Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Complementing Semantic Roles with Temporally-Anchored Spatial Knowledge: Crowdsourced Annotations and Experiments

Alakananda Vempala and **Eduardo Blanco**
Human Intelligence and Language Technologies Lab
University of North Texas
Denton, TX 76203
AlakanandaVempala@my.unt.edu, eduardo.blanco@unt.edu

Abstract

This paper presents a framework to infer spatial knowledge from semantic role representations. We infer whether entities are or are not located somewhere, and temporally anchor this spatial information. A large crowdsourcing effort on top of OntoNotes shows that these temporally-anchored spatial inferences are ubiquitous and intuitive to humans. Experimental results show that inferences can be performed automatically and semantic features yield performance improvements.

1 Introduction

Extracting meaning from text has received considerable attention in the last decade. In particular, semantic role labeling has become popular, including both corpora development and automatic role labelers. Semantic roles capture semantic links between predicates and their arguments; they capture who did what to whom, how, when and where.

There are several corpora with semantic role annotations. FrameNet (Baker, Fillmore, and Lowe 1998) annotates frame elements (semantic roles) defined in semantic frames, which are triggered by lexical units. PropBank (Palmer, Gildea, and Kingsbury 2005) and NomBank (Meyers et al. 2004) annotate semantic roles for verbal and nominal predicates respectively. More recently, OntoNotes (Hovy et al. 2006) includes PropBank-style semantic roles. Semantic role labelers trained with PropBank have matured in the last decade (Carreras and Màrquez 2005; Zhou and Xu 2015), with state-of-the-art F-measures around 0.81.

While semantic roles encode useful information, there is much more meaning in all but the simplest statements. Consider the sentence *John drove to San Francisco for a doctor's appointment* and the semantic roles annotated in OntoNotes (Figure 1, solid arrows). On top of these valuable roles, one can infer that *John* had LOCATION *San Francisco* for a relatively short period of time after *drove* (more precisely, during the *doctor's appointment*), but probably not long after, long before or during *drove*. This additional knowledge is intuitive to humans, even though it is disregarded by existing tools and highly ambiguous: if *John drove home to San Francisco after a vacation in Colorado*, it is reasonable to believe that he had LOCATION *San Francisco* well after *drove*,

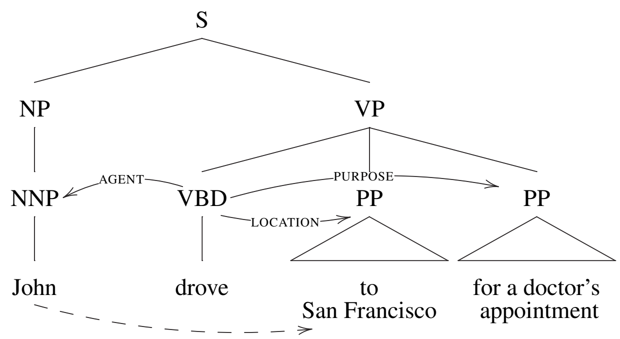


Figure 1: Semantic roles in OntoNotes (solid arrows) and additional spatial knowledge (dashed arrow).

i.e., he did not leave *San Francisco* shortly after *drove* took place because he lives in *San Francisco*.

This paper presents a framework to infer temporally-anchored spatial knowledge from semantic roles. The main contributions are: (1) analysis of missing spatial knowledge in OntoNotes; (2) crowdsourced annotations on top of OntoNotes;[1] (3) experimental results detailing results with gold-standard and predicted linguistic annotations, and using lexical, syntactic and semantic features.

2 Semantic Roles and Additional Spatial Knowledge

We represent a semantic relation R between x and y as R(x, y). R(x, y) can be read "x has R y", e.g., AGENT(*bought*, *Bill*) can be read "*bought* has AGENT *Bill*." Semantic roles are relations R(x, y) such that (1) x is a predicate and (2) y is an argument of x. In this paper, we work on top of OntoNotes semantic roles, which only account for verbal predicates, i.e., for all semantic roles R(x, y), x is a verb.

We use the term *additional spatial knowledge* to refer to relations LOCATION(x, y) such that (1) x is not a predicate or (2) x is a predicate and y is not an argument of x. In other words, additional spatial knowledge is spatial meaning not captured with semantic roles. As we shall see, the framework presented here not only infers plain LOCATION(x, y), but also temporally anchors this additional knowledge.

[1] Available at http://hilt.cse.unt.edu/

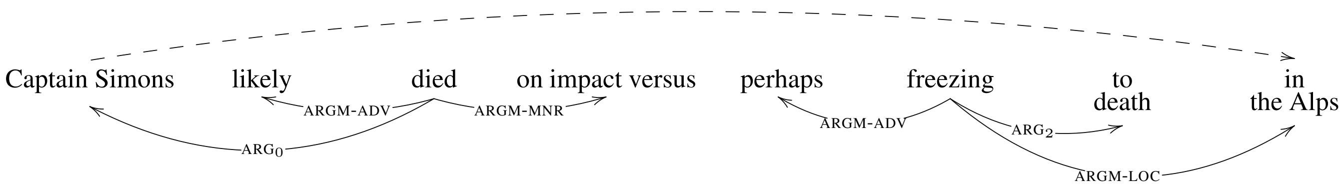


Figure 2: Semantic roles in OntoNotes (solid arrows) and additional spatial knowledge of type (1b) (dashed arrow).

2.1 Semantic Roles in OntoNotes

OntoNotes is a large corpus with 1,302,342 tokens and 63,918 sentences from several genres including newswire, broadcast news and conversations, magazines and the web.[2] It includes POS tags, word senses, parse trees, speaker information, named entities, semantic roles and coreference.

OntoNotes semantic roles follow PropBank framesets. It uses a set of numbered arguments (ARG_0–ARG_5) whose meanings are verb-dependent, and argument modifiers which share a common meaning across verbs (ARGM-LOC, ARGM-TMP, ARGM-MNR, ARGM-PRP, ARGM-CAU, etc.). For a detailed description of the semantic roles used in OntoNotes, we refer the reader to the LDC catalog[3] and PropBank (Palmer, Gildea, and Kingsbury 2005).

Throughout this paper, semantic roles are drawn with solid arrows. To improve readability, we often rename numbered arguments, e.g., AGENT instead of ARG_0 in Figure 1.

2.2 Additional Spatial Knowledge

OntoNotes semantic roles only capture a portion of all spatial knowledge. They capture locations of verbal predicates with (1) ARGM-LOC for all verbs, and (2) numbered arguments for a few verbs, e.g., the start and end point of *go.01* are encoded with ARG_3 and ARG_4.

There are 2 types of additional relations LOCATION(x, y): (1) those whose arguments x and y are semantic roles of some verb, and (2) those whose arguments x and y are not semantic roles of any verb. Type (1) can be further divided into type (1a) if x and y are roles of the same verb, and type (1b) if x and y are roles of different verbs.

Figure 1 exemplifies an inference of type (1a): *drove* has AGENT *John* and LOCATION *San Francisco*, the additional spatial knowledge between *John* and *San Francisco* is inferred between roles of the same verb. Figure 2 presents an inference of type (1b): *died* has ARG_0 *Captain Simons* and *freezing* has ARGM-LOC *in the Alps*, the additional relation LOCATION(*Captain Simons*, *in the Alps*) links roles of different verbs: ARG_0 of *died* and ARGM-LOC of *freezing*.

The following statement exemplifies type (2): *[Palm Beach estate owners]*$_{\text{AGENT}}$ *drive [Bentleys and other luxury cars]*$_{\text{THEME}}$. Semantic roles indicate the AGENT and THEME of *drive*; additional spatial knowledge includes LOCATION(*Bentleys and other luxury cars*, *Palm Beach*). Note that the AGENT is *estate owners*, and that *Palm Beach* indicates their location—it is not an argument of *drive*.

[2]We use the CoNLL-2011 Shared Task distribution (Pradhan et al. 2011), available at http://conll.cemantix.org/2011/.

[3]https://catalog.ldc.upenn. edu/LDC2013T19

foreach *sentence s* **do**
  foreach *semantic role* ARGM-LOC(y_{verb}, y) $\in s$ **do**
    foreach *semantic role* ARG_i(x_{verb}, x) $\in s$ **do**
      if *is_valid(x, y)* **then**
        generate potential relation LOCATION(x, y)
      end
    end
  end
end

Algorithm 1: Procedure to generate all potential additional spatial knowledge targeted in this paper.

3 Corpus Creation

Annotating all additional spatial knowledge in OntoNotes is outside the scope of this paper. We focus on additional relations LOCATION(x, y) of type (1) (Section 2.2) such that ARG_i(x_{verb}, x) and ARGM-LOC(y_{verb}, y) exist, i.e., x is a numbered role (ARG_0–ARG_5) of some verb x_{verb} and y is ARGM-LOC of some verb y_{verb} (x_{verb} and y_{verb} need not be the same). We also enforce that:

1. x and y belong to the same sentence and do not overlap;
2. the head of x is a noun and one of these named entity types: `fac`, `gpe`, `loc`, or `org`;[4] and
3. the head of y is a noun subsumed by *physical_entity* in WordNet, or one of these named entity types: `person`, `org`, `work_of_art`, `fac`, `norp`, `product` or `event`.

These restrictions are designed to reduce the annotation effort and automatically generate the least amount of invalid potential additional spatial knowledge. For example, locations that have as head an adverb (*here*, *there*, etc.) are unlikely to grant inferences. Similarly, it is almost surely the case that neither x nor y can be a named entity such as `date`, `percent` or `cardinal`. All potential additional spatial knowledge targeted in this paper is generated with Algorithm 1; *is_valid(x, y)* enforces the above restrictions. The number of potential LOCATION relations generated is 1,732.

3.1 Crowdsourcing Annotations

Once potential relations LOCATION(x, y) are generated with Algorithm 1, they must be validated. After pilot in-house annotations, it became clear that it is suboptimal to (1) ask whether *x is located at y*, and (2) force annotators to answer `YES`, `NO` or `UNKNOWN`. First, unlike objects such as *bridges* and *houses*, most entities change their location frequently;

[4]For a description and examples of these named entity types, refer to (Weischedel and Brunstein 2005).

	certYES		probYES		certNO		probNO		UNK		INV		Maj. Label	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Day Before	481	27.77	200	11.54	589	34.00	145	8.37	94	5.42	223	12.87	1311	75.69
During	1066	61.54	61	3.52	293	16.91	44	2.54	56	3.23	212	12.24	1424	82.21
Day After	647	37.35	191	11.02	436	25.17	141	8.14	99	5.71	218	12.58	1293	74.65
All	2194	42.22	452	8.69	1318	25.36	330	6.35	249	4.79	653	12.56	4028	77.52

Table 1: Label counts per temporal anchor, and number of questions with a majority label in the crowdsourced annotations.

considering temporally anchored spatial knowledge is intuitive. Second, while there is often evidence that something is (or is not) located somewhere, it is difficult to fully commit.

Based on these observations, we first generate three questions for each potential LOCATION(*x*, *y*):

1. Is *x* located at *y* the day before y_{verb}?
2. Is *x* located at *y* during y_{verb}?
3. Is *x* located at *y* the day after y_{verb}?

Then, we allow annotators to answer from six labels inspired by previous work (Saurí and Pustejovsky 2012):

- `certYES`: I am certain that the answer is yes.
- `probYES`: The answer is probably yes, but I am not sure.
- `certNO`: I am certain that the answer is no.
- `probNO`: The answer is probably no, but I am not sure.
- `UNK`: There is not enough information to answer.
- `INV`: The question is invalid.

Annotations were gathered using Amazon Mechanical Turk. We created Human Intelligence Tasks (HITs) consisting of the three questions regarding a potential additional LOCATION(*x*, *y*). The only information available to annotators was the sentence from which the additional LOCATION(*x*, *y*) was generated, they did not see semantic role information, the previous or next sentence, etc. Following previous work (Callison-Burch and Dredze 2010), we recruited annotators with previous approval rate $\geq$ 90% and past approved HIT count over 5,000. We also discarded submissions that took unusually short time compared to other submissions, and work done by annotators who always chose the same label. Workers received $0.03 per HIT. We requested 5 annotations per HIT. 150 annotators participated in the task, on average they annotated 57.33 HITs (minimum number of HITs per annotator: 1, maximum: 1,409). We assigned the final answer to each question by calculating the majority label among all annotations.

3.2 Annotation Analysis

Columns 2–13 in Table 1 summarize the counts for each label. Overall, 42.22% of questions are answered with `certYES` and 25.36% with `certNO`, i.e. 67.58% of potential additional spatial knowledge can be inferred with certainty (annotators are sure that *x* is or is not located at *y*). Percentages for `probYES` and `probNO` are substantially lower, 8.69% and 6.35% respectively. It is worth noting that 61.54% of questions for *during* temporal anchor are answered with `certYES`. This is due to the fact that some events (almost always) require their participants to be at the LOCATION of the event *during* the event, e.g., participants in meetings. The last column in Table 1 indicates the percentage of questions for which a majority label exists. The percentage is larger for *during* questions (82.21%), as they are easier to annotate, and 77.52% overall.

	Pearson	% of annotators that agree			
		≥ 5	≥ 4	≥ 3	≥ 2
Day Before	0.80	2.9	15.3	54.9	98.4
During	0.87	12.4	35.1	68.4	98.6
Day After	0.79	3.4	16.3	52.5	98.5
All	0.83	6.2	22.2	58.6	98.5

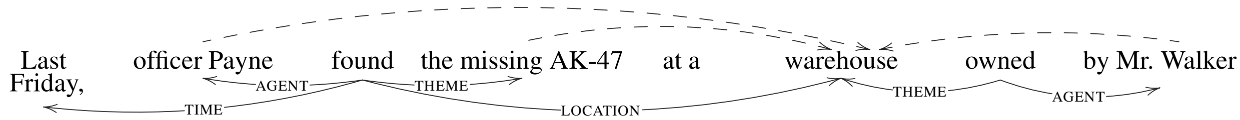
Table 2: Pearson correlations between crowdsourced and control annotations, and percentage of instances for which at least 5, 4, 3 and 2 annotators agree (out of 5 annotators).

Top 20 most certain verbs
leave explode begin march stand bear teach discuss arrest discover carry receive raise bury establish appear live die base open
Top 20 least ambiguous verbs
hear hire begin lead bear locate march conduct call receive bury provide attack retire lock draw teach base execute stop

Table 3: Top 20 most certain verbs (i.e., with the most `certYES` and `certNO` annotations) and top 20 least ambiguous verbs (i.e., with highest inter-annotator agreements).

In order to ensure quality, we manually annotated 10% of questions in each genre, and calculated Pearson correlations with the majority label after mapping labels as follows: `certYES`: 2, `probYES`: 1, `certNO`: -2, `probNO`: -1, `UNK`: 0, `INV`: 0. Overall correlation is 0.83 (Table 2), and *during* questions show a higher correlation of 0.87. Correlations per genre (not shown) are between high 0.70s and mid 0.80s, i.e., all genres achieved high agreements. We also calculated the raw inter-annotator agreements (Table 2). At least 3 annotators agreed (perfect label match) in 58.6% of questions and at least 2 annotators in 98.5%. Note that Pearson correlation is a better indicator of agreement, since not all label mismatches are the same, e.g., `certYES` vs. `probYES` and `certYES` vs. `certNO`. Also, a majority label may exists even if only 2 annotators agree (last column Table 1), e.g., {`probYES`, `UNK`, `INV`, `probYES`, `certYES`}.

Finally, Table 3 indicates the top 20 most certain verbs, i.e., with the highest ratio of `certYES` and `certNO` labels, and the top 20 least ambiguous verbs, i.e., with the overall highest inter-annotator agreements.



x	y	y_{verb}	Day Before	During	Day After
officer Payne	warehouse	found	probYES	certYES	probNO
the missing AK-47	warehouse	found	certYES	certYES	certNO
Mr. Walker	warehouse	found	UNK	UNK	UNK

Figure 3: Semantic roles in OntoNotes (solid arrows) and additional spatial knowledge annotations (dashed arrows).

Type	No.	Name	Description
	0	temporal tag	are we predicting the LOC(x, y) a day before, during or a day after y_{verb}?
Lexical	1–4	first word, POS tag	first word and POS tag in x and y
	5–8	last word, POS tag	last word and POS tag in x and y
Syntactic	9, 10	syntactic node	syntactic node of x and y
	11	common subsumer	syntactic node subsuming x and y
Semantic	12–15	predicate, POS tag	word surface form and POS tag of x_{verb} and y_{verb}
	16	same predicate	whether x_{verb} and y_{verb} are the same token
	17	ARGM-LOC count	number of ARGM-LOC semantic roles in the sentence
	18	ARGM-TMP count	number of ARGM-TMP semantic roles in the sentence
	19, 20	NE type	named entity types of head of x and y, if any

Table 4: Lexical, syntactic and semantic features to infer potential additional relation LOCATION(x, y).

3.3 Annotation Examples

Figure 3 presents a sample sentence with the semantic role annotations in OntoNotes (solid arrows) and all potential additional spatial knowledge generated (dashed arrows) along with the annotations. This sentence has 4 semantic roles for verb *found* (TIME: *Last Friday*, AGENT: *officer Payne*, THEME: *the missing AK-47*, and LOCATION: *warehouse*), and 2 semantic roles for verb *owned* (THEME: *warehouse*, and AGENT: *Mr. Walker*).

Annotators were asked to determine whether *officer Payne*, *the missing AK-47* and *Mr. Walker* are (or are not) located at the *warehouse* the day before, during and the day after *found*. Annotators interpreted that *officer Payne* was (1) certainly located at the *warehouse* during event *found* (certYES), (2) probably located there the day before (probYES), (3) and probably not located there the day after (probNO). In other words, they understood that a search took place at the *warehouse*, the search (probably) lasted a few days, *officer Payne* was at the *warehouse* daily until he found *the missing AK-47*, and then he (probably) didn't go back the day after. Regarding *the missing AK-47*, they annotated that the *AK-47* was certainly located at the *warehouse* the day before and during *found*, but not the day after (presumably, it was processed as evidence and moved away from the *warehouse*). Regarding *Mr. Walker*, they annotated that there is not enough evidence (UNK) to determine whether he was at the *warehouse*—property owners need not be located at their properties at any point of time.

4 Inferring Spatial Knowledge

We follow a standard supervised machine learning approach. Out of the 5,196 generated questions (1,732 LOCATION(x, y) $\times$ 3 temporal anchors), those with label INV were discarded, leaving 4,545 valid instances. We follow the CoNLL-2011 Shared Task (Pradhan et al. 2011) split into train, development and test. We trained an SVM model with RBF kernel using scikit-learn (Pedregosa et al. 2011). The feature set and parameters C and γ were tuned using 10-fold cross-validation with the train and development sets, and results are calculated using the test instances.

4.1 Feature Selection

Selected features (Table 4) are a combination of lexical, syntactic and semantic features extracted from words, POS tags, parse trees and semantic role representations. Our lexical and syntactic features are standard in semantic role labeling (Gildea and Jurafsky 2002) and thus we do not elaborate on them. We discarded many more well-known lexical and syntactic features that did not yield performance improvements during cross validation, e.g., path, subcategory, head.

Semantic features are derived from the verb-argument structures from which the potential additional relation LOCATION(x, y) was generated (Algorithm 1). Features 12–15 correspond to the surface form and part-of-speech tag of the verbs to which x and y attach (i.e., x_{verb} and y_{verb}). Feature 16 indicates whether x_{verb} and y_{verb} are the same, it differentiates between inferences of type (1a) and (1b). Features 17 and 18 are the number of ARGM-LOC and ARGM-TMP semantic roles in the sentence. Finally, features 19 and 20 are the named entity types, if any, of x and y.

Inspired by our previous work (Blanco and Vempala 2015), we tried additional semantic features, e.g., flags indicating semantic role presence, count for each semantic role attaching to x_{verb} and y_{verb}, numbered semantic role between x_{verb} and x, but discarded them because they did not improve performance during the tuning process.

System		All instances						Instances with majority label					
		DB	D	DA	All			DB	D	DA	All		
		F	F	F	P	R	F	F	F	F	P	R	F
most frequent per temporal anchor baseline	`certYES`	0.00	0.83	0.62	0.48	1.00	0.65	0.00	0.84	0.61	0.48	1.00	0.65
	`probYES`	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	`certNO`	0.59	0.00	0.00	0.00	0.00	0.00	0.74	0.00	0.00	0.00	0.00	0.00
	`probNO`	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	`UNK`	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	All	0.24	**0.58**	0.28	0.23	0.48	**0.31**	0.43	0.60	0.27	0.23	0.48	**0.31**
lexical features	`certYES`	0.34	0.83	0.57	0.59	0.75	0.66	0.23	0.86	0.61	0.68	0.78	0.73
	`probYES`	0.12	0.00	0.00	0.09	0.06	0.07	0.00	0.00	0.20	0.09	0.08	0.09
	`certNO`	0.58	0.21	0.46	0.48	0.52	0.50	0.75	0.36	0.58	0.65	0.64	0.64
	`probNO`	0.12	0.00	0.09	0.25	0.06	0.09	0.00	0.00	0.00	0.00	0.00	0.00
	`UNK`	0.14	0.00	0.17	0.50	0.06	0.11	0.00	0.00	0.00	0.00	0.00	0.00
	All	0.38	**0.61**	0.41	0.48	0.52	**0.48**	0.49	0.69	0.52	0.58	0.63	**0.60**
lexical + syntactic features	`certYES`	0.39	0.82	0.52	0.59	0.72	0.65	0.33	0.85	0.55	0.67	0.75	0.71
	`probYES`	0.08	0.00	0.00	0.06	0.03	0.04	0.00	0.00	0.25	0.12	0.08	0.10
	`certNO`	0.55	0.29	0.45	0.45	0.51	0.48	0.74	0.39	0.62	0.62	0.67	0.64
	`probNO`	0.11	0.00	0.09	0.20	0.06	0.09	0.00	0.00	0.00	0.00	0.00	0.00
	`UNK`	0.27	0.00	0.12	0.27	0.10	0.14	0.00	0.00	0.00	0.00	0.00	0.00
	All	0.38	**0.62**	0.38	0.45	0.51	**0.47**	0.50	0.69	0.51	0.57	0.63	**0.60**
lexical + syntactic + semantic features	`certYES`	0.41	0.82	0.62	0.66	0.74	0.69	0.23	0.85	0.67	0.69	0.80	0.74
	`probYES`	0.23	0.00	0.17	0.26	0.14	0.18	0.25	0.00	0.25	0.40	0.17	0.24
	`certNO`	0.57	0.19	0.48	0.46	0.54	0.50	0.75	0.24	0.71	0.66	0.69	0.67
	`probNO`	0.09	0.00	0.24	0.22	0.11	0.15	0.00	0.00	0.00	0.00	0.00	0.00
	`UNK`	0.24	0.00	0.30	0.31	0.16	0.21	0.00	0.00	0.00	0.00	0.00	0.00
	All	**0.41**	**0.61**	**0.48**	0.51	0.54	**0.52**	0.51	0.67	0.60	0.61	0.66	**0.63**

Table 5: Results obtained with the baseline, and using several combinations of features derived from gold-standard linguistic annotations. Results are provided per temporal anchor (DB: Day Before, D: During, DA: Day After).

5 Experimental Results

Table 5 presents results obtained with a baseline and several combinations of features using machine learning. Results are provided for all instances (Columns 3–8) and for instances for which a majority label exists (Columns 9–14). These results where obtained using gold-standard linguistic annotations for both generation of potential additional knowledge and feature extraction.

Overall, performance is better with instances for which a majority label exists (overall F-measure 0.63 vs. 0.52). This is not surprising, as these instances are easier to annotate manually. General performance trends per label, temporal anchor, and combinations of features are similar when using all instances and only instances with a majority label. The rest of this section describes results using all test instances.

The baseline simply predicts the most likely label for each question depending on the temporal anchor: `certYES` for *during* and *day after* and `certNO` for *day before* (Table 1). Overall F-measure is 0.31, but it is worth noting that the baseline obtains an F-measure of 0.58 for *during* instances.

The last block in Table 5 presents results using all features. Overall F-measure is 0.52, results are again better for *during* instances (0.61) than for *day before* (0.41) and *day after* (0.48). Results are higher for `certYES` and `certNO` (0.69 and 0.50 respectively) than for other labels (0.15–0.21). This is probably because most instances are labeled with `certYES` and `certNO` (Table 1). Better performance with these labels is desirable because they allow us to infer where entities are (and are not) located with certainty.

5.1 Feature Ablation

The bottom 3 blocks in Table 5 detail results using (1) lexical, (2) lexical and syntactic, and (3) lexical, syntactic and semantic features. Lexical features yield better performance than the baseline (0.48 vs. 0.31 overall F-measure), and including syntactic features does not have an impact (0.47). But considering lexical, syntactic and semantic features improves overall F-measure from 0.48 to 0.52.

Results for *during* instances are virtually the same with all feature combinations (lexical: 0.61, lexical and syntactic: 0.62, lexical, syntactic and semantic: 0.61). But results with all features for *day before* instances, and especially *day after* instances, is better (0.41 vs. 0.38 and 0.48 vs. 0.41).

During instances are the easiest to predict. As a matter of fact, lexical features alone perform as well as all features, and only slightly better than the baseline. Regarding labels, `certYES` and `certNO` are easier to predict with all feature combinations, and other labels (`probYES`, `probNO`, `UNK`) are the ones that benefit the most from complementing lexical features with syntactic and semantic features.

5.2 Gold-Standard vs. Predicted Linguistic Information

The last batch of results (Table 6) presents results using gold-standard and predicted linguistic annotations (POS tags, named entities, parse trees and semantic roles). Gold-standard and predicted annotations are used as present in the CoNLL-2011 Shared Task release (gold and auto files). All

	DB	D	DA	All		
	F	F	F	P	R	F
gold	0.41	0.61	0.48	0.51	0.54	0.52
predicted∩gold	0.45	0.54	0.58	0.60	0.58	0.55
predicted	0.25	0.33	0.29	0.58	0.20	0.29

Table 6: Results obtained with instances derived from (1) gold-standard annotations, (2) predicted annotations which are also in gold, and (3) predicted annotations.

experiments in this section are carried out using all features. Models are always trained with gold annotations, but tested with test instances generated as described below.

The evaluation presented in the first row (gold) is equivalent to the last row in Table 5: potential additional LOCATION(x, y) relations are generated using gold semantic roles and features are extracted from gold-standard linguistic annotations. The evaluation in the second row (gold ∩ predicted) generates potential additional LOCATION(x, y) relations using predicted semantic roles, but then filters overgenerated relations (i.e., those which are not generated from gold-standard roles). This system extracts features from predicted linguistic annotations, but as a result of the filtering, the number of test instances decreases from 444 to 155. The evaluation in the third row (predicted) generates potential additional LOCATION(x, y) from predicted semantic roles, and extracts features from predicted linguistic annotations.

Not surprisingly, *predicted* evaluation is the lowest: while precision is similar, recall suffers due to missing semantic roles in the predicted annotations, which unequivocally lead to potential spatial knowledge not being generated by Algorithm 1. The *gold* ∩ *predicted* evaluation may look surprisingly good, but the high F-measure is justified by the fact that it is restricted to the potential additional LOCATION(x, y) relations that are generated with both gold-standard and predicted semantic roles. Intuitively, roles are predicted more accurately in simpler sentences (shorter, without complex syntax), which in turn are also easier to infer from.

6 Related Work

Tools to extract the PropBank-style semantic roles we infer from have been studied for years (Carreras and Màrquez 2005; Hajič et al. 2009; Lang and Lapata 2010). These systems only extract semantic links between predicates and their arguments, not between arguments of predicates. In contrast, this paper complements semantic role representations with spatial knowledge for numbered arguments.

There have been several proposals to extract semantic links not annotated in well-known corpora such as Nombank (Meyers et al. 2004), FrameNet (Baker, Fillmore, and Lowe 1998) or PropBank (Palmer, Gildea, and Kingsbury 2005). Gerber and Chai (2010) augmented NomBank annotations with additional numbered arguments appearing in the same or previous sentences; Laparra and Rigau (2013) presented an improved algorithm for this task. The SemEval-2010 Task 10: Linking Events and their Participants in Discourse (Ruppenhofer et al. 2009) targeted cross-sentence missing numbered arguments in FrameNet and PropBank. Blanco and Moldovan (2014) inferred additional argument modifiers for verbs in PropBank. Unlike the framework presented in this paper, these previous efforts reveal implicit semantic links involving predicates. None of them infer semantic links between predicate arguments or target temporally-anchored spatial knowledge.

We have previously proposed an unsupervised approach that does not account for temporal anchors or uncertainty to infer semantic relations between predicate arguments (Blanco and Moldovan 2011). We have also presented preliminary experiments with 200 sentences following the framework presented here (Blanco and Vempala 2015).

Attaching temporal information to semantic relations is uncommon. In the context of the TAC KBP temporal slot filling track (Garrido et al. 2012; Surdeanu 2013), relations common in information extraction (e.g., SPOUSE, COUNTRY_OF_RESIDENCY) are assigned a temporal interval indicating when they hold. Unlike this line of work, the approach presented in this paper builds on top of semantic roles, targets temporally-anchored LOCATION relations, and accounts for uncertainty (e.g., `certYES` vs. `probYES`).

The task of spatial role labeling (Hajič et al. 2009; Kolomiyets et al. 2013) aims at thoroughly representing spatial information with so-called spatial roles, e.g., trajector, landmark, spatial and motion indicators, path, direction, distance, and spatial relations. Unlike us, the task does not consider temporal anchors or certainty. But as the examples throughout this paper show, doing so is useful because (1) spatial information does not hold for good for most entities and (2) humans sometimes can only state that it is probably the case that an entity is (or is not) located somewhere. In contrast to this task, we infer temporally-anchored spatial knowledge as humans intuitively understand it.

7 Conclusions

Semantic roles in OntoNotes capture semantic links between a verb and its arguments—they capture who did what to whom, how, when and where. This paper takes advantage of OntoNotes semantic roles in order to infer temporally-anchored spatial knowledge. Namely, we combine semantic roles within a sentence in order to infer whether entities are or are *not* located somewhere, and assign temporal anchors and certainty labels to this additional knowledge.

A crowdsourcing annotation effort shows that annotations can be done reliably by asking plain English questions to non-experts. Experimental results show moderate F-measure using gold-standard linguistic annotations (0.52), and relatively poor performance (0.29) in a more realistic scenario, when the additional spatial knowledge is inferred after extracting semantic roles automatically.

The essential conclusion of this paper is that semantic roles are a reliable semantic layer from which additional meaning can be inferred. While this paper focuses on temporally-anchored spatial knowledge, we believe that many more semantic relations (CAUSE, TIME, etc.) between arguments of verbs can be inferred using a similar strategy.

References

Baker, C. F.; Fillmore, C. J.; and Lowe, J. B. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, 86–90. Montreal, Quebec, Canada: Association for Computational Linguistics.

Blanco, E., and Moldovan, D. 2011. Unsupervised Learning of Semantic Relation Composition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1456–1465. Portland, OR: Association for Computational Linguistics.

Blanco, E., and Moldovan, D. 2014. Leveraging Verb-Argument Structures to Infer Semantic Relations. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 145–154. Gothenburg, Sweden: Association for Computational Linguistics.

Blanco, E., and Vempala, A. 2015. Inferring Temporally-Anchored Spatial Knowledge from Semantic Roles. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, 452–461. Denver, CO: Association for Computational Linguistics.

Callison-Burch, C., and Dredze, M. 2010. Creating Speech and Language Data With Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 1–12. Los Angeles, CA: Association for Computational Linguistics.

Carreras, X., and Màrquez, L. 2005. Introduction to the CoNLL-2005 Shared Task: Ssemantic Role Labeling. In *CONLL '05: Proceedings of the Ninth Conference on Computational Natural Language Learning*, 152–164. Morristown, NJ: Association for Computational Linguistics.

Garrido, G.; Peñas, A.; Cabaleiro, B.; and Rodrigo, A. 2012. Temporally Anchored Relation Extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL '12, 107–116. Jeju Island, Korea: Association for Computational Linguistics.

Gerber, M., and Chai, J. 2010. Beyond NomBank: A Study of Implicit Arguments for Nominal Predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1583–1592. Uppsala, Sweden: Association for Computational Linguistics.

Gildea, D., and Jurafsky, D. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics* 28(3):245–288.

Hajič, J.; Ciaramita, M.; Johansson, R.; Kawahara, D.; Martí, M. A.; Màrquez, L.; Meyers, A.; Nivre, J.; Padó, S.; Štěpánek, J.; Straňák, P.; Surdeanu, M.; Xue, N.; and Zhang, Y. 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*, CoNLL '09, 1–18. Boulder, CO: Association for Computational Linguistics.

Hovy, E.; Marcus, M.; Palmer, M.; Ramshaw, L.; and Weischedel, R. 2006. OntoNotes: the 90% Solution. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Semantics*, 57–60. Morristown, NJ: Association for Computational Linguistics.

Kolomiyets, O.; Kordjamshidi, P.; Moens, M.-F.; and Bethard, S. 2013. Semeval-2013 task 3: Spatial role labeling. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 255–262. Atlanta, GA: Association for Computational Linguistics.

Lang, J., and Lapata, M. 2010. Unsupervised Induction of Semantic Roles. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 939–947. Los Angeles, CA: Association for Computational Linguistics.

Laparra, E., and Rigau, G. 2013. ImpAr: A Deterministic Algorithm for Implicit Semantic Role Labelling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1180–1189. Sofia, Bulgaria: Association for Computational Linguistics.

Meyers, A.; Reeves, R.; Macleod, C.; Szekely, R.; Zielinska, V.; Young, B.; and Grishman, R. 2004. The NomBank Project: An Interim Report. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, 24–31. Boston, MA: Association for Computational Linguistics.

Palmer, M.; Gildea, D.; and Kingsbury, P. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics* 31(1):71–106.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Pradhan, S.; Ramshaw, L.; Marcus, M.; Palmer, M.; Weischedel, R.; and Xue, N. 2011. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, 1–27. Portland, OR: Association for Computational Linguistics.

Ruppenhofer, J.; Sporleder, C.; Morante, R.; Baker, C.; and Palmer, M. 2009. SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, 106–111. Boulder, CO: Association for Computational Linguistics.

Saurí, R., and Pustejovsky, J. 2012. Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text. *Computational Linguistics* 38(2):261–299.

Surdeanu, M. 2013. Overview of the TAC2013 Knowledge Base Population Evaluation: English Slot Filling and Temporal Slot Filling. In *Proceedings of the TAC-KBP 2013 Workshop*.

Weischedel, R., and Brunstein, A. 2005. BBN Pronoun Coreference and Entity Type Corpus. Technical report, Linguistic Data Consortium, Philadelphia.

Zhou, J., and Xu, W. 2015. End-to-end Learning of Semantic Role Labeling Using Recurrent Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 1127–1137. Beijing, China: Association for Computational Linguistics.