgren, 2008). Although the absolute meaning of distributional representations remains unclear, they have proven highly successful for modeling relative aspects of meaning, as required for instance in word similarity tasks (Mikolov et al., 2013; Pennington et al., 2014). Given their ability to model lexical similarity, it is not surprising that such representations are also successful at representing similar words in semantic tasks related to role labeling (Pennacchiotti et al., 2008; Croce et al., 2010; Zapirain et al., 2013).

Although distributional representations can be used directly as features for role labeling (Padó et al., 2008; Gorinski et al., 2013; Roth and Woodsend, 2014, inter alia), further gains should be possible when considering document-specific properties such as genre and context. This is particularly true in the context of FrameNet, where different senses are observed across a diverse range of texts including spoken dialogue and debate transcripts as well

Country	Frame	Frame Element
Iran	Supply	RECIPIENT
	Commerce_buy	BUYER
China	Supply	SUPPLIER
	Commerce_sell	SELLER
Iraq	Locative_relation	GROUND
	Arriving	GOAL

Table 2: Most frequent roles assigned to country names appearing FrameNet texts: whereas Iran and China are mostly mentioned in an economic context, references to Iraq are mainly found in a news article about a politician's visit to the country.

as travel guides and newspaper articles. Country names, for example, can be observed as fillers for different roles depending on the text genre and its perspective. Whereas some text may talk about a country as an interesting holiday destination (e.g., "Berlitz Intro to Jamaica"), others may discuss what a country is good at or interested in (e.g., "Iran [Nuclear] Introduction"). A list of the most frequent roles assigned to different country names are displayed in Table 2.

Previous approaches model word meaning in context (Thater et al., 2010; Dinu and Lapata, 2010, inter alia) using sentence-level information which is already available in traditional SRL systems in the form of explicit features. Here, we go one step further and define a simple model in which word meaning representations are adapted to each document. As a starting point, we use the GloVe toolkit (Pennington et al., 2014) for learning representations[4] and apply it to the Wikipedia corpus made available by the Westbury Lab.[5] The learned representations can be seen as word vectors whose components encode basic bits of related encyclopaedic knowledge. We adapt these general representations to the actual meaning of a word in a particular text by running additional iterations of the GloVe toolkit using document-specific co-occurrences as input and Wikipedia-based representations for initialization. To make up for the large difference in data size between the Wikipedia corpus and a single document, we normalize co-occurrence counts based on the ratio between the absolute numbers of co-occurrences in both resources.

Given co-occurrence matrices C_{wiki} and C_d, and the vocabulary V, we formally define the features of our SRL model as the components of the vector space $\vec{w}_i$ of words w_i ($1 \leq i \leq |V|$) occurring in document d. The representations are learned by applying GloVe to optimize the following objective for n iterations ($1 \leq t \leq n$):

$$J_t = \sum_{i,j} f(X_{ij})(\vec{w}_i^T \vec{w}_j - \log X_{ij})^2, \quad (2)$$

$$\text{where } X = \begin{cases} C_{\text{wiki}} \text{ if } t < t_d \\ C_d \text{ otherwise} \end{cases} \quad (3)$$

The weighting function f scales the impact of each word pair such that unseen pairs do not contribute to the overall objective and frequent co-occurrences are not overweighted. In our experiments, we use the same weighting function and parametrization as defined in Pennington et al. (2014). We further set the number of iterations to be performed on each co-occurrence matrix following results of an initial cross-validation experiment on our training data ($t_d = 50, n = 100$).

5.2 Co-occurring Roles

If an entity is mentioned several times in discourse, it is likely that it also fills several roles. Whereas the distributional model described in Section 5.1 provides us with information regarding the role assignments suitable for an entity given co-occurring words, we can also can explicitly consider previous role assignments to the same entity. As shown in Table 2, a country that fills the SUPPLIER role is more likely to also fill the role of a SELLER than that of a BUYER. Given the high number of different frame elements in FrameNet, only a small fraction of pairs can be found in the training data, which entails that directly utilizing role co-occurrences might not be helpful. In order to benefit from previous role assignments in discourse, we follow related work on resolving implicit arguments (Ruppenhofer et al., 2011; Silberer and Frank, 2012) and consider the semantic types of role assignments (see Section 3) as

[4] We selected this toolkit in our work due to its flexibility: as it directly operates over co-occurrence matrices, we can manipulate counts prior to word vector computation and easily take into account multiple matrices.

[5] http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html

features instead of the role labels themselves. This tremendously reduces the feature space from more than 8,000 options (number of defined frame elements) to just 27 (number of semantic types observed for frame elements in the training data).

In practice, we define one binary indicator feature f_s for each semantic type s observed at training time. Given a potential filler, we set the feature value of f_s to 1 (otherwise 0) if and only if there exists a co-referent entity mention annotated as a frame element filler with semantic type s. Since texts in FrameNet do not contain any manual mark-up of coreference relations, we rely on entity mentions and coreference chains predicted by the Stanford Coreference Resolution system (Lee et al., 2013).

5.3 Discourse Newness

Our third contextual feature type is based on the observation that the salience of a discourse entity and its semantic prominence are interrelated. Previous work (Rose, 2011) showed that semantic prominence, as signal-led by semantic roles, can better explain subsequent phenomena related to discourse salience (such as pronominalization) than syntactic indicators. Our question here is whether this insight can be also applied in reverse. Can information on discourse salience be useful as an indicator for semantic roles?

For this feature, we make use of the same coreference chains as predicted for determining co-occurring roles. Unfortunately, automatically predicted mentions and coreference chains are noisy. To identify particularly reliable indicators for discourse salience, we inspected held-out development data. One such indicator is whether an entity is mentioned for the first time (discourse-new) or has been mentioned before (discourse-old). Let w denote an entity and $R_1...R_n$ the set of all co-reference chains with mentions $r_1...r_m \in R_i$ $(1 \leq i \leq n)$ ordered by their appearance in text. We define discourse newness based on head words r.head as:

$$\text{new}(w) = \begin{cases} 0 \text{ if } \exists r_j \in R_i : j > 1 \wedge r_j.\text{head} \equiv w \\ 1 \text{ else} \end{cases} \quad (4)$$

Although this feature is a simple binary indicator, it can be very useful for distinguishing between roles that are more or less likely to be assigned to new entities. For example, it is easy to imagine that the RESULT of a CAUSATION is more likely to be discourse-new than the EFFECT that caused it. Table 3 provides an overview of frames found in the training and development data which have roles with substantially different likelihoods for discourse-new fillers.

Frame	Frame Element	new/old
Statement	SPEAKER	43.8
	MESSAGE	99.1
	MEDIUM	80.0
Leadership	LEADER	78.0
	GOVERNED	93.4
Intensionally_create	CREATOR	58.8
	CREATED_ENTITY	90.1

Table 3: Frequent frames that have elements with different likelihoods of discourse-new vs. discourse-old fillers; new/old ratios as observed on the development set.

5.4 Frame-based Reranking

Our goal is to learn a better model for FrameNet-based semantic role labeling using linguistically inspired features such as those described in the previous sections. To do this, we need a framework for representing single role assignments *and* a model of how such assignments depend on each other within a frame instance. Inspired by previous work on reranking in SRL, we assume that we can find the correct filler of a frame element based on the top k roles predicted for each candidate word sequence. We leverage this assumption to train a reranking model that considers the top predictions for each candidate and uses all relevant features to select the best overall structure.

Our implementation of the reranking model is an adaptation of the reranker made available in the mate-tools (see Section 4), which we extend to deal with frame-specific features and arbitrary role labels. As features for the global component, we apply all local features and additionally use the following two types of indicator features on the whole frame structure:

- Total number of roles in the predicted structure
- Ordered set of predicted role labels

Frames	SRL model	P	R	F_1
gold	SEMAFOR[7]	78.4	**73.1**	75.7*
gold	Framat	80.3	71.7	75.8*
gold	Framat$^{+context}$	**80.4**	73.0	**76.5**
SEMAFOR	SEMAFOR	69.2	**65.1**	67.1*
SEMAFOR	Framat	71.1	63.7	67.2*
SEMAFOR	Framat$^{+context}$	**71.1**	64.8	**67.8**

Table 4: Full structure prediction results using gold (top) and predicted frames (bottom). All numbers are percentages. $*$ Significantly different ($p<0.05$) from Framat$^{+context}$.

At test time, the reranker takes as input the n-best labels for the m-best fillers of a frame structure, computes a global score for each of the $n \times m$ possible combinations and returns the structure with the highest overall score as its prediction output. Based on initial experiments on our training data, we set these parameters to $m = 8$ and $n = 4$.

6 Experiments

In this section, we demonstrate the usefulness of contextual features for FrameNet-based SRL models. Our hypothesis is that contextual information can considerably improve an existing semantic role labeling system. Accordingly, we test this hypothesis based on the output of three different systems. The first system, henceforth called *Framat* (short for *Fra*meNet-adapted *mat*e-tools) is the baseline system described in Section 4. The second system, henceforth *Framat$^{+context}$*, is an enhanced version of the baseline that additionally uses all extensions described in Section 5. Finally, we also consider the output of SEMAFOR (Das et al., 2014), a state-of-the-art model for frame-semantic role labeling. Although all systems are provided with entire documents as input, SEMAFOR and Framat process each document sentence-by-sentence whereas Framat$^{+context}$ also uses features over *all* sentences.

For evaluation, we use the same FrameNet training and evaluation texts as established in Das and Smith (2011). We compute precision, recall and F_1-score using the modified SemEval-2007 scorer from the SEMAFOR website.[6]

[6] http://www.ark.cs.cmu.edu/SEMAFOR/eval/

[7] Results produced by running SEMAFOR on the exact same frame instances for training and testing as our own models.

Model/added feature	P	R	F_1
Framat w/o reranker	77.5	72.5	74.9
+discourse newness	77.6	72.3	74.9
+word meaning vectors	77.9	72.7	75.2
+cooccurring roles	77.9	72.8	75.3
+reranker	**80.6**	72.7	76.4
+frame structure	80.4	**73.0**	**76.5**

Table 5: Full structure prediction results using gold frames, Framat and different sets of context features. All numbers are percentages.

Results Table 4 summarizes our results with Framat, Framat$^{+context}$, and SEMAFOR using gold and predicted frames (see the upper and lower half of the table, respectively). Although differences in system architecture lead to different precision/recall trade-offs for Framat and SEMAFOR, both systems achieve comparable F_1 (for both gold and predicted frames). Compared to Framat, we can see that the contextual enhancements implemented in our Framat$^{+context}$ model lead to immediate gains of 1.3 points in recall, corresponding to a significant increase of 0.7 points in F_1. Framat$^{+context}$'s recall is slightly below that of SEMAFOR (73.0% vs. 73.1%), however, it achieves a much higher level of precision (80.4% vs. 78.4%).

We examined whether differences in performance among the three systems are significant using an approximate randomization test over sentences (Yeh, 2000). SEMAFOR and Framat perform significantly worse ($p<0.05$) compared to Framat$^{+context}$ both when gold and predicted frames are used. In the remainder of this section we discuss results based on gold frames, since the focus of this work lies primarily on the role labeling task.

Impact of Individual Features We demonstrate the effect of adding individual context-based features to the Framat model in a separate experiment. Whereas all models in the previous experiment used a reranker for direct comparability, here we start with the Framat baseline (without a reranker) and add each enhancement described in Section 5 incrementally. As summarized in Table 5, the baseline without a reranker achieves a precision and

recall of 77.5% and 72.5%, respectively. Addition of our discourse new feature increases precision (+0.1%), but also reduces recall (−0.2%). Adding word meaning vectors compensates for the loss in recall (+0.4%) and further increases precision (+0.3%). Information about role assignments to coreferring mentions increases recall (+0.1%) while retaining the same level of precision. Finally, we can see that jointly considering role labeling decisions in a global reranker with additional features on frame structure leads to the strongest boost in performance, with combined additional gains in precision and recall of +2.5% and +0.2%, respectively. Interestingly, the gains realized here are much higher compared to when adding the reranker to the Framat model without contextual features, which corresponds to a +2.8% increase in precision but a −0.8% reduction in recall.

General vs. Document-specific Vectors We also assessed the impact of adapting vectors to documents (see Table 6). Specifically, we compared a version of the Framat$^{+\text{context}}$ model without any vectors against a model using the adaptation technique presented in Section 5.1 and a simpler alternative which obtains GloVe representations trained on the Wikipedia corpus and FrameNet texts. The latter model does not explicitly take document information into account, but it should be able to yield vectors representative of the FrameNet domains, merely by being trained on them. As shown in Table 6, our adaptation technique is superior to learning word representations based on Wikipedia and all FrameNet texts at once. Using the components of document-specific vectors as features improves precision and recall by +0.7 percentage points over Framat$^{+\text{context}}$ without vectors. Word representations trained on Wikipedia and FrameNet improve precision by +0.2 percentage points and recall by +0.6.

Qualitative Improvements In addition to quantitative gains, we also observe qualitative improvements when considering contextual features. A set of example predictions by different models are listed in Table 7. The annotations show that Framat and SEMAFOR mislabel several cases that are correctly classified by Framat$^{+\text{context}}$.

In the first example, only Framat$^{+\text{context}}$ is able to predict that *on Dec. 1* fills the frame element

Model/word representations	P	R	F_1
Framat$^{+\text{context}}$ without vectors	79.7	72.2	75.8
+document-specific vectors	**80.4**	**73.0**	**76.5**
+general (Wiki+FN) vectors	79.9	72.8	76.2

Table 6: Full structure prediction results using gold frames, Framat$^{+\text{context}}$ and different vector representations. All numbers are percentages.

TIME. This may seem trivial at first glance but is actually remarkable as the word token *Dec* neither occurs in the training data nor is well represented as a time expression in Wikipedia. The only way the model is able to label this phrase correctly is by finding that corresponding word tokens are similarly distributed across the test document as other time expressions are in the training data. In the second and third examples, correct assignments require some form of world knowledge which is not expressed within the respective sentences but might be approximated based on context. For example, knowing that *aunt*, *uncle* and *grandmother* are role fillers of a KINSHIP frame means that they are of the semantic type *human* and thus only compatible with the frame element RECIPIENT, not with GOAL. Similarly, correctly classifying the relation between *Clinton* and *stooge* in the last example is only possible if the model has access to some information that makes *Clinton* a likely filler of the SUPERIOR role. We conjecture that document-specific word vector representations provide such information given that *Clinton* co-occurs in the document with words such as *president*, *chief*, and *claim*.

Overall, we find that the features introduced in Section 5 model a fair amount of contextual information which can help a semantic role labeling model to perform better decisions.

7 Discussion

In this section, we discuss the extent to which our model leverages the full potential of contextual features for semantic role labeling. We manually examine role assignments to frame elements which seem particularly sensitive to context. We analyze such frame elements based on differences in label assignment between Framat and Framat$^{+\text{context}}$ that can be traced back to factors such as *agency* in dis-

SEMAFOR	*Can [he]$_{\text{THEME}}$ go$_{\text{MOTION}}$ [to Paris]$_{\text{GOAL}}$ on Dec. 1 ?
Framat	*Can [he]$_{\text{THEME}}$ go$_{\text{MOTION}}$ [to Paris on Dec. 1]$_{\text{GOAL}}$?
Framat$^{+\text{context}}$	Can [he]$_{\text{THEME}}$ go$_{\text{MOTION}}$ [to Paris]$_{\text{GOAL}}$ [on Dec. 1]$_{\text{TIME}}$?
SEMAFOR	*Send$_{\text{SENDING}}$ [my regards]$_{\text{THEME}}$ to my aunt , uncle and grandmother .
Framat	*Send$_{\text{SENDING}}$ [my regards]$_{\text{THEME}}$ [to my aunt , uncle and grandmother]$_{\text{GOAL}}$.
Framat$^{+\text{context}}$	Send$_{\text{SENDING}}$ [my regards]$_{\text{THEME}}$ [to my aunt , uncle and grandmother]$_{\text{RECIPIENT}}$.
SEMAFOR	*Stephanopoulos does n't want to seem a Clinton stooge$_{\text{SUBORDINATES_AND_SUPERIORS}}$
Framat	*Stephanopoulos doesn't want to seem a [Clinton]$_{\text{DESCRIPTOR}}$ stooge$_{\text{SUBORDINATES_AND_SUPERIORS}}$
Framat$^{+\text{context}}$	Stephanopoulos does n't want to seem a [Clinton]$_{\text{SUPERIOR}}$ stooge$_{\text{SUBORDINATES_AND_SUPERIORS}}$

Table 7: Examples of frame structures that are labeled incorrectly (marked by asterisks) without contextual features.

course and *word sense* in context. We investigate whether our model captures these factors successfully and showcase examples while reporting absolute changes in precision and recall.

7.1 Agency and Discourse

Many frame elements in FrameNet indicate agency, a property that we expect to highly correlate with contextual features on semantic types of assigned roles (see Section 5.2) and discourse salience (see Section 5.3). Analysis of system output revealed that such features indeed affect and generally improve role labeling. Considering all AGENT elements across frames, we observe absolute improvements of 4% in precision and 3% in recall. In the following, we provide a more detailed analysis of two specific frame elements: the low frequent AGENT element of the PROJECT frame and the highly frequent SPEAKER element in the STATEMENT frame.

The AGENT of a PROJECT is defined as the "individual or organization that carries out the PROJECT". The main difficulty in identifying instances of this frame element is that the frame-evoking target word is typically a noun such as *project*, *plan*, or *program* and hence syntactic features on word-word dependencies do not provide sufficient cues. We found several cases where context provided missing cues, leading to an increase in recall from 56% to 78%. In cases where additional features did not help, we identified two types of errors: firstly, the filler was too far from the target word and therefore could not be identified as a filler at all ("[North Korea]$_{\text{AGENT}}$ is developing ... program$_{\text{PROJECT}}$"), and secondly, earlier mentions indicating agency were not detected by the coreference resolution system ("The IAEA assisted Syria (...) This study was part of an IAEA$_{\text{AGENT}}$ .. program$_{\text{PROJECT}}$).

The SPEAKER of a STATEMENT is defined as "the *sentient* entity that produces [a] MESSAGE". Instances of the STATEMENT frame are frequently evoked by verbs such as *say*, *mention*, and *claim*. The SPEAKER role can be hard to identify in subject position as an unknown entity could also fill the MEDIUM role. For example, "a report claims that ..." should be analyzed differently from "a person claims". Our contextual features improve role labeling in cases where the subject can be classified based on previous role assignments. On the negative side, we found our model to be too conservative in some cases where a subject is discourse new. Additional gains would be possible with improved coreference chains that include pronouns such as *some* and *I*. Such chains could be established through a better preprocessing pipeline or by utilizing additional linguistic resources.

7.2 Word Meaning and Context

As discussed earlier, we expect that the meaning of a word in context provides valuable cues regarding potential frame elements. Two types of words are of particular interest here: *ambiguous* words, for which different senses might apply depending on context, and *out-of-vocabulary* words, for which no clear sense could be established during training. In the following, we take a closer look at differences in role assignment between Framat and Framat$^{+\text{context}}$ for such fillers.

Ambiguous words that occur as fillers of different frame elements in the test set include *party*,

power, *program*, and *view*. We find occurrences of these words in two broad types of contexts: political and non-political. Within political contexts, *party* and *power* fill frame elements such as POSSESSION and LEADER. Outwith political contexts, we find frame elements such as ELECTRICITY and SOCIAL_EVENT to be far more likely. The Framat model exhibits a general bias towards the political domain, often missing instances of frame elements that are more common in non-political contexts (e.g., "the six-[party]$_{\text{INTERLOCUTORS}}$ talks$_{\text{DISCUSSION}}$"). Framat$^{+\text{context}}$, in contrast, shows less of a bias and provides better classification based on context features for all frame elements. Overall, precision for the four ambiguous words is improved from 86% to 93%, with a few errors remaining due to rare dependency paths (e.g., [*program*]$_{\text{ACT}}$ $\xleftarrow{\text{NMOD}}$ *which* $\xleftarrow{\text{SBAR}}$ *is* $\xleftarrow{\text{PRD}}$ *violation*$_{\text{COMPLIANCE}}$) and differences between frame elements that depend on factors such as number (COGNIZER vs. COGNIZER_1).

A frequently observed error by the baseline model is to assign peripheral frame elements such as TIME to role fillers that actually are not time expressions. This happens because words which have not been seen frequently during training but appear in adverbial positions are generally likely to fill the frame element TIME. We find that the use of document-specific word vector representations drastically reduces the number of such errors (e.g., "to give$_{\text{GIVING}}$ [generously]$_{\text{MANNER vs. *TIME}}$"), with absolute gains in precision and recall of 14% and 9%, respectively, presumably because non-time expressions are often distributed differently across a document than time expressions. Document-specific word vector representations also improve recall for out-of-vocabulary words, as seen with the example of *Dec* discussed in Section 6. However, such representations by themselves might be insufficient to determine which aspects of a word sense are applicable across a document as occurrences in specific contexts may also be misleading (e.g., "...changes [throughout the *community*]" vs. "... [throughout the *ages*]$_{\text{TIME}}$"). Some of these cases could be resolved using higher level features that explicitly model interactions between (predicted) word meaning in context and other factors, however we leave this to future work.

8 Conclusions

In this paper, we enriched a traditional semantic role labeling model with additional information from context. The corresponding features we defined can be grouped into three categories: (1) discourse-level features that directly utilize discourse knowledge in the form of coreference chains (newness, prior role assignments), (2) sentence-level features that model properties of a frame structure as a whole, and (3) lexical features that can be computed using methods from distributional semantics and an adaptation to model document-specific word meaning.

To implement our discourse-level enhancements, we modified a semantic role labeling system developed for PropBank/NomBank which we found to achieve competitive performance on FrameNet-based annotations. Our main contribution lies in extending this system to the discourse level. Our experiments revealed that discourse aware features can significantly improve semantic role labeling performance, leading to gains of over $+2.0$ percentage points in precision and state-of-the-art results in terms of F_1. Analysis of system output revealed two reasons for improvement. Firstly, contextual features provide necessary additional information to understand and assign roles on the sentence level, and secondly, some of our discourse-level features generalize better than traditional lexical and syntactic features. We further found that additional gains can be achieved using improved preprocessing tools and a more sophisticated model for feature interactions. In the future, we are planning to assess whether discourse-level features generalize cross-linguistically. We would also like to investigate whether semantic role labeling can benefit from recognizing textual entailment and high-level discourse relations. Our code is publicly available under `http://github.com/microth/mateplus`.

Acknowledgements

We are grateful to Diana McCarthy and three anonymous referees whose feedback helped to substantially improve the present paper. The research presented in this paper was funded by a DFG Research Fellowship (RO 4848/1-1).

References

Apoorv Agarwal, Sriramkumar Balasubramanian, Anup Kotalwar, Jiehan Zheng, and Owen Rambow. 2014. Frame semantic tree kernels for social network extraction from text. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 211–219, Gothenburg, Sweden, 26–30 April 2014.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstration Volume*, pages 33–36, Beijing, China.

Wanxiang Che, Ting Liu, and Yongqiang Li. 2010. Improving semantic role labeling with word sense. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 246–249, Los Angeles, California, 1–6 June 2010.

Bob Coyne, Alex Klapheke, Masoud Rouhizadeh, Richard Sproat, and Daniel Bauer. 2012. Annotation tools and knowledge representation for a text-to-scene system. In *Proceedings of 24th International Conference on Computational Linguistics*, pages 679–694, Mumbai, India, 8–15 December 2012.

Danilo Croce, Cristina Giannone, Paolo Annesi, and Roberto Basili. 2010. Towards open-domain semantic role labeling. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 237–246, Uppsala, Sweden, 11–16 July 2010.

Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2011. Structured lexical similarity via convolution kernels on dependency trees. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1034–1046, Edinburgh, United Kingdom.

Dipanjan Das and Noah A. Smith. 2011. Semi-supervised frame-semantic parsing for unknown predicates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* Portland, Oregon, 19–24 June 2011.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-Semantic Parsing. *Computational Linguistics*, 40(1):9–56.

Koen Deschacht and Marie-Francine Moens. 2009. Semi-supervised semantic role labeling using the Latent Words Language Model. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 21–29, Singapore, 2–7 August 2009.

Georgiana Dinu and Mirella Lapata. 2010. Measuring distributional similarity in context. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1162–1172, Cambridge, Massachusetts, 9–11 October 2010.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32.

William Foland and James Martin. 2015. Dependency-based semantic role labeling using convolutional neural networks. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 279–288, Denver, Colorado.

Matthew Gerber and Joyce Chai. 2012. Semantic Role Labeling of Implicit Arguments for Nominal Predicates. *Computational Linguistics*, 38(4):755–798.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Philip Gorinski, Josef Ruppenhofer, and Caroline Sporleder. 2013. Towards weakly supervised resolution of null instantiations. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 119–130, Potsdam, Germany, 19–22 March 2013.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1448–1458, Baltimore, Maryland, 23–25 June 2014.

Fei Huang and Alexander Yates. 2010. Open-domain semantic role labeling by modeling word spans. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 968–978, Uppsala, Sweden, 11–16 July 2010.

Richard Johansson and Pierre Nugues. 2008. The effect of syntactic representation on semantic role labeling. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 393–400, Manchester, United Kingdom, 18–22 August 2008.

Egoitz Laparra and German Rigau. 2013. Sources of evidence for implicit argument resolution. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 155–166, Potsdam, Germany, 19–22 March 2013.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Tao Lei, Yuan Zhang, Lluís Màrquez, Alessandro Moschitti, and Regina Barzilay. 2015. High-order low-rank tensors for semantic role labeling. In *Proceedings*

of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1150–1160, Denver, Colorado.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, 9–15 June 2013.

Alessandro Moschitti. 2004. A study on convolution kernels for shallow statistic parsing. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 335–342, Barcelona, Spain.

Sebastian Padó, Marco Pennacchiotti, and Caroline Sporleder. 2008. Semantic role assignment for event nominalisations by leveraging verbal data. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 665–672, Manchester, United Kingdom.

Marco Pennacchiotti, Diego De Cao, Roberto Basili, Danilo Croce, and Michael Roth. 2008. Automatic induction of FrameNet lexical units. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 457–465, Honolulu, Hawaii, USA, 25–27 October 2008.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar, 25–29 October 2014.

Ralph L Rose. 2011. Joint information value of syntactic and semantic prominence for subsequent pronominal reference. *Salience: Multidisciplinary Perspectives on Its Function in Discourse*, 227:81–103.

Michael Roth and Kristian Woodsend. 2014. Composition of word representations improves semantic role labelling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 407–413, Doha, Qatar, 25–29 October 2014.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. FrameNet II: Extended Theory and Practice. Technical report, International Computer Science Institute, 14 September 2010.

Josef Ruppenhofer, Philip Gorinski, and Caroline Sporleder. 2011. In search of missing arguments: A linguistic approach. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 331–338, Hissar, Bulgaria, 12–14 September 2011.

Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 12–21, Prague, Czech Republic.

Carina Silberer and Anette Frank. 2012. Casting implicit role linking as an anaphora resolution task. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, pages 1–10, Montréal, Canada, 7-8 June.

Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015. Efficient inference and structured learning for semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:29–41.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 948–957, Uppsala, Sweden, 11–16 July 2010.

Kristina Toutanova, Aria Haghighi, and Christopher Manning. 2005. Joint learning improves semantic role labeling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 589–596, Ann Arbor, Michigan, 29–30 June 2005.

Boyi Xie, Rebecca J. Passonneau, Leon Wu, and Germán G. Creamer. 2013. Semantic frames to predict stock price movement. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 873–883, Sofia, Bulgaria, 4–9 August 2013.

Nianwen Xue and Martha Palmer. 2004. Calibrating features for semantic role labeling. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 88–94, Barcelona, Spain, July.

Alexander Yeh. 2000. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 947–953, Saarbrücken, Germany.

Beñat Zapirain, Eneko Agirre, Lluís Màrquez, and Mihai Surdeanu. 2013. Selectional preferences for semantic role classification. *Computational Linguistics*, 39(3):631–663.