K-SRL: Instance-based Learning for Semantic Role Labeling

Alan Akbik
IBM Research Almaden
650 Harry Road, San Jose
CA 95120, USA
akbika@us.ibm.com

Yunyao Li
IBM Research Almaden
650 Harry Road, San Jose
CA 95120, USA
yunyaoli@us.ibm.com

Abstract

Semantic role labeling (SRL) is the task of identifying and labeling predicate-argument structures in sentences with semantic frame and role labels. A known challenge in SRL is the large number of low-frequency exceptions in training data, which are highly context-specific and difficult to generalize. To overcome this challenge, we propose the use of *instance-based learning* that performs no explicit generalization, but rather extrapolates predictions from the most similar instances in the training data. We present a variant of *k-nearest neighbors* (kNN) classification with composite features to identify nearest neighbors for SRL. We show that high-quality predictions can be derived from a very small number of similar instances. In a comparative evaluation we experimentally demonstrate that our instance-based learning approach significantly outperforms current state-of-the-art systems on both in-domain and out-of-domain data, reaching F_1-scores of 89,28% and 79.91% respectively.

1 Introduction

Semantic role labeling (SRL) is the task of annotating predicate-argument structures in sentences with shallow semantic information. One prominent labeling scheme for the English language is the Proposition Bank (Palmer et al., 2005), which annotates predicates with *frame* labels and arguments with *role* labels (see Figure 1 for examples). Frame labels disambiguate the predicate meaning in the context of the sentence. Role labels roughly correspond to simple questions (*who, when, how, why, with whom*) with regards to the disambiguated predicate. SRL has been found useful for a wide range of applications such as information extraction (Fader et al., 2011), question answering (Shen and Lapata, 2007; Maqsud et al., 2014) and machine translation (Lo et al., 2013).

Current state-of-the-art SRL approaches train classifiers with bags of features (Johansson and Nugues, 2008; Björkelund et al., 2009; Choi and Palmer, 2011) to predict semantic labels for each constituent in a sentence. These approaches typically employ classifiers such as logistic regression or SVM that learn for each feature a measure of impact on the classification decision and abstract away from local contexts in specific training examples.

Local bias. We argue that such approaches are not ideal for SRL due to a strong local bias of features within specific contexts. Low-frequency examples in SRL are often not noise to be abstracted away, but rather correspond to exceptions that require explicit handling.

For example, consider the task of argument labeling: Arguments that are syntactically realized as passive subjects are typically labeled **A1**[1]. However, there exist numerous low-frequency exceptions to this rule. For instance, passive subjects of certain frames (such as the frame TELL.01) are most commonly labeled **A2**. See Figure 1 for an example. Other examples of local bias include different types of diathesis alternation which affect specific frames and argument types (Kipper et al., 2008), the syntactic realization of higher order roles (A2 to A5) which is highly irregular among frames (Palmer et al., 2005), and the realization of roles in non-agentive frames. These phenomena are observed only in specific and often low-frequency contexts of composite features, but are highly relevant to SRL.

[1]This corresponds to the linguistic intuition that active subjects are commonly thematic *agents*, while direct objects and passive subjects are most commonly the thematic *patient* or *theme* of a frame (van der Plas et al., 2014)

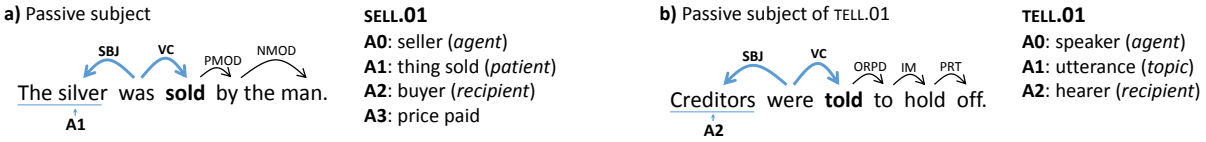

Figure 1: Example sentences with a passive subject (underlined). Passive subjects are typically labeled **A1** (e.g. Sentence a), but there are exceptions to this rule (e.g. frame TELL.01 in Sentence b).

	STATEMENT	COMPOSITE	SUPPORT
(1)	57% of all subjects are labeled **A0**	**P**	17,788 instances
(2)	33% of all subjects are labeled **A1**	**P**	17,788 instances
(3)	74% of *active* subjects are labeled **A0**	**P+V**	13,737 instances
(4)	86% of *passive* subjects are labeled **A1**	**P+V**	4,051 instances
(5)	100% of passive subjects of SELL.01 are labeled **A1**	**P+V+F**	137 instances
(6)	88% of passive subjects of TELL.01 are labeled **A2**	**P+V+F**	53 instances

Table 1: Observations based on CoNLL09 training data: The more atomic features we include in a composite, the more discriminative (and descriptive) it becomes, but generally with lower support.

Feature contexts. We propose to explicitly capture local bias with feature contexts by constructing composites of standard SRL features. Refer to Table 1 for a list of observations over the CoNLL09 shared task gold data (Hajič et al., 2009) for different numbers of features combined into composites: Statements 1 and 2 involve only the syntactic path feature P that models the syntactic function of an argument. Statements 3 and 4 use a composite feature of P and V, the predicate voice feature (either *active* or *passive*). Finally, statements 5 and 6 use a composite feature of P, V and F, the specific frame context.

We make three observations in Table 1: First, the more atomic features we include in a composite feature, the more discriminative it becomes and the more explicitly it captures local bias. For instance, statement 6 is a composite of three atomic features and explicitly captures the exception for passive subjects of TELL.01 discussed earlier. Second, higher order composite features tend to have lower support (i.e. number of training examples that share this combination of atomic features). The use of composite features can therefore aggravate sparsity issues in training data. Finally, since composite features make the interplay of features explicit, they can be rendered as human readable statements. Classification decisions using such features can be easily interpretable for error analysis and extension.

Instance-based Learning for SRL Based on these observations, we propose to use *instance-based learning* (Aha et al., 1991; Daelemans and Van den Bosch, 2005) for SRL. Such learning does not abstract away from specific feature contexts, but rather considers the overall similarity of a test instance to instances in the training data. It has been shown to be applicable to a range of NLP tasks such as PoS tagging (Daelemans et al., 1999), dependency parsing (Nivre et al., 2004) and word sense disambiguation (Veenstra et al., 2000). The arguably most well-known approach of this kind is *k-nearest neighbors* classification (kNN) in which the class label is determined as the majority label of the k most similar training examples (Cover and Hart, 1967).

We propose to identify nearest neighbors using composite features, i.e. instances that share the most similar combination of atomic features. We use a function to assign to each composite of atomic features a discrete distance value, effectively rendering the search for nearest neighbors as a search within a *Parzen* window (Parzen, 1962). The variable k represents the minimum support within this window that we require.

Figure 2 illustrates our proposed approach: To classify the underlined argument in the sample sentence, we search for nearest neighbors using composite features. A distance 1 composite feature consists of P+V+F and AL, the lemma of the argument head. Nearest neighbors at this distance are therefore all training instances in which "creditor" is a passive subject of TELL.01. As the diagram on the right hand side in Figure 2 shows, this finds only one match, labeled **A2**, which is below the minimum support k that we require. We therefore increase the window to distance 2, which relaxes the argument head lemma

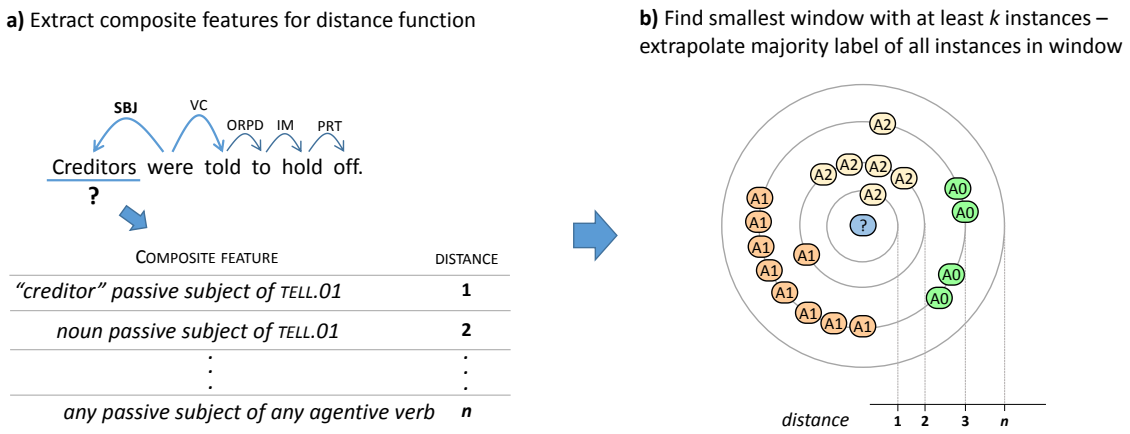

Figure 2: Example of argument labeling with nearest neighbor classification and a composite feature distance function.

restriction on composites features. Nearest neighbors at this distance are all passive subjects of TELL.01 in the training data. As the diagram shows, there are six nearest neighbors within a distance 2 window. From this neighborhood, we extrapolate the label **A2** as prediction.

Contributions We propose a simple and highly effective instance-based learning model for semantic role labeling[2]. We develop a *nearest k-window* variant of kNN in which we use a composite feature distance function to explicitly capture local contexts in sparse data. We give a full description of our SRL system, dubbed K-SRL, motivate and illustrate the atomic and composite features we use, and describe an easy-first algorithm for modeling global argument labeling constraints (Section 2). We present a detailed experimental evaluation that shows that our proposed approach significantly outperforms existing state-of-the-art systems (Section 3).

2 Instance-based Learning for SRL

We use instance-based learning for SRL as a sequence of two classification tasks: First, joint predicate identification and classification (referred to as *predicate labeling*), followed by joint argument identification and classification (referred to as *argument labeling*). See Figure 3 for an illustration. In this section, we describe the proposed nearest *k*-window classifier (Section 2.1) and discuss the specific features used (Sections 2.3 and 2.2), before presenting how we include global constraints into argument labeling (Section 2.4).

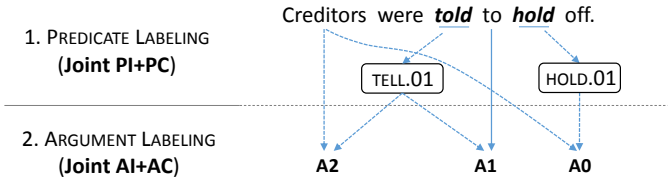

Figure 3: K-SRL system outline.

2.1 Nearest *k*-Window Classification

Algorithm 1 outlines our nearest *k*-window classification algorithm. It takes as input unlabeled instances and performs feature extraction. A distance function assigns each feature an integer distance value, with 1 as the closest distance. The search for nearest neighbors begins with a window of distance 1. The algorithm retrieves for each unlabeled sample all training examples whose distance from the current sample lie within this window. If the window contains fewer than *k* training instances, we increase the window size until it passes a threshold. Among the instances in the window, the algorithm determines

[2]In this work, we focus on verbal predicates since they are comprehensively and consistently labeled in available PropBank releases. We aim to revisit SRL for other types of predicates once current efforts to consistently annotate noun predicates, light verb constructions and adjectives (Bonial et al., 2014) are completed.

the relative frequencies of each label. For instance, if the window contains 10 similar instances of which 9 are labeled **A0** and one is labeled **A1**, the relative frequencies of **A0** and **A1** are 90% and 10% respectively. We interpret this relative frequency as a measure of *confidence*. The label with the highest relative frequency is returned as the most confident prediction. A label is returned only if its associated confidence is higher than a threshold (denoted by θ). If either insufficient examples in the nearest neighborhood are found, or the associated confidence is below the threshold, we return a fallback label. For the subtask of predicate labeling, the fallback label is the most common sense of a verb. For argument labeling, the fallback is to not label the word as an argument.

Algorithm 1 Nearest k-Window Classification

```
ws ← 1
if ws ≤ maximal distance then
  for each x ∈ Unknown Sample do
    Add to I all y ∈ Training Set, where distance(x, y) ≤ ws
    if |I| ≥ k then
      Determine majority class label I
      if the relative frequency of the label for x ≥ θ then
        Return I, the label and its relative frequency for x
      end if
    else
      ws = ws + 1
    end if
  end for
end if
```

2.2 Features for Instance-based Predicate Labeling

Atomic Features The Proposition Bank distinguishes different frame options for a verb based on syntactic subcategorization and coarse-grained polysemy (Palmer et al., 2005). For instance, the verb *return* may evoke the frames RETURN.01 (*return to a place*, as in *John returned to Boulder*) and RETURN.02 (*return an item to someone*, as in *John returned the stapler to Mary*). The key difference of the two in subcategorization is that RETURN.02 may take a syntactic object while RETURN.01 may not. Besides objects, other differentiators in subcategorization involve particles, complements and prepositional objects mediated by different prepositions.

We use each component of a subcategorization frame as one feature: The subject lemma S, the verb lemma VB, the verb particle VP, the object lemma O and the prepositional object PP . Each of these features may also be empty if unobserved. In addition to such lexical features, we define a set of binary features that indicate whether a subcategorization frame component is observed or not: S? for subjects, O? for objects and PP? for prepositional objects. Finally, we use the verb voice V since some frames are more commonly observed in passive voice.

Composite Feature The set of atomic features described above constructs one single composite feature, denoted as F_x for a given instance x. The distance between x_{test} (test instance) and x_{train} (training instance) corresponds to the total number of non-empty features that the two instances do not share, defined as $d(x, y) = |F_{x_{test}} \cup F_{x_{train}}| - |F_{x_{test}} \cap F_{x_{train}}|$. This distance function in essence corresponds to Jaccard distance, without normalizing to a value between 0 and 1.

2.3 Features for Instance-based Argument Labeling

Atomic Features Table 2 includes a list of atomic features. Besides four well-established features from previous work (i.e. the predicate frame F, the predicate voice V, the argument head lemma AL and the argument head PoS tag AP), we define the following two novel atomic features:
(1) *Syntactic-Semantic Path* P: A variant of the syntactic path that, instead of traversing only the syntactic tree, traverses semantic arcs from preceding classifications whenever possible. It is designed for the

FEATURE	SHORTHAND
Predicate frame	F
Predicate frame class	FC
Predicate voice	V
Syntactic-semantic path	P
Argument head lemma	AL
Argument head pos	AP

Table 2: Argument labeling features, with novel features in bold.

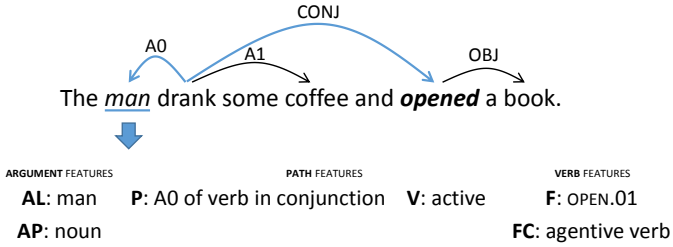

Figure 4: Features extracted for the underlined word (*man*) with regards to the predicate OPEN.01.

COMPOSITE	DISTANCE	EXAMPLE
AL+V+P+F	1	"man" ∧ A0 of verb in conjunction ∧ active ∧ OPEN.01
AP+V+P+F	2	any noun ∧ A0 of verb in conjunction ∧ active ∧ OPEN.01
V+P+F	3	any word ∧ A0 of verb in conjunction ∧ active ∧ OPEN.01
AL+V+P+FC	4	"man" ∧ A0 of verb in conjunction ∧ active ∧ agentive verb
AL+V+P+FC	5	any noun ∧ A0 of verb in conjunction ∧ active ∧ agentive verb
V+P+FC	6	any word ∧ A0 of verb in conjunction ∧ active ∧ agentive verb

Table 3: Distance value assigned to each composite feature, with an example for features extracted in Figure 4.

phenomenon of *raised* arguments, defined as syntactic constituents of a preceding verb. For instance, in Figure 4, *man* is a constituent of the verb *drank* as well as a raised argument for the verb *open*. In this example, we build the path from *man* to *open* by first traversing the semantic arc (A0) from *man* to *drank* and then the syntactic arc (CONJ) from *drank* to *open*. The resulting syntactic-semantic path is verbalized as "A0 of verb in conjunction".
(2) *Frame Class* FC: This feature categorizes frames based on whether they may take a thematic *agent* as argument. Some frames, such as FESTER.01 and HOVER.01, cannot and are therefore considered non-agentive. Non-agentive frames define no **A0** and therefore realize semantic roles differently[3].

Composite Features Table 3 lists all composite features along with their associated distances, including features extracted for the sample sentence in Figure 4. As described in Section 1, the relevant context for argument labeling includes: 1) The syntactic-semantic relationship between predicate and argument (P+V), 2) The frame-specific context of this syntactic-semantic relationship (F or FC), and 3) The argument-specific context (AL or AP). We require each of the three components to be represented in each composite feature in order to capture argument contexts. We define a distance function that assigns the closest distance 1 to the most discriminative of these composite features (i.e. AL+P+V+F). The function assigns higher distances to composites with fewer or less specific features (AP is a less specific representation of the argument context than Al).

This approach draws inspiration from *backoff* modeling, a well known method for addressing sparsity in language modeling with n-grams (Katz, 1987): If insufficient training data exists, such models commonly backoff to lower histories (for instance, a 3-gram model may back off to a 2-gram language model). The six distance values assigned to composite features in Table 3 may be interpreted in a similar spirit since our approach broadens the search to nearest neighbors with less specific composite features if insufficient training data exists.

2.4 Easy-First Argument Labeling

While argument labeling decisions are made locally, each core semantic role (labels **A0** through **A5**) may only be assigned once per predicate (Che et al., 2009). To include this global constraint, we use a greedy approach in which already assigned core labels are removed from consideration for the remaining predictions. Our approach follows an easy-first philosophy (Goldberg and Elhadad, 2010) where clas-

[3]For example, while active subjects are most commonly labeled **A0** (agent) of a verb ("the dog ate", "the bird sang"), they are typically the **A1** (theme) of non-agentive frames ("the wound festered").

Algorithm 2 Easy-First Argument Labeling

```
for each predicate p ∈ Unknown Sample do
  A ← ∅
  C ← Candidate arguments of p, their labels and confidence value in sorted order by confidence
  for c ∈ C with the highest confidence value in C do
    Remove c from C
    if Label of c ∉ Set of labels in A then
      Add c to A
    end if
  end for
end for
```

sifications for all predicate arguments are ordered by confidence and highest confidence predictions are made first. Algorithm 2 outlines this approach.

3 Experiments

In this section we evaluate K-SRL, our proposed instance-based labeling approach for SRL. We conduct a comparison study to evaluate its performance against previously published state-of-the-art systems. We also examine how different parameters of K-SRL impact its performance, including the minimum support variable *k*, different components in composite features, and our interpretation of relative label frequencies in the nearest neighborhood as an indication of confidence for assigning labels.

3.1 Experimental Setup

We use the benchmark data sets from the CoNLL-2009 shared task (Hajič et al., 2009) and compare our results against the top two scoring systems of the CoNLL-2009 shared task as well as two recent state-of-the-art systems: 1) CHEN (Zhao et al., 2009), which uses maximum entropy classifiers. 2) CHE (Che et al., 2009), which uses SVM classifiers. 3) MATEPLUS (Roth and Woodsend, 2014a), a state-of-the-art extension of a previous system (Björkelund et al., 2009) that uses logistic regression classifiers and word embeddings. 4) PATHLSTM (Roth and Lapata, 2016), the current state-of-the-art which uses logistic regression classifiers for predicates and neural network models with word embeddings for arguments. Our default settings for K-SRL are $k = 3$ and confidence threshold $\theta = 0$, both determined through experimentation. For both settings, we present parameter sweep experiments.

We compute the precision, recall and F_1 to measure the quality of the systems. In our study, we focus on verbal predicates and their roles, which we evaluate using the scoring metric of the CoNLL-2009 shared task. We recomputed the measures for the previous state-of-art systems using their published results to focus on verbal predicates and their roles[4].

3.2 Evaluation Results

Tables 4 and 5 summarize the results for our comparison study on in-domain and out-of-domain data respectively. As can be seen, K-SRL outperforms all previous approaches by a significant margin on both data sets. In the in-domain setting, K-SRL achieves 89.28% F_1-score, outperforming PATHLSTM, the currently published state-of-the-art approach, by 1.1 percentage points. In the out-domain setting, K-SRL achieves 79.91% F_1-score, outperforming MATEPLUS, the best previous system on out-of-domain data in our evaluation, by over 3 percentage points, and PATHLSTM by an even larger margin.

3.3 Additional Experiments

We conduct a detailed empirical examination to evaluate different aspects of our approach and make the following observations:

[4] As a result, the numbers for previous work reported here are slightly different from the published numbers.

SYSTEM	PRECISION	RECALL	F_1
CHE	87.43	83.92	85.64
CHEN	88.45	84.22	86.28
MATEPLUS	89.59	86.07	87.79
PATHLSTM	90.24	86.24	88.19
K-SRL	**91.21**	**87.42**	**89.28**
K-SRL$_{local}$	90.19	87.15	88.64
K-SRL$_{(-AL)}$	90.33	86.55	88.4
K-SRL$_{(-F)}$	88.74	85.11	86.89
K-SRL$_{(-FC)}$	91.17	87.53	89.31

Table 4: Evaluation results on in-domain data.

SYSTEM	PRECISION	RECALL	F_1
CHE	76.25	71.24	73.66
CHEN	78.1	71.64	74.73
MATEPLUS	79.46	74.21	76.74
PATHLSTM	79.92	73.31	76.47
K-SRL	**82.09**	**77.84**	**79.91**
K-SRL$_{local}$	80.38	77.78	79.06
K-SRL$_{(-AL)}$	81.69	77.28	79.42
K-SRL$_{(-F)}$	80.96	76.88	78.86
K-SRL$_{(-FC)}$	81.69	77.71	79.65

Table 5: Evaluation results on out-of-domain data.

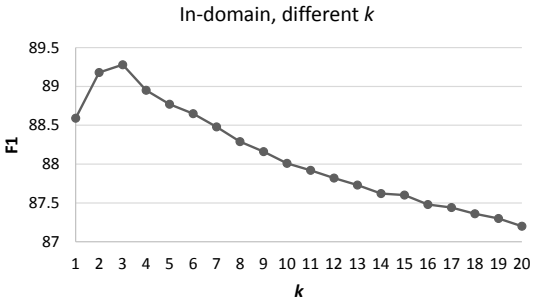

Figure 5: Parameter sweep over k on in-domain data.

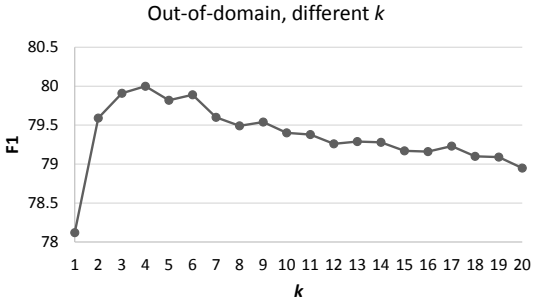

Figure 6: Parameter sweep over k on out-of-domain data.

Highest quality predictions from small neighborhoods. To determine the best setting for *k*, we conduct a parameter sweep experiment. Figures 5 and 6 summarize results of SRL for *k* from 1 to 20 on both in-domain and out-domain data. The results show that the best F_1 scores are achieved at relatively low settings for *k*, with best results obtained with $k = 3$ for in-domain and $k = 4$ for out-of-domain data. At higher *k*, F_1-score drops gradually, indicating that this approach loses its ability to capture local bias. At lower *k*, the approach overfits, decreasing F_1-score especially in the out-of-domain scenario (↓1.8 pp). These observations confirm our initial conjecture that SRL is affected by a strong local bias and that a small nearest neighborhood suffices to make high quality predictions.

Global constraints improve argument labeling. We run an ablation test in which we make only local predictions without modeling global constraints as described in Section 2.4, which reduces the F_1-score by .6 and .8 percentage points respectively on in-domain and out-of-domain data (see K-SRL$_{local}$ in Tables 4 and 5). These results are in line with previous evaluations on the impact of modeling global argument constraints (Toutanova et al., 2008; Roth and Lapata, 2016).

Frame and argument contexts are important. To assess the importance of individual features in their contexts, we run ablation tests in which we remove individual atomic features from composites, as summarized in Tables 4 and 5. Specifically, removing the frame feature F from argument labeling (K-SRL$_{(-F)}$)), which causes all argument labeling predictions to be made without frame-specific contests, leads to the most significant decrease on F_1 scores (↓2.5 pp and ↓1 pp) in our ablation tests. Omitting argument head lemma feature AL (K-SRL$_{(-AL)}$), the only feature capturing argument-level selectional preference (Resnik, 1997) in our approach, results in evident reduction on F_1 scores (↓0.8 pp and ↓0.5 pp). Meanwhile, the removal of frame class feature (K-SRL$-FC$) impacts only the out-of-domain scenario slightly (↓0.3 pp). This observation indicates that small neighborhoods with the frame feature often suffice to capture exceptions for non-agentive verbs.

Relative frequencies measures confidence. We assess our interpretation of relative frequencies in the nearest neighborhood as a measure of confidence by running a parameter sweep over θ. The results are depicted in Figures 7 and 8. As can be seen, precision improves at higher θ, while recall decreases, indicating that the quality of label prediction positively correlates with the associated confidence. We measure the best F_1-scores at $\theta = 0.5$ and $\theta = 0.6$ respectively, but F_1-score remains relatively stable between $\theta = 0.0$ and $\theta = 0.7$, indicating a balanced trade-off within these parameters. These observa-

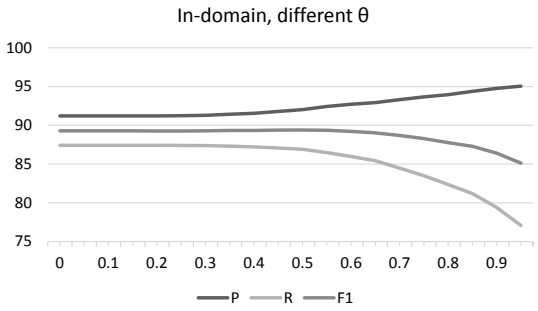

Figure 7: Parameter sweep over θ on in-domain data.

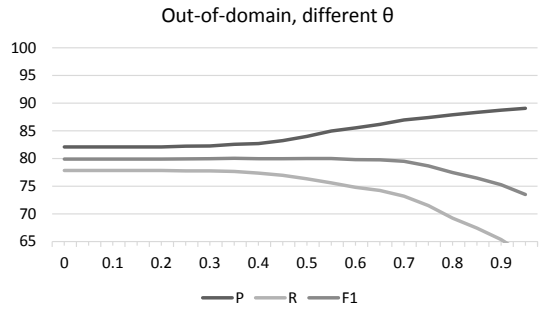

Figure 8: Parameter sweep over θ on out-of-domain data.

tions indicate that relative frequencies can serve as a good measure of confidence and be used to influence the precision-recall trade-off.

3.4 Discussion

Our instance-based learning approach is designed to capture the strong local bias of SRL. We note that even in out-of-domain evaluation scenarios, a very small nearest neighborhood suffices to make high quality predictions. Our experimental results demonstrate the effectiveness of this approach compared to previous state-of-the-arts for both in-domain and out-of-domain scenarios.

The state-of-the-art results are also remarkable in light of the relatively simple feature set we used. While previous work has investigated the use of word clusters (Choi and Palmer, 2011), word embeddings (Roth and Woodsend, 2014b; Roth and Lapata, 2016) and explicit learning of selectional preference (Zapirain et al., 2013) for better generalization across the training data, such features are absent in our current approach. Instead, for predicate labeling we use only the subcategorization frame and for argument labeling a simple set of 6 basic atomic features. This is in stark contrast to previous works that often employ dozens of different features classes (Johansson and Nugues, 2008; Björkelund et al., 2009; Choi and Palmer, 2011).

Interpretability of classification decisions. Our approach has the advantage of interpretability since each classification is determined through a specific composite feature that can be translated into a human readable statement (as illustrated in Table 3). This enables us to easily understand classification results and debug misclassifications, and thus facilitates the process of defining atomic features and composites for SRL. At the same time, we note that explicitly modeling composites does add another layer of complexity in feature engineering to this task. We plan to further investigate this in future work.

4 Conclusion and Outlook

We introduced an instance-based learning approach for semantic role labeling that is designed to address the large number of low-frequency exceptions in training data. We proposed to construct composites based on a few existing well-known features to identify similar instances. Our experimental results indicates that our model built on top of this approach significantly outperform existing systems, leading to new state-of-the-art results in SRL for verbal predicates and their roles.

We intend to focus more specifically on feature engineering for instance-based SRL. In particular, we plan to explore automatic feature selection methods especially in the context of composite features. We also plan to evaluate generalization features such as word clusters or word embeddings in the context of instance-based SRL.

Finally, we plan to extend our system to different types of predicates including nouns and complex predicates (Bonial et al., 2014), as well as evaluate its applicability to SRL in different languages (Xue and Palmer, 2005).

References

[Aha et al.1991] David W Aha, Dennis Kibler, and Marc K Albert. 1991. Instance-based learning algorithms. *Machine learning*, 6(1):37–66.

[Björkelund et al.2009] Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 43–48. Association for Computational Linguistics.

[Bonial et al.2014] Claire Bonial, Julia Bonn, Kathryn Conger, Jena D Hwang, and Martha Palmer. 2014. Propbank: Semantics of new predicate types. In *LREC*, pages 3013–3019.

[Che et al.2009] Wanxiang Che, Zhenghua Li, Yongqiang Li, Yuhang Guo, Bing Qin, and Ting Liu. 2009. Multilingual dependency-based syntactic and semantic parsing. In *Proceedings of the thirteenth conference on computational natural language learning: shared task*, pages 49–54. Association for Computational Linguistics.

[Choi and Palmer2011] Jinho D Choi and Martha Palmer. 2011. Transition-based semantic role labeling using predicate argument clustering. In *Proceedings of the ACL 2011 Workshop on Relational Models of Semantics*, pages 37–45. Association for Computational Linguistics.

[Cover and Hart1967] Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

[Daelemans and Van den Bosch2005] Walter Daelemans and Antal Van den Bosch. 2005. *Memory-based language processing*. Cambridge University Press.

[Daelemans et al.1999] Walter Daelemans, Sabine Buchholz, and Jorn Veenstra. 1999. Memory-based shallow parsing. In *Proceedings of the Third Conference on Computational Natural Language Learning*.

[Fader et al.2011] Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.

[Goldberg and Elhadad2010] Yoav Goldberg and Michael Elhadad. 2010. An efficient algorithm for easy-first non-directional dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 742–750. Association for Computational Linguistics.

[Hajič et al.2009] Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18. Association for Computational Linguistics.

[Johansson and Nugues2008] Richard Johansson and Pierre Nugues. 2008. Dependency-based semantic role labeling of propbank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 69–78. Association for Computational Linguistics.

[Katz1987] Slava Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401.

[Kipper et al.2008] Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1):21–40.

[Lo et al.2013] Chi-kiu Lo, Meriem Beloucif, and Dekai Wu. 2013. Improving machine translation into chinese by tuning against chinese meant. In *Proceedings of the Tenth International Workshop on Spoken Language Translation (IWSLT 2013)*.

[Maqsud et al.2014] Umar Maqsud, Sebastian Arnold, Michael Hülfenhaus, and Alan Akbik. 2014. Nerdle: Topic-specific question answering using wikia seeds. In *COLING (Demos)*, pages 81–85.

[Nivre et al.2004] Joakim Nivre, Johan Hall, and Jens Nilsson. 2004. Memory-based dependency parsing. In *Proceedings of the Eighth Conference on Computational Natural Language Learning*.

[Palmer et al.2005] Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

[Parzen1962] Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

[Resnik1997] Philip Resnik. 1997. Selectional preference and sense disambiguation. In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How*, pages 52–57. Washington, DC.

[Roth and Lapata2016] Michael Roth and Mirella Lapata. 2016. Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany. To appear.

[Roth and Woodsend2014a] Michael Roth and Kristian Woodsend. 2014a. Composition of word representations improves semantic role labelling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 407–413, Doha, Qatar, 25–29 October.

[Roth and Woodsend2014b] Michael Roth and Kristian Woodsend. 2014b. Composition of word representations improves semantic role labelling. In *EMNLP*, pages 407–413. ACL.

[Shen and Lapata2007] Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *EMNLP-CoNLL*, pages 12–21.

[Toutanova et al.2008] Kristina Toutanova, Aria Haghighi, and Christopher D Manning. 2008. A global joint model for semantic role labeling. *Computational Linguistics*, 34(2):161–191.

[van der Plas et al.2014] Lonneke van der Plas, Marianna Apidianaki, Rue John von Neumann, and Chenhua Chen. 2014. Global methods for cross-lingual semantic role and predicate labelling.

[Veenstra et al.2000] Jorn Veenstra, Antal Van den Bosch, Sabine Buchholz, Walter Daelemans, et al. 2000. Memory-based word sense disambiguation. *Computers and the Humanities*, 34(1-2):171–177.

[Xue and Palmer2005] Nianwen Xue and Martha Palmer. 2005. Automatic semantic role labeling for chinese verbs. In *IJCAI*, volume 5, pages 1160–1165. Citeseer.

[Zapirain et al.2013] Benat Zapirain, Eneko Agirre, Lluis Marquez, and Mihai Surdeanu. 2013. Selectional preferences for semantic role classification. *Computational Linguistics*, 39(3):631–663.

[Zhao et al.2009] Hai Zhao, Wenliang Chen, Chunyu Kit, and Guodong Zhou. 2009. Multilingual dependency learning: A huge feature engineering method to semantic dependency parsing. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 55–60. Association for Computational Linguistics.