Annotating Temporally-Anchored Spatial Knowledge on Top of OntoNotes Semantic Roles

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

Abstract

This paper presents a two-step methodology to annotate spatial knowledge on top of OntoNotes semantic roles. First, we manipulate semantic roles to automatically generate potential additional spatial knowledge. Second, we crowdsource annotations with Amazon Mechanical Turk to either validate or discard the potential additional spatial knowledge. The resulting annotations indicate whether entities are or are not located somewhere with a degree of certainty, and temporally anchor this spatial information. Crowdsourcing experiments show that the additional spatial knowledge is ubiquitous and intuitive to humans, and experimental results show that it can be inferred automatically using standard supervised machine learning techniques.

Keywords: spatial knowledge, temporally-anchored knowledge, semantic inference

1. Introduction

Extracting meaning from text has received considerable attention in the last decade. In particular, semantic role labeling and efforts focused on spatial meaning—both corpora development and automatic tools—have become popular. Semantic roles capture semantic links between predicates and their arguments; they capture who did what to whom, how, when and where (Baker et al., 1998; Palmer et al., 2005). Efforts targeting spatial meaning use specialized relations such as TRAJECTOR and LANDMARK (Kordjamshidi et al., 2011; Kolomiyets et al., 2013), or define subtasks such as identifying spatial elements and spatial signals (Pustejovsky et al., 2015).

There are several corpora with semantic role annotations, e.g., FrameNet (Baker et al., 1998), PropBank (Palmer et al., 2005), and OntoNotes (Hovy et al., 2006). While semantic roles are useful, 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 semantic roles, one can infer that *John* had LOCATION *San Francisco* for a 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 although 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*.

This paper presents (1) annotations of temporally-anchored spatial knowledge on top of OntoNotes semantic roles, and (2) experiments to extract this knowledge automatically. We release a new resource[1] that annotates where entities are and are *not* located, and temporally anchor this information. Additionally, we incorporate certainty levels since there is often evidence that something is (or is not) located somewhere, but one cannot fully commit.

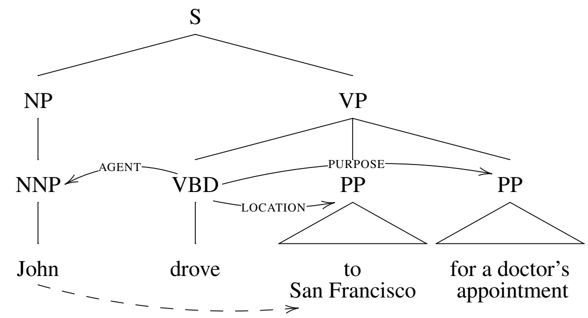


Figure 1: Semantic roles in OntoNotes (solid arrows) and additional spatial knowledge (dashed arrow).

2. OntoNotes 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. We use the term *additional spatial knowledge* to refer to relations LOCATION(x, y) that are not semantic roles, i.e., when (1) x is not a predicate or (2) x is a predicate and y is not an argument of x.

2.1. Semantic Roles in OntoNotes

OntoNotes (Hovy et al., 2006) is a large corpus of 63,918 sentences from several genres including newswire, broadcast news and conversations, and magazines.[2] It includes POS tags, word senses, parse trees, speaker information, named entities, semantic roles and coreference.

OntoNotes semantic roles follow PropBank framesets and only account for verbal roles, i.e., for all semantic roles R(x, y), x is a verb. The role set consists of numbered arguments and argument modifiers. Numbered arguments, also referred to as core arguments, range from ARG_0 to ARG_5, and their meanings are verb-dependent, e.g., ARG_3 is used to indicate the INSTRUMENT with *apply.03* and the

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

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

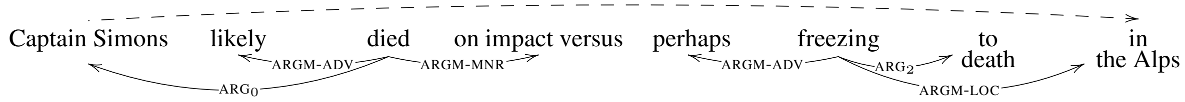


Figure 2: Semantic roles in OntoNotes (solid arrows) and additional spatial knowledge of type (1b) (dashed arrow).

[The paint]$_{\text{ARG}_1}$ was [applied]$_v$ [with a hard-bristle brush]$_{\text{ARG}_3}$.
[In '69]$_{\text{ARGM-TMP}}$ [at the age of 11]$_{\text{ARGM-TMP}}$ [you]$_{\text{ARG}_0}$ [went]$_v$ [from Beijing]$_{\text{ARG}_3}$ [to Shanghai]$_{\text{ARG}_4}$.

Table 1: Examples of PropBank-style semantic roles.

ARGM-LOC: location	ARGM-CAU: cause
ARGM-EXT: extent	ARGM-TMP: time
ARGM-DIS: discourse connectives	ARGM-PNC: purpose
ARGM-ADV: general-purpose	ARGM-MNR: manner
ARGM-NEG: negation marker	ARGM-DIR: direction
ARGM-MOD: modal verb	

Table 2: Argument modifiers in PropBank and OntoNotes.

START_POINT with *go.01* (Table 1). Argument modifiers have a common meaning across verbs, the list of modifiers provided by Palmer et al. (2005) is reproduced verbatim in Table 2. For a more detailed description of the semantic roles used in OntoNotes, we refer the reader to the LDC catalog[3] and PropBank (Palmer et al., 2005).

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

2.2. Additional Spatial Knowledge

OntoNotes annotates spatial information with (1) ARGM-LOC for all verbs, and (2) numbered arguments for a few verbs, e.g., the START_POINT and END_POINT of *go.01* are annotated with ARG_3 and ARG_4 respectively.

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

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*).

In this paper, we focus on annotating and extracting additional spatial knowledge LOCATION(x, y) of type (1) when x and y satisfy certain constraints (Section 3.1.).

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

```
foreach sentence s do
    foreach semantic role ARGM-LOC(y_verb, y) ∈ s do
        foreach semantic role ARG_i(x_verb, x) ∈ s do
            if is_valid(x, y) then
                is x located at y a day before y_verb?
                is x located at y a during y_verb?
                is x located at y a day after y_verb?
```

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

3. Corpus Creation

We follow a two-step methodology to annotate temporally-anchored spatial knowledge on top of OntoNotes. First, we manipulate semantic roles to generate potential spatial knowledge. Second, we gather crowdsourced annotations to either discard or validate the potential knowledge.

3.1. Generating Potential Additional Spatial Knowledge

All potential spatial knowledge inferable from OntoNotes semantic roles (i.e., spatial knowledge of type 1, Section 2.2.) can be generated by calculating all combinations of semantic roles. Such a brute-force approach generates a lot of potential knowledge that is later discarded during the annotation process. In order to make the annotation effort more efficient, we target additional LOCATION(x, y) inferable from intra-sentential numbered arguments $\text{ARG}_i(x_{verb}, x)$ and ARGM-LOC(y_{verb}, y), and impose the following restrictions:

1. x and y must not overlap;
2. the head of x must be a named entity of type `person`, `org`, `work_of_art`, `fac`, `norp`, `product` or `event`;[4]
3. the head of y must be a noun subsumed by *physical_entity.n.01* in WordNet (Miller, 1995) or a named entity of type `fac`, `gpe`, `loc`, or `org`; and
4. the heads of x and y must be different than the heads of all previously generated pairs from the same sentence.

We defined these restrictions with two goals in mind: to ease the annotation effort and generate the least amount of invalid potential knowledge possible. ARGM-LOC is the most likely role to indicate spatial information in OntoNotes and the vast majority of roles (71%) are numbered roles. When x is a named entity, the additional spatial knowledge is more intuitive. When y does not satisfy restriction (3), e.g., *here*, *in my brain*, *under him*, potential additional knowledge is almost always invalid.

All potential spatial knowledge targeted in this paper is generated using Algorithm 1. *is_valid(x,y)* returns true if

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

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

Table 3: Label counts and percentages per temporal anchor.

Restrictions	Number of pairs (*x*, *y*) generated			
	Train	Dev	Test	Total
None	34,362	5,217	5,418	44,997
1	28,480	4,329	4,463	37,272
2	3,007	483	400	3,890
3	15,155	2,568	2,475	20,198
4	32,431	4,884	5,118	42,433
1, 2	2,856	460	381	3,697
1, 3	13,922	2,341	2,252	18,515
1, 4	28,446	4,287	4,458	37,191
1, 2, 3	1,424	263	176	1,863
1, 2, 3, 4	1,321	247	164	1,732

Table 4: Number of pairs (*x*, *y*) generated after enforcing different restrictions (Section 3.1.).

restrictions 1–4 above are satisfied. The total number of ARGM-LOCs is 9,612, and the total number of pairs (*x*, *y*) prior to enforcing any restriction is 44,997. Table 4 shows the number of pairs (*x*, *y*) generated using several combinations of restrictions. After enforcing all restrictions, we generate 1,732 pairs; for each pair, we generate 3 questions to gather temporally-anchored spatial knowledge:

- Is *x* located at *y* the day before y_{verb}?
- Is *x* located at *y* during y_{verb}?
- Is *x* located at *y* the day after y_{verb}?

3.2. Crowdsourcing Annotations

Once potential additional spatial knowledge is generated via simple plain English questions, it is time to gather answers. After pilot annotations (Blanco and Vempala, 2015), it became clear that it is suboptimal to force annotators to answer YES, NO or UNKNOWN—often times there is evidence that something is (or is not) located somewhere, but it is difficult to fully commit. Inspired by previous work (Saurí and Pustejovsky, 2012), we considered 6 labels:

- certYES: I am certain that the answer is yes.
- probYES: The answer is probably yes, but I am unsure.
- certNO: I am certain that the answer is no.
- probNO: The answer is probably no, but I am unsure.
- UNK: There is not enough information to choose one of the labels above.
- INV: The question is invalid, I can't understand it.

Annotations were gathered using Amazon Mechanical Turk. We created Human Intelligence Tasks (HITs) consisting of the 3 questions regarding a potential additional LOCATION(*x*, *y*). The only information available to annotators was the source sentence, they did not have access to

Genre	#pairs	Day Before	During	Day After	All
nw	685	0.85	0.87	0.82	0.86
bn	437	0.85	0.85	0.82	0.84
bc	161	0.92	0.94	0.89	0.92
mz	306	0.66	0.92	0.79	0.78
wb	143	0.84	0.87	0.85	0.87
All	1,732	0.80	0.87	0.79	0.83

Table 5: Pearson correlation between crowdsourced annotations and control sentences (10% of annotated sentences).

semantic role information or any additional linguistic information. Figure 3 shows the interface including instructions, an example, and the radio buttons that force annotators to chose one option per temporal anchor.

We created 1,732 HITs (5,196 questions) and published them in batches based on the genre of the source text. We recruited annotators with previous approval rate $\geq 90\%$ and past approved HIT count over 5,000. We discarded submissions that took unusually short time compared to other submissions, and work done by annotators who always chose the same label. We requested 5 annotations per HIT and paid annotators \$0.03 per HIT. 150 annotators participated in the task, on average they annotated 57.33 HITs (minimum: 1, maximum: 1,409). The final labels were assigned using the mode among the 5 annotations (the label that occurs most often). Ties had to be broken randomly for 22.48% of questions.

4. Corpus Analysis

Columns 2–13 in Table 3 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, people standing somewhere.

4.1. Annotation Quality

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 5), and *during* questions show a higher correlation (0.87)

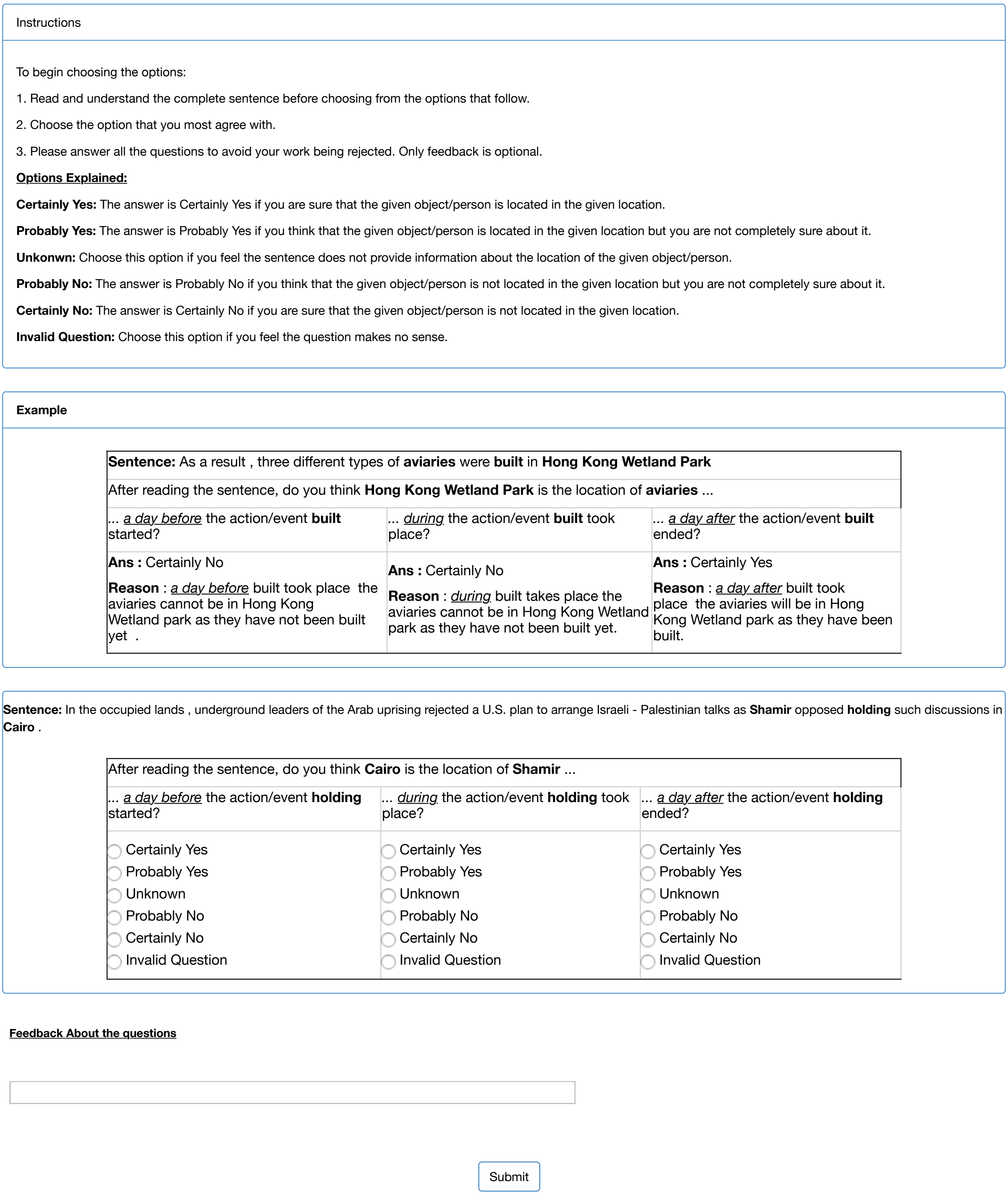


Figure 3: Amazon Mechanical Turk instructions, example and interface used to crowdsource annotations.

than *before* and after (0.80, 0.79). Correlations per genre are between 0.78 and 0.92, i.e., all genres achieved high agreements. The highest Pearson correlation is obtained with sentences from broadcast conversations (bc, 0.92), followed by web data (wb, 0.87), newswire (nw, 0.86), broadcast news (bn, 0.84), and magazine (mz, 0.78)

We also calculated the raw inter-annotator agreements and the percentages of questions for which there is no tie (Table 6). At least 3 annotators agreed (perfect match) in 58.6% of questions and at least 2 annotators in 98.5%. Overall, there were no ties in 77.52% of questions. 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, note that the final labels can sometimes be calculated without breaking ties if a majority label does not exist but 2 annotators agree (at least 3 annotators agree: 58.6%, no tie: 77.52%), e.g., {`probYES`, `UNK`, `INV`, `probYES`, `certYES`}.

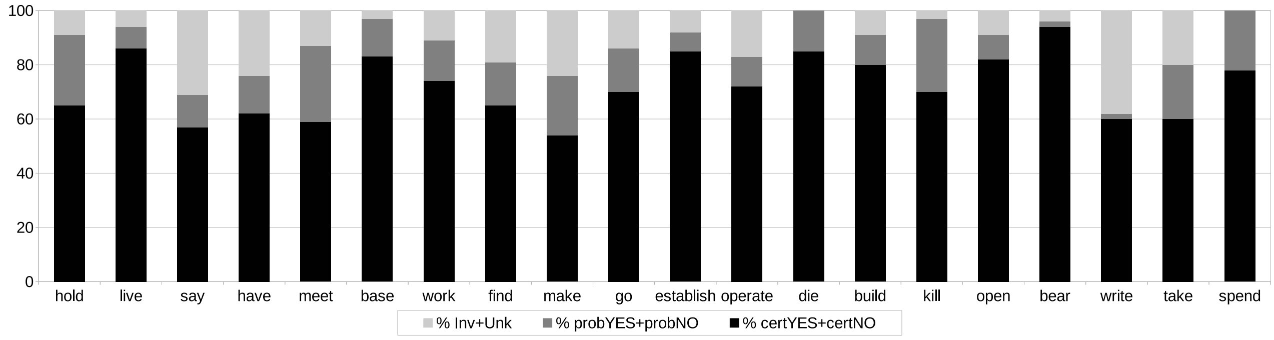


Figure 4: Label distribution for the top 20 most frequent verbs to which y attaches (y_{verb}).

	% of annotators that agree				% No Tie
	≥ 5	≥ 4	≥ 3	≥ 2	
Day Before	2.9	15.3	54.9	98.4	75.69
During	12.4	35.1	68.4	98.6	82.21
Day After	3.4	16.3	52.5	98.5	74.65
All	6.2	22.2	58.6	98.5	77.52

Table 6: Percentage of questions for which at least 5, 4, 3 or 2 annotators agree (out of 5), and percentage of questions without a tie.

Top 20 most certain verbs
leave, *explode*, *begin*, *march*, *stand*, *bear*, *teach*, *discuss*, *arrest*, *discover*, *carry*, *receive*, *raise*, *bury*, *establish*, *appear*, *live*, *die*, *base* and *open*
Top 20 verbs with highest inter-annotator agreement
hear, *hire*, *begin*, *lead*, *bear*, *locate*, *march*, *conduct*, *call*, *receive*, *bury*, *provide*, *attack*, *retire*, *lock*, *draw*, *teach*, *base*, *execute* and *stop*

Table 7: Top 20 most certain verbs (i.e., with the most `certYES` and `certNO` labels) and top 20 verbs with the highest inter-annotator agreements sorted by frequency.

Out of the top 50 most frequent verbs to which y attaches (y_{verb}), the ones with the most `certYES` and `certNO` labels and the ones with highest inter-annotator agreements are presented in Table 7. Finally, Figure 4 depicts the label distribution for the top 20 most frequent verbs. Note that most labels are either `certYES` or `certNO`, i.e., additional spatial knowledge can be inferred with certainty.

4.2. Annotation Examples

In this section, we provide samples of easy and difficult annotations based on annotator agreement.

Consider sentence *[Officer Payne]*$_{\text{AGENT}}$ *[collected]*$_{\text{verb}}$ *[the AK-47]*$_{\text{THEME}}$ *[at the warehouse]*$_{\text{LOCATION}}$. Annotators interpreted that the *AK-47* certainly had LOCATION *warehouse* the day before (`certYES`: 3, `probYES`:2) and during *collected* (`certYES`: 5), but not the day after (`certNO`: 5).

Consider now sentence *[Reporter Garith McClain]*$_{\text{ARG}_0,\, v_1}$ *is* *[covering]*$_{v_1}$ *[the story]*$_{\text{ARG}_1,\, v_1}$ *[for the [London]*$_{\text{ARGM-LOC},\, v_2}$ *[based]*$_{v_2}$ *[Guardian Newspaper]*$_{\text{ARG}_1,\, v_2}$*]*$_{\text{ARGM-ADV},\, v_1}$. While there is not enough information to determine whether *Garith McClain* has LOCATION *London* at any point of time, some annotators interpreted that it is probable (`probYES`:2, `UNK`: 3).

5. Experimental Results

We follow a standard supervised machine learning approach. Each of the 5,196 questions becomes an instance. In this paper, we experiment with instances whose majority label is not `INV` (Invalid) and for which at least 3 annotators agreed (2,725 intances, 52%). We follow the CoNLL-2011 Shared Task (Pradhan et al., 2011) split into train, development and test, and train 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 were calculated using test instances. All features are derived from gold standard linguistic annotations (POS tags, parse trees, semantic roles, etc.). We have previously presented results including instances for which less than 3 annotators agreed and using predicted linguistic annotations (Vempala and Blanco, 2016).

Feature selection. Table 8 presents the feature set. Lexical and syntactic features are standard in semantic role labeling (Gildea and Jurafsky, 2002). We added several features extracted from the semantic role representations we infer from (Features 12–20).

Semantic features are derived from the verb-argument structures from which the potential additional relation LOCATION(x, y) is 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 the heads of x and y. Figure 5 exemplifies all features.

We also tried several additional semantic features, e.g., flags indicating presence of all semantic roles (not only ARGM-LOC and ARGM-TMP), counts 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 using cross-validation with train and development instances.

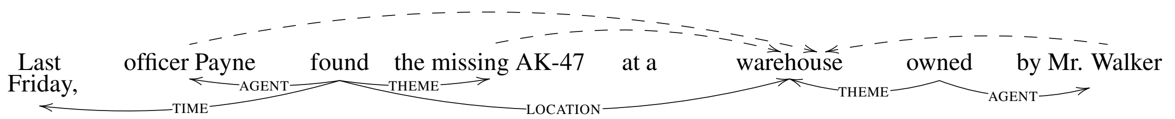
Results. Table 9 presents results obtained using a baseline and all features. The baseline predicts the most likely label per temporal anchor (day before: `certNO`, during: `certYES`, day after: `certYES`) and obtains an F-measure of 0.31. It is worth noting that *during* instances obtain a relatively high overall F-measure with the baseline, 0.60.

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 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 8: Lexical, syntactic and semantic features to infer potential additional relations LOCATION(x, y).

System		Day Before			During			Day After			All		
		P	R	F	P	R	F	P	R	F	P	R	F
most frequent per temporal anchor baseline	`certYES`	0.00	0.00	0.00	0.72	1.00	0.84	0.44	1.00	0.61	0.59	0.45	0.51
	`certNO`	0.58	1.00	0.74	0.00	0.00	0.00	0.00	0.00	0.00	0.58	0.45	0.51
	Other	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	0.34	0.58	0.43	0.52	0.72	0.60	0.19	0.44	0.27	0.51	0.59	**0.54**
lexical + syntactic + semantic features	`certYES`	0.33	0.18	0.23	0.74	1.00	0.85	0.65	0.69	0.67	0.69	0.80	0.74
	`probYES`	1.00	0.14	0.25	0.00	0.00	0.00	0.25	0.25	0.25	0.40	0.17	0.24
	`certNO`	0.63	0.91	0.75	1.00	0.14	0.24	0.68	0.76	0.71	0.66	0.69	0.67
	`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	0.53	0.58	0.51	0.75	0.75	0.67	0.57	0.62	0.60	0.61	0.66	**0.63**

Table 9: Results obtained with the baseline and all features. We report results using instances whose majority label is not `INV` and for which at least 3 annotators agree.



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`

Type	No.	Feature Name	Value
	0	temporal tag	Day Before, During or Day After
Lexical	1–4	first word, POS tag	x: officer, NN; y: warehouse, NN
	5–8	last word, POS tag	x: Payne, NNP; y: warehouse, NN
Syntactic	9, 10	syntactic node	x: NNP; y: NN
	11	common subsumer	PP
Semantic	12–15	predicate, POS tag	x_{verb}: find, VBD; y_{verb}: find, VBD
	16	same predicate	True
	17	ARGM-LOC count	1
	18	ARGM-TMP count	1
	19, 20	NE type	x: PER; y : NONE

Figure 5: Semantic role labels (solid arrows), all potential additional spatial knowledge (dashed arrows), annotations per temporal anchor, and feature values extracted for pair (*officer Payne*, *warehouse*).

Using all features, the overall F-measure is 0.63. *During* instances obtain higher F-measure (0.67) than *before* (0.51) and *after* (0.60). This is not surprising, as *during* instances obtained higher inter-annotator agreements. F-measures are higher for the labels that allow us to infer spatial knowledge with certainty (`certYES`: 0.74, `certNO`: 0.67) than other labels (`probYES`: 0.24; `probNO`, `UNK`: 0.00). Previously, we have presented feature ablation experiments (Vempala and Blanco, 2016).

6. Previous 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). These systems extract semantic links between verbs and their arguments. In contrast, the work presented here complements semantic role representations with temporally-anchored spatial knowledge.

There have been several proposals to extract semantic links not annotated in well-known corpora such as Propbank (Palmer et al., 2005), FrameNet (Baker et al., 1998) or Nombank (Meyers et al., 2004). Gerber and Chai (2010) augmented NomBank annotations with additional numbered arguments appearing in the same or previous sentences, and Laparra and Rigau (2013) presented an improved algorithm for the same task. The SemEval-2010 Task 10: Linking Events and their Participants in Discourse (Ruppenhofer et al., 2009) targeted cross-sentence missing numbered arguments in PropBank and FrameNet. We have previously proposed an unsupervised framework to compose semantic relations out of previously extracted relations (Blanco and Moldovan, 2011a; Blanco and Moldovan, 2011b), and a supervised approach to infer additional argument modifiers (ARGM) for verbs in PropBank (Blanco and Moldovan, 2014). Unlike the current work, these previous efforts improve the semantic representation of predicates. None of them infer semantic links between arguments of predicates, target temporally-anchored spatial knowledge or account for degrees of certainty.

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. In contrast, the approach presented in this paper builds on top of semantic roles, targets temporally-anchored LOCATION relations, and accounts for uncertainty (certYES / certNO vs. probYES / probNO).

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, i.e., trajector, landmark, spatial and motion indicators, path, direction, distance, etc. Unlike us, the task does not consider temporal anchors or uncertainty. As the examples throughout this paper illustrate, doing so is useful because (1) spatial information does not hold forever 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.

This paper is an extension of our previous work. We have presented preliminary annotations and experiments following the same approach to generate potential additional spatial knowledge (Section 3.1.), but only enforcing restriction 1 and using 200 sentences (Blanco and Vempala, 2015). We have also presented additional results using the same crowdsourced annotations detailed in this paper (Vempala and Blanco, 2016).

7. Conclusions

Semantic roles capture who did what to whom, how, when and where. Among other role labels, PropBank uses numbered arguments (ARG_0, ARG_1, etc.) to encode the core arguments of a verb, and ARGM-LOC to encode the location. This work takes advantage of OntoNotes semantic roles in order to infer temporally-anchored spatial knowledge. Semantic role representations within a sentence are combined in order to infer whether entities are or are *not* located somewhere, and assign a certainty label to this additional knowledge.

A new resource with additional spatial knowledge annotated on top of OntoNotes is presented with detailed analysis. Most potential additional spatial knowledge automatically generated can be inferred with certainty (certYES: 42.22% , certNO: 25.36%). Crowdsourcing experiments show that the additional knowledge is intuitive to humans, the overall Pearson between final labels and control sentences is 0.83.

Experimental results show that inferring additional spatial knowledge can be done with a modest weighted F-measure of 0.63. Results are higher for certYES and certNO (0.74 and 0.67), the labels that indicate that something is certainly located somewhere or not. Inferring spatial knowledge for the day before or after an event occurred is harder than during the event (0.51 and 0.60 vs. 0.67).

The most important conclusion of this work is the fact that given an ARGM-LOC semantic role, temporally-anchored spatial knowledge can be inferred for numbered arguments in the same sentence. Indeed, annotators answered 50.91% of questions with certYES or probYES, and 31.71% of questions with certNO or probNO (Table 3). Another important observation is that spatial knowledge can be inferred from most verbs, not only motion verbs. While it is fairly obvious to infer from *John moved to Paris* that he had LOCATION *Paris* the day after *moved* but (probably) not the day before or during, we can also infer the location of an entity with respect to verbs such as *found* (Figure 5). Indeed, several of the top 20 most certain verbs (Table 7) are non-motion verbs, e.g., *explode*, *begin*, *stand*, *teach*.

8. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 17th international conference on Computational Linguistics*, Montreal, Canada.

Blanco, E. and Moldovan, D. (2011a). A model for composing semantic relations. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 45–54. Association for Computational Linguistics.

Blanco, E. and Moldovan, D. (2011b). Unsupervised learning of semantic relation composition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1456–1465, Portland, Oregon, USA, June. 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*, pages 145–154, Gothenburg, Sweden, April. Association for Computational Linguistics.

Blanco, E. and Vempala, A. (2015). Inferring temporally-anchored spatial knowledge from semantic roles. In *Pro-*

ceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 452–461, Denver, Colorado, May–June. Association for Computational Linguistics.

Carreras, X. and Màrquez, L. (2005). Introduction to the CoNLL-2005 shared task: semantic role labeling. In *CONLL '05: Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 152–164, Morristown, NJ, USA. 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: Long Papers - Volume 1*, pages 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*, pages 1583–1592, Uppsala, Sweden, July. Association for Computational Linguistics.

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

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, pages 1–18, Boulder, Colorado. Association for Computational Linguistics.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: the 90% Solution. In *NAACL '06: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, Morristown, NJ, USA. Association for Computational Linguistics.

Kolomiyets, O., Kordjamshidi, P., Moens, M.-F., and Bethard, S. (2013). Semeval-2013 task 3: Spatial role labeling. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 255–262. Association for Computational Linguistics.

Kordjamshidi, P., Van Otterlo, M., and Moens, M.-F. (2011). Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Trans. Speech Lang. Process.*, 8(3):4:1–4:36, December.

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 (Volume 1: Long Papers)*, pages 1180–1189, Sofia, Bulgaria, August. 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 A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May. Association for Computational Linguistics.

Miller, G. A. (1995). Wordnet: A lexical database for english. In *Communications of the ACM*, volume 38, pages 39–41.

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*, pages 1–27, Portland, Oregon, USA, June. Association for Computational Linguistics.

Pustejovsky, J., Kordjamshidi, P., Moens, M.-F., Levine, A., Dworman, S., and Yocum, Z. (2015). Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 884–894, Denver, Colorado, June. 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)*, pages 106–111, Boulder, Colorado, June. 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, June.

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*.

Vempala, A. and Blanco, E. (2016). Complementing semantic roles with temporally-anchored spatial knowledge: Crowdsourced annotations and experiments. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, Phoenix, AZ.

Weischedel, R. and Brunstein, A. (2005). BBN pronoun coreference and entity type corpus. Technical report, Linguistic Data Consortium, Philadelphia.