Beyond Plain Spatial Knowledge: Determining Where Entities Are and Are Not Located, and For How Long

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 complements semantic role representations with spatial knowledge beyond indicating plain locations. Namely, we extract where entities are (and are not) located, and for how long (seconds, hours, days, etc.). Crowdsourced annotations show that this additional knowledge is intuitive to humans and can be annotated by non-experts. Experimental results show that the task can be automated.

1 Introduction

Extracting meaning from text is crucial for true text understanding and an important component of several natural language processing systems. Among many others, previous efforts have focused on extracting causal relations (Bethard and Martin, 2008), semantic relations between nominals (Hendrickx et al., 2010), spatial relations (Kordjamshidi et al., 2011) and temporal relations (Pustejovsky et al., 2003; Chambers et al., 2014).

In terms of corpora development and automated approaches, semantic roles are one of the most studied semantic representations (Toutanova et al., 2005; Màrquez et al., 2008). They have been proven useful for, among others, coreference resolution (Ponzetto and Strube, 2006) and question answering (Shen and Lapata, 2007). While semantic roles provide a useful semantic layer, they capture a portion of the meaning encoded in all but the simplest statements. Consider the sentence in Figure 1 and the semantic roles of *drove* (solid arrows). In addition to these roles, humans intuitively understand that (dashed arrow) (1) *John* was not located in *Berlin* before or during *drove*, (2) he was located in *Berlin* after *drove* for a short period of time (presumably, until he was done picking up the package, i.e., for a few minutes to an hour), and then left *Berlin* and thus (3) was not located there anymore. Some of this additional spatial knowledge is inherent to the motion verb *drive*: people cannot drive to the location where they are currently located, and they will be located at the destination of driving after *driving* takes place. But determining for how long the agent of *drive* remains at the destination depends on the arguments of *drive*: from *[John]*$_{\text{AGENT}}$ *[drove]*$_{v}$ *[home]*$_{\text{DESTINATION}}$ *[after an exhausting work day]*$_{\text{TIME}}$, it is reasonable to believe that *John* will be located at *home* overnight.

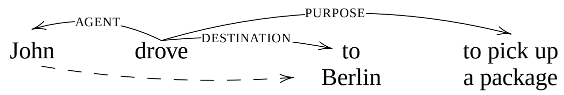


Figure 1: Semantic roles (solid arrows) and additional spatial knowledge (dashed arrow).

This paper manipulates semantic roles in order to extract temporally-anchored spatial knowledge. We extract where entities are and are *not* located, and temporally anchor this information. Temporal anchors indicate for how long something is (or is not) located somewhere, e.g., for 5 minutes before (or after) an event. We target additional spatial knowledge not only between arguments of motion verbs as exemplified above, but also between intra-sentential arguments of any verb. The main contributions are: (1) crowdsourced annotations on top of OntoNotes[1] indicating where something is and is not located (polarity), and for how long (temporal anchors); (2) detailed annotation analysis using coarse- and fine-grained labels (yes / no vs. seconds, minutes, years, etc.); and (3) experiments detailing results with several feature combinations, and using gold-standard and predicted linguistic information.

[1] Available at http://www.cse.unt.edu/~blanco/

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1502–1512,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

2 Definitions and Background

We use R(x, y) to denote a semantic relationship R between x and y. R(x, y) can be read "x has R y", e.g., AGENT(*drove*, *John*) can be read "*drove* has AGENT *John*." By definition, semantic roles are semantic relationships between predicates and their arguments—for all semantic roles R(x, y), x is a predicate and y is an argument of x. Generally speaking, semantic roles capture who did what to whom, how, when and where.

We use the term *additional spatial knowledge* to refer to spatial knowledge not captured with semantic roles, i.e., spatial meaning between x and y where (1) x is not a predicate or (2) x is a predicate and y is not an argument of x. As we shall see, we go beyond extracting "x has LOCATION y" with plain LOCATION(x, y) relations. We extract where entities are and are not located, and for how long they are located (and not located) somewhere.

2.1 Semantic Roles in OntoNotes

OntoNotes (Hovy et al., 2006) is large corpus (≈64K sentences) that includes verbal semantic role annotations, i.e., the first argument x of any role R(x, y) is a verb.[2] OntoNotes semantic roles follow PropBank framesets (Palmer et al., 2005). It uses a set of numbered arguments (ARG$_0$–ARG$_5$) whose meanings are verb-dependent, e.g., ARG$_2$ is used for "*employer*" with verb *work.01* and "*expected terminus of sleep*" with verb *sleep.01*. Additionally, it uses argument modifiers which share a common meaning across verbs (ARGM-LOC, ARGM-TMP, ARGM-PRP, ARGM-CAU, etc.). For a detailed description of OntoNotes semantic roles, we refer the reader to the LDC catalog[3] and PropBank (Palmer et al., 2005). To improve readability, we often rename numbered arguments, e.g., AGENT instead of ARG$_0$ in Figure 1.

3 Related Work

Approaches to extract PropBank-style semantic roles have been studied for years (Carreras and Màrquez, 2005), state-of-the-art tools sobtain F-measures of 83.5 (Lewis et al., 2015). In this paper, we complement semantic role representations with temporally-anchored spatial knowledge.

Extracting additional meaning on top of popular corpora is by no means a new problem. Gerber and Chai (2010) augmented NomBank (Meyers et al., 2004) 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 (Ruppenhofer et al., 2009) targeted cross-sentence missing arguments in FrameNet (Baker et al., 1998) and PropBank (Palmer et al., 2005). Silberer and Frank (2012) casted the SemEval task as an anaphora resolution task. We have previously proposed an unsupervised framework to compose semantic relations out of previously extracted relations (Blanco and Moldovan, 2011), 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 (1) improve the semantic representation of verbal and nominal predicates, or (2) infer relations between arguments of the same predicate.

More recently, we showed that spatial relations can be inferred from PropBank-style semantic roles (Blanco and Vempala, 2015; Vempala and Blanco, 2016). In this paper, we expand on this idea as follows. First, we not only extract whether "x has LOCATION y" before, during or after an event, but also specify for how long before and after (seconds, minutes, hours, days, weeks, months, years, etc.). Second, we release crowdsourced annotations for 1,732 potential additional spatial relations. Third, we experiment with both gold and predicted linguistic information.

Spatial semantics has received considerable attention in the last decade.

The task of spatial role labeling (Kordjamshidi et al., 2011; Kolomiyets et al., 2013) aims at representing spatial information with so-called spatial roles, e.g., trajector, landmark, spatial and motion indicators, etc. Unlike us, spatial role labeling does not aim at extracting where entities are *not* located or temporally-anchored spatial information. But doing so is intuitive to humans, as the examples and crowdsourced annotations in this paper show. Spatial knowledge is intuitively associated with motion events, e.g., *drive*, *go*, *fly*, *walk*, *run*. Hwang and Palmer (2015) presented a classifier to detect caused motion constructions triggered by non-motion verbs, e.g., *The crowd laughed the clown off the stage* (i.e., the crowd made the clown leave the stage). Our work does not target motion verbs or motion constructions,

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

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

as the examples in Table 3 show, non-motion constructions triggered by non-motion verbs also allow us to infer temporally-anchored spatial meaning, e.g., *played*, *honored*, *taught*, *fighting*.

4 Corpus Creation and Analysis

Our goal is to complement semantic role representations with additional spatial knowledge. Specifically, our goal is to infer temporally-anchored spatial knowledge between *x* and *y*, where semantic roles ARG$_i$(x_{verb}, *x*) and ARGM-LOC(y_{verb}, *y*) exists in the same sentence. In order to achieve this goal, we follow a two-step methodology. First, we automatically generate potential additional spatial knowledge by combining selected semantic roles. Second, we crowdsource annotations, including polarity and temporal anchors, to validate or discard the potential additional knowledge.

4.1 Generating Potential Spatial Knowledge

We generate potential additional relations LOCATION(*x*, *y*) by combining all ARG$_i$(x_{verb}, *x*) and ARGM-LOC(y_{verb}, *y*) semantic roles within a sentence (x_{verb} and y_{verb} need not be the same). Then, we enforce the following restrictions:

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

These restrictions were designed after manual analysis of randomly selected combinations of ARG$_i$ and ARGM-LOC semantic roles with two goals in mind: to (1) reduce the annotation effort and (2) generate the least amount of invalid potential additional spatial knowledge without arbitrarily discarding any predicates (e.g., focus only on motion verbs). Additional relations not satisfying restriction 1 are nonsensical, and restriction 4 simply discards potential additional relations that have already been generated. Restrictions 2 and 3 are designed to improve the likelihood that the potential additional spatial knowledge will not be discarded when crowdsourcing annotations, e.g., locations whose head is an adverb such as *here* and *there* (11% of all ARGM-LOC roles) do not yield valid additional spatial knowledge.

OntoNotes annotates $9{,}612$ ARGM-LOC semantic roles, and the number of potential LOCATION relations generated is $1{,}732$. Thus, our methodology aims at adding 18% of additional spatial relations on top of OntoNotes semantic roles. If we consider each temporal anchor as a different spatial relation, we aim at adding 54% additional spatial relations. As we shall see, over 69% of the additional potential relations are valid (Section 4.3).

4.2 Crowdsourcing Spatial Knowledge

Once potential spatial knowledge is generated, it must be validated or discarded. We are interested in additional spatial knowledge as intuitively understood by humans, so we avoid lengthy annotation guidelines and ask simple questions to non-experts via Amazon Mechanical Turk.

After in-house pilot annotations, it became clear that asking "Is *x* located in/at *y*" for each potential LOCATION(*x*, *y*) and forcing annotators to answer *yes* or *no* is suboptimal. For example, consider again Figure 1 and question "Is *John* located in *Berlin*?". An unabridged natural answer would be "not before or during *drove*, but certainly after *drove* for a few minutes until he was done *picking up the package*." In other words, it is intuitive to consider polarity (whether *x* is or is not located at *y*) and temporal anchors (for how long?).

We designed the interface in Figure 2 to gather annotations including polarity and temporal anchors, and accounting for granularity levels. Answers map to the following coarse-grained labels:

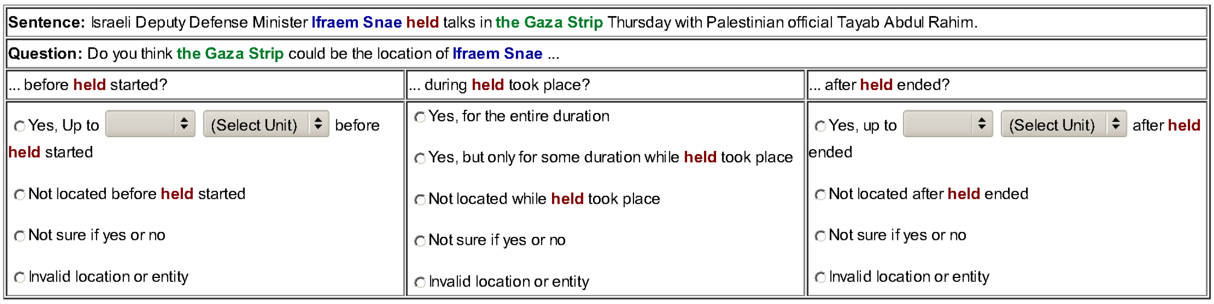
- Before and after: `yes`, `no`, `unk` and `inv`.
- During: `yes` (first 2 options), `no`, `unk` and `inv`.

Label `unk` stands for *unknown* and `inv` for *invalid*. Furthermore, `yes` maps to these fine-grained labels indicating specific periods of time:

- Before and after: an integer and a unit of time (`secs`, `mins`, `hours`, `days`, `weeks`, `months` or `years`)[5], or `inf` for infinity.
- During: `entire` or `some`.

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

[5] The interface restricts the range of valid integers, e.g., numbers selectable with `secs` range from 1 to 59.



Sentence: Israeli Deputy Defense Minister **Ifraem Snae** **held** talks in **the Gaza Strip** Thursday with Palestinian official Tayab Abdul Rahim.

Question: Do you think **the Gaza Strip** could be the location of **Ifraem Snae** ...

... before **held** started?	... during **held** took place?	... after **held** ended?
○ Yes, Up to [ ] (Select Unit) before **held** started	○ Yes, for the entire duration	○ Yes, up to [ ] (Select Unit) after **held** ended
	○ Yes, but only for some duration while **held** took place	
○ Not located before **held** started	○ Not located while **held** took place	○ Not located after **held** ended
○ Not sure if yes or no	○ Not sure if yes or no	○ Not sure if yes or no
○ Invalid location or entity	○ Invalid location or entity	○ Invalid location or entity

Figure 2: Amazon Mechanical Turk interface to collect temporally-anchored spatial annotations. Annotators were also provided with a description and examples of all answers (not shown).

	`secs`	`mins`	`hours`	`days`	`weeks`	`months`	`years`	`inf`	`entire`	`some`
Before	0.20	7.55	11.33	7.36	3.78	8.15	46.72	14.91	n/a	n/a
During	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	97.77	2.23
After	0.50	6.48	11.29	6.48	3.34	6.29	29.47	36.15	n/a	n/a

Table 1: Percentage of fine-grained labels for instances annotated with coarse-grained label `yes`.

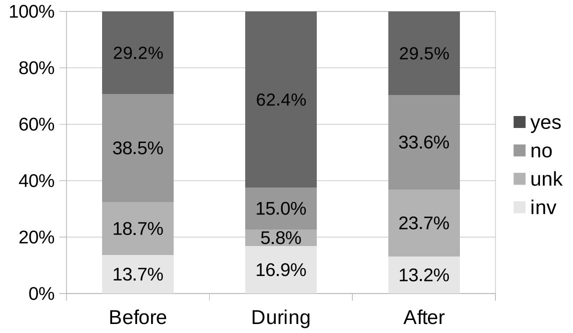


Figure 3: Percentage of coarse-grained labels per temporal anchor. Total number of annotations is $1{,}732 \times 3 = 5{,}196$.

We created one Human Intelligence Task (HIT) per potential LOCATION(x, y), and recruited annotators with previous approval rate $\geq 95\%$ and 5,000 or more previous approved HITs. A total of 74 annotators participated in the task, on average, they annotated 163.24 HITs (maximum: 1,547, minimum: 1). We rejected submissions that took unusually short time compared to other submissions, and those from annotators who always chose the same label. Overall, we only rejected 1.2% of submissions. We collected 7 annotations per HIT and paid $0.05 per HIT.

4.3 Annotation Analysis

Figure 3 shows the percentage of coarse-grained labels per temporal anchor. Labels `yes` and `no` combined account for 67.7% of labels (before), 77.4% (during) and 63.1% (after). Note that both `yes` and `no` yield valid additional spatial knowledge: whether x is (or is not) located at y. Annotators could not commit to `yes` or `no` in 16.1% of questions on average (`unk`), with a much smaller percentage for during temporal anchor (5.8%; before: 18.7%, after: 23.7%). This is not surprising, as arguments of some verbs, e.g., AGENT of *play*, must be located at the location of the event during the event, but not necessarily before or after. Finally, `inv` only accounts for 14.6% of labels (before: 13.7%, during: 16.9%, after: 13.2%), thus most potential additional knowledge automatically generated (Section 4.1) can be understood.

Percentages of fine-grained labels per temporal span, i.e., refinements of `yes` coarse-grained labels, are shown in Table 1. The vast majority of times (97.77%) annotators believe an entity is at a location during an event, the entity is there for the entire duration of the event (`entire`). Annotators barely used label `secs` (before: 0.20% and after: 0.50%), but percentages range between 3.34% and 46.72% for other units of time (uniform distribution would be $1/8 = 12.5\%$). Labels `years` and `inf`, which indicate that an entity is located somewhere for years or indefinitely before (or after) an event, are the most common fine-grained labels for *before* and *after* $(14.91–46.72\%)$.

4.3.1 Annotation Quality

Table 2 presents agreement measures. Pearson correlations are the weighted averages between each annotator and the majority label and are calculated following this mapping: (coarse labels): `yes`: 1, `unk`/`inv`: 0, `no`: -1; (fine labels): before/after: `secs`: 1, `mins`: $1 + 1/7$, `hours`:

	Coarse-grained labels						Fine-grained labels					
	Pearson	% instances s.t. a annotators agree					Pearson	% instances s.t. a annotators agree				
		$a = 7$	$a \geq 6$	$a \geq 5$	$a \geq 4$	$a \geq 3$		$a = 7$	$a \geq 6$	$a \geq 5$	$a \geq 4$	$a \geq 3$
Before	0.73	0.9	8.0	30.0	65.8	97.2	0.67	0.8	6.0	21.6	49.3	85.2
During	0.81	8.9	39.9	59.1	81.4	98.3	0.79	2.1	19.5	45.8	71.8	94.1
After	0.66	0.7	6.0	27.0	62.9	96.6	0.62	0.5	4.6	19.8	49.9	87.1
All	0.67	3.5	18.0	38.8	70.0	97.4	0.64	1.1	10.0	29.0	57.0	88.8

Table 2: Weighed Pearson correlations between annotators and the majority label, and percentage of instances for which at least 7, 6, 5, 4 and 3 annotators (out of 7) agree.

	Before		During		After	
Statement	C	F	C	F	C	F
Statement 1: [. . .] [Hsia]$_{\text{ARG}_0,\, v_1, v_2}$ [stopped]$_{v_1}$ off [in Milan]$_{\text{ARGM-LOC}, v_1}$ [to [visit]$_{v_2}$ [Hsiao Chin]$_{\text{ARG}_1, v_2}$]$_{\text{ARGM-PRP}, v_1}$.						
x: *Hsia*, y: *Milan*, y_{verb}: *stopped*	yes	mins	yes	entire	yes	hours
x: *Hsiao Chin*, y: *Milan*, y_{verb}: *stopped*	yes	years	yes	entire	yes	years
Statement 2: [President Clinton]$_{\text{ARG}_0, v_1}$ [played]$_{v_1}$ [a supporting role]$_{\text{ARG}_1, v_1}$ [today]$_{\text{ARGM-TMP}, v_1}$ [in [New York City where]$_{\text{ARGM-LOC}, v_2}$ [the first lady, Senator Clinton]$_{\text{ARG}_1, v_2}$, was [honored]$_{v_2}$ [at Madison Square Garden]$_{\text{ARGM-LOC}, v_2}$]$_{\text{ARGM-LOC}, v_1}$.						
x: *(President) Clinton*, y: *New York City*, y_{verb}: *played*	yes	hours	yes	entire	yes	hours
x: *(President) Clinton*, y: *Madison Square Garden*, y_{verb}: *honored*	yes	mins	yes	entire	yes	mins
x: *(Senator) Clinton*, y: *New York City*, y_{verb}: *played*	yes	hours	yes	entire	yes	hours
x: *(Senator) Clinton*, y: *Madison Square Garden*, y_{verb}: *honored*	yes	mins	yes	entire	yes	mins
Statement 3: [Before [joining]$_{v_2}$ [Maidenform]$_{\text{ARG}_1, v_2}$ [in 1972]$_{\text{ARGM-TMP},\, \text{V}_2}$]$_{\text{ARGM-TMP},\, v_1}$, [[Mr. Brawer, who]$_{\text{ARG}_0, v_3}$ [holds]$_{v_3}$ [a doctoral degree in English]$_{\text{ARG}_1, v_3}$]$_{\text{ARG}_0,\, v_1, v_2}$, [taught]$_{v_1}$ [at the University of Wisconsin]$_{\text{ARGM-LOC},\, v_1}$.						
x: *Maidenform*, y: *University of Wisconsin* , y_{verb}: *taught*	no	n/a	no	n/a	no	n/a
x: *Mr. Brawer*, y: *University of Wisconsin*, y_{verb}: *taught*	no	n/a	yes	entire	no	n/a
Statement 4: [...] [George Koskotas, self-confessed embezzler]$_{\text{ARG}_0,\, v_1}$, [now]$_{\text{ARGM-TMP}, v_1}$ [residing]$_{v_1}$ [in [a jail cell in Salem, Mass., from where]$_{\text{ARGM-LOC},\, v_2}$ [he]$_{\text{ARG}_0,\, v_2}$ is [fighting]$_{v_2}$ [extradition proceedings]$_{\text{ARG}_1, v_2}$]$_{\text{ARG}_1, v_1}$.						
x: George Koskotas, y: *a jail cell in Salem, Mass.*, y_{verb}: *fighting*	yes	months	yes	entire	unk	n/a

Table 3: Annotation examples. For each statement, we indicate semantic roles with square brackets, all potential additional spatial knowledge (is x located at y?), and annotations with respect to y_{verb} (coarse- (C) and fine-grained (F) labels per temporal anchor: before, during and after).

$1 + 2/7$, `days`: $1 + 3/7$, `weeks`: $1 + 4/7$, `months`: $1 + 5/7$, `years`: $1 + 6/7$, `inf`: 2; during: `some`: 1 `entire`:2. Calculating the weighted average of individual Pearson correlations allows us to take into account the number of questions answered by each annotator.

Correlations range between 0.66 and 0.81 with coarse-grained labels, and are slightly lower with fine-grained labels (0.67 vs. 0.73, 0.79 vs. 0.81, and 0.62 vs. 0.66). Questions for *during* temporal anchor are easier to answer with both kinds of labels (coarse: 0.81, fine: 0.79).

Table 2 also shows how many annotators (out of 7) chose the same label (exact match). At least 4 annotators agreed with coarse-grained labels in most instances (70%), and at least 3 annotators agreed virtually always (97.4%). Percentages are lower with fine-grained labels: 57.0% and 88.8%.

4.4 Annotation Examples

Table 3 presents several annotation examples. We include all potential additional spatial knowledge (Section 4.1) and annotations per temporal anchor.

Two additional LOCATION(x, y) can be inferred from Statement (1): whether *Hsia* and *Hisao Chin* are located in *Milan* before, during and after *stopped.* Annotators understood that *Hsia* was in *Milan* temporarily: for a few minutes before *stopped*, during the full duration of *stopped* and for a few hours after *stopped.* In other words, *Hsia* was elsewhere, then went to *Milan* and left after visiting with *Hsiao* for a few hours. Regarding *Hsiao*, annotators interpreted that *Milan* is her permanent location: for years before and after *Hsia stopped* to visit her. While somehow ambiguous, these annotations are reasonably intuitive.

Statement (2) has 2 ARGM-LOC roles and 4 potential additional relations. Annotations for *during* are straightforward: both *President Clinton* and *Senator Clinton* were located in *New York City* during *played* and at *Madison Square Garden* during *honored.* Annotations for *before* and *after* are more challenging: both Clintons where located in *New York City* for hours (not days) before and after *played*, but at *Madison Square Garden* for a few minutes (not hours) before and after *honored.*

Feature		Description
basic	1–4	x_{verb}, y_{verb} and their part-of-speech tags
lexical	5–12	first and last words of *x* and *y*, and their part-of-speech tags
	13	whether *x* occurs before or after *y*
heads	14–17	heads of *x* and *y*, and their part-of-speech tags
	18–19	named entity types of the heads of *x* and *y*
semantic	20	semantic role label linking x_{verb} and *x*
	21–24	number of ARGM-TMP and ARGM-LOC roles in x_{roles} and y_{roles}
	25–26	number of ARGM-TMP and ARGM-LOC roles in the sentence to which *x* and *y* belong
	27	whether x_{verb} and y_{verb} are the same verb

Table 4: Feature set to determine whether *x* is (or is not) located at *y*, and for how long. x_{verb} (y_{verb}) denote the verbs to which *x* (y) attach, and x_{roles} (y_{roles}) denote the semantic roles of x_{verb} (y_{verb}).

In other words, they arrived to *Madison Square Garden* shortly before *honored* and left shortly after, but stayed in *New York City* for some hours.

Statement (3) exemplifies `no` label. Potential additional spatial knowledge includes whether *Maidenform* is located at *University of Wisconsin*, which is never true (`no`). Additionally, *University of Wisconsin* was a location of *Mr. Brawer* while he *taught* there (*during*), but nor *before* or *after*.

Statement (4) exemplifies contrastive coarse-grained labels and `unk` label. Annotators interpreted that *George Koskotas* was in the *jail cell* for months before and during *fighting extradition*, and that it is unknown (`unk`) *after fighting* because the outcome of the fight is unknown.

5 Inferring Temporally-Anchored Spatial Knowledge

We follow a standard machine learning approach, and use the training, development and test sets released by the organizers of the CoNLL-2011 Shared Task (Pradhan et al., 2011). We first generate additional spatial knowledge deterministically as described in Section 4.1. Then, for each additional LOCATION(*x*, *y*), we generate one instance per temporal anchor and discard those annotated `inv`. The total number of instances is $1{,}732 \times 3 - 754 = 4{,}442$. We trained SVM models with RBF kernel using scikit-learn (Pedregosa et al., 2011). The feature set and SVM parameters were tuned using 10-fold cross-validation with the train and development sets, and results are calculated using the test set. During the tuning process, we discovered that it is beneficial to train one SVM per temporal anchor instead of a single model for the 3 temporal anchors.

5.1 Feature Selection

We use a mixture of standard features from semantic role labeling, and semantic features designed for extracting temporally-anchored spatial knowledge from semantic roles. In order to determine whether *x* is (or is not) located at *y* and for how long, we extract features from *x* and *y*, the verbs to which they attach (x_{verb} and y_{verb}) and all semantic roles of x_{verb} and y_{verb} (x_{roles} and y_{roles}).

Basic, *lexical* and *heads* features are standard in role labeling (Gildea and Jurafsky, 2002). *Basic* features are the word form and part-of-speech of x_{verb} and y_{verb}. *Lexical* features capture the first and last words of *x* and *y* and their part-of-speech tags, as well as a binary flag indicating whether *x* occurs before or after *y*. *Heads* features capture the heads of *x* and *y* and their part-of-speech tags, as well as their named entity types, if any.

Semantic features include features 20–27. Feature 20 indicates the semantic role linking *x* and x_{verb} (ARG$_0$, ARG$_1$, ARG$_2$, etc.); recall that the semantic role between *y* and y_{verb} is always ARGM-LOC (Section 4.1). Features 21–24 are counts of ARGM-TMP and ARGM-LOC semantic roles in the verb-argument structures to which *x* and *y* attach. Features 25–26 are the same counts of roles, but taking into account all the roles in the sentence to which *x* and *y* belong. Finally, feature 27 signals whether *x* and *y* attach to the same verb.

We tried many other features, including counts of all roles, heads of all semantic roles present, semantic role ordering, VerbNet (Schuler, 2005) and Levin (Levin, 1993) verb classes, and WordNet hypernyms (Miller, 1995), but they did not yield any improvements during the tuning process.

We exemplify features with pair (*x*: *George Koskotas, self-confessed embezzler*, *y*: *a jail cell in [...], from where*) from Statement 4 in Table 3:

- Basic: features 1–4: {residing, VBG, fighting, VBG}.
- Lexical: feature 5–12: {George, NNP, Koskotas, NNP, a, DT, where, WRB}, features 13: {before}.

		Before			During			After			All		
		P	R	F	P	R	F	P	R	F	P	R	F
baseline	yes	0.00	0.00	0.00	0.77	1.00	0.87	0.00	0.00	0.00	0.77	0.55	0.64
	no	0.49	1.00	0.65	0.00	0.00	0.00	0.40	1.00	0.57	0.44	0.86	0.58
	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
	Weighted avg.	0.24	0.49	0.32	0.60	0.77	0.67	0.16	0.40	0.23	0.51	0.55	0.50
basic	yes	0.48	0.34	0.40	0.82	0.94	0.88	0.53	0.50	0.52	0.71	0.71	0.71
	no	0.52	0.63	0.57	0.62	0.36	0.46	0.47	0.49	0.48	0.51	0.54	0.52
	unk	0.23	0.21	0.22	0.33	0.10	0.15	0.29	0.29	0.29	0.26	0.23	0.25
	Weighted avg.	0.44	0.45	0.44	0.76	0.79	0.76	0.44	0.44	0.44	0.55	0.56	**0.56**
basic + lexical + heads	yes	0.68	0.37	0.48	0.83	0.92	0.87	0.73	0.38	0.50	0.79	0.67	0.73
	no	0.59	0.80	0.68	0.47	0.32	0.38	0.53	0.66	0.59	0.56	0.68	0.61
	unk	0.36	0.29	0.32	0.20	0.10	0.13	0.32	0.39	0.35	0.33	0.32	0.32
	Weighted avg.	0.56	0.56	0.54	0.73	0.77	0.74	0.54	0.50	0.50	0.62	0.61	**0.61**
basic + lexical + heads + semantics	yes	0.86	0.44	0.58	0.85	0.94	0.90	0.79	0.38	0.51	0.84	0.70	**0.77**
	no	0.63	0.80	0.71	0.56	0.47	0.50	0.55	0.69	0.61	0.59	0.71	**0.64**
	unk	0.37	0.38	0.38	0.50	0.10	0.17	0.33	0.42	0.37	0.35	0.37	**0.36**
	Weighted avg.	0.64	0.60	**0.60**	0.78	0.81	**0.78**	0.57	0.52	**0.52**	0.66	0.64	**0.65**

Table 5: Results obtained with gold-standard linguistic annotations and coarse-grained labels using the baseline and several feature combinations (basic, lexical, heads and semantic features).

- Head: features 14–17: {Koskotas, NNP, cell, NN}, features 18–19: {person, none},
- Semantic feature 20: {ARG$_0$}, features 21–24: {1, 0, 0, 1}, feature 25–26: {1, 1}, feature 27: {no}.

6 Experiments and Results

We present results using gold-standard (Section 6.1) and predicted (Section 6.2) linguistic annotations. POS tags, parse trees, named entities and semantic roles are taken directly from *gold* or *auto* files in the CoNLL-2011 Shared Task release.

6.1 Gold-Standard Linguistic Annotations

Using gold-standard linguistic annotations has two advantages. First, because we have gold semantic roles and named entities, we generate the same potential additional spatial knowledge generated while creating our annotations (Section 4.1). Second, feature values are guaranteed to be correct.

6.1.1 Predicting Coarse-Grained Labels

Table 5 presents results with coarse-grained labels using a baseline and learning with several combinations of features extracted from gold-standard linguistic annotations (POS tags, parse trees, semantic roles, etc.). The baseline predicts the most frequent label per temporal anchor, i.e., yes for *during*, and no for *before* and *after* (Figure 3).

Best results for all labels and temporal anchors are obtained with all features (basic, lexical, heads and semantics). Overall F-measure is 0.65, and *during* instances obtain higher F-measure (0.78) than *before* (0.60) and *after* (0.52). Regarding labels, yes obtains best results (overall 0.77), followed by no (0.64) and unk (0.36). Not surprisingly, the most frequent label per temporal anchor obtains the best results with all features (*before*: no, 0.71; *during*: yes, 0.90; *after*: no, 0.61).

Before and *after* instances benefit the most from learning with all features with respect to the baseline (before: 0.32 vs. 0.60, after: 0.23 vs. 0.52). While *during* instances also benefit, the difference in F-measure is lower (0.67 vs. 0.78).

Feature Ablation. The bottom 3 blocks in Table 5 present results using several feature types incrementally. *Basic* features yield an overall F-measure of 0.56, and surprisingly good results for *during* instances (0.76). Indeed, the best performance obtained with *during* instances is 0.78 (all features), suggesting that the verbs to which x and y attach are very strong features.

Lexical and *heads* features are most useful for *before* (0.44 vs. 0.54, +22.7%) and *after* (0.44 vs. 0.50, +13.6%) instances, and are actually detrimental for *during* instances (0.76 vs. 0.74, -2.6%). Including *semantic* features, however, improves results with respect to *basic* features for all temporal anchors: before: 0.44 vs. 0.60, 36.4% during: 0.76 vs. 0.78, 2.6% after: 0.44 vs. 0.52, 18.2%.

Differences in overall F-measure are not statistically significant between *basic* and *basic + lexical + heads* (0.56 vs. 0.61, Z-test, two-tailed, p-value $= 0.05$), but the difference including *semantic* features is significant (0.50 vs. 0.65, Z-test, two-tailed, p-value $= 0.009$).

		Before			During			After			All		
		P	R	F	P	R	F	P	R	F	P	R	F
baseline	spurious	0.50	1.00	0.66	0.50	1.00	0.66	0.50	1.00	0.66	0.50	1.00	0.66
	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
	Weighted avg.	0.25	0.50	0.33	0.25	0.50	0.33	0.25	0.50	0.33	0.25	0.50	0.33
basic	yes	0.00	0.00	0.00	0.47	0.56	0.51	0.00	0.00	0.00	0.52	0.36	0.43
	no	0.55	0.33	0.42	0.40	0.20	0.27	0.43	0.38	0.41	0.43	0.32	0.37
	unk	0.26	0.30	0.28	0.11	0.07	0.08	0.37	0.25	0.30	0.24	0.18	0.21
	spurious	0.68	0.91	0.78	0.68	0.71	0.70	0.67	0.93	0.78	0.69	0.88	0.77
	Weighted avg.	0.51	0.58	0.53	0.53	0.55	0.54	0.49	0.58	0.52	0.54	0.58	**0.55**
basic + lexical + heads + semantics	yes	1.00	0.07	0.13	0.74	0.87	0.80	0.00	0.00	0.00	0.74	0.56	**0.64**
	no	0.64	0.48	0.55	0.67	0.20	0.31	0.43	0.46	0.44	0.53	0.49	**0.51**
	unk	0.41	0.78	0.54	0.50	0.47	0.48	0.41	0.61	0.49	0.51	0.68	0.58
	spurious	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	Weighted avg.	0.82	0.75	0.73	0.84	0.84	0.83	0.66	0.71	0.68	0.80	0.79	**0.79**

Table 7: Results obtained with predicted linguistic annotations and coarse-grained labels. spurious is a new label indicating overgenerated pairs not present in the gold standard.

	Label	P	R	F
Before	mins	1.00	0.67	0.80
	days	1.00	0.50	0.67
	years	0.15	0.17	0.16
	inf	1.00	0.33	0.50
	no	0.63	0.80	0.71
	unk	0.37	0.38	0.38
	other	0.00	0.00	0.00
	Weighted avg.	0.52	0.54	**0.51**
During	entire	0.84	0.94	0.88
	some	0.00	0.00	0.00
	no	0.56	0.45	0.50
	unk	0.50	0.10	0.17
	Weighted avg.	0.77	0.80	**0.77**
After	years	0.27	0.25	0.26
	inf	0.56	0.24	0.33
	no	0.55	0.69	0.61
	unk	0.33	0.42	0.37
	other	0.00	0.00	0.00
	Weighted avg.	0.41	0.44	**0.41**
All	mins	0.50	0.50	0.50
	days	1.00	0.29	0.44
	years	0.21	0.21	0.21
	inf	0.67	0.27	0.38
	entire	0.84	0.94	0.89
	no	0.59	0.71	0.64
	unk	0.35	0.37	0.36
	other	0.00	0.00	0.00
	Weighted avg.	0.56	0.59	**0.57**

Table 6: Results obtained with gold linguistic annotations and fine-grained labels using all features.

6.1.2 Predicting Fine-Grained Labels

Table 6 presents results using fine-grained labels and all features. Overall F-measure is lower than with coarse-grained labels (0.57 vs. 0.65). Results for *during* instances barely decreases (0.78 vs. 0.77) because almost 98% of fine-grained labels are entire (Table 1).

Most fine-grained labels for *before* and *after* are infrequent (Table 1), our best model is unable to predict labels secs, hours, weeks and months for *before*, and secs, mins, hours, days, weeks and months for *after* (*other* rows). But these labels account for relatively few instances: individually, between 0.2% and 11.33%, and among all of them, 23.46% for *before* and 34.38% for *after* instances.

It is worth noting that mins, days and inf obtain relatively high F-measures for *before*: 0.80, 0.67 and 0.50 respectively. In other words, we can distinguish whether an entity is somewhere only for a few minutes or days (but not longer) before an event, or at all times before an event.

6.2 Predicted Linguistic Annotations

In order to make an honest evaluation in a realistic environment, we also experiment with predicted linguistic annotations. The major disadvantage of doing so is that predicted semantic roles and named entities are often incorrect or missing, thus we generate spurious additional spatial knowledge and miss some additional spatial knowledge because the potential relation cannot be generated.

Table 7 presents results using predicted linguistic annotations. The additional label spurious is used for instances generated from incorrect semantic roles or named entities, as these instances do not appear in the crowdsourced annotations (Section 4). Due to space constraints, we only present results using coarse-grained labels, but provide results per temporal anchor.

The baseline, which predicts the most likely label per temporal anchor, always predicts spurious since 50% of generated additional potential knowledge does not appear in the crowdsourced annotations. Using all features clearly outperforms *basic* features (overall F-measure:

0.79 vs 0.55), thus we focus on the former.

Using all features, spurious is always predicted correctly. While useful to discard additional spatial knowledge that should not have been generated, spurious does not allow us to make meaningful inferences. The labels that we are most interested in, yes and no, obtain overall F-measures of 0.64 and 0.51 (compared to 0.77 and .64 with gold linguistic annotations). Regarding labels, yes can only be reliably predicted for *during* instances (F-measure: 0.80), and no is predicted with modest F-measures for all temporal anchors: before: 0.55, during: 0.31, after: 0.44.

7 Conclusions

This paper demonstrates that semantic roles are a reliable semantic layer from which one can infer whether entities are located or not located somewhere, and for how long (seconds, minutes, days, years, etc.). Crowdsourced annotations show that this kind of inferences are intuitive to humans. Moreover, most potential additional spatial knowledge generated following a few simple deterministic rules was validated by annotators (yes and no; before: 67.7%, during: 77.4%, after: 63.1%).

Experimental results with gold-standard semantic roles and named entities show that inference can be done with standard supervised machine learning (overall F-measure: 0.65, yes: 0.77, no: 0.64). Using predicted linguistic information, results decrease substantially (yes: 0.64, no: 0.51). This is mostly due to the fact that predicted semantic roles and named entities are often wrong or missing, and this fact unequivocally makes the inference process more challenging.

We believe that combining semantic roles and other semantic representation in a similar fashion to the one used in this paper could be useful to infer knowledge beyond spatial inferences. For example, one could infer who is in POSSESSION of something over time by manipulating the events in which the object in question participates in.

References

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

Steven Bethard and James H. Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eduardo Blanco and Dan Moldovan. 2011. Unsupervised learning of semantic relation composition. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, pages 1456–1465, Portland, Oregon.

Eduardo Blanco and Dan Moldovan. 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.

Eduardo Blanco and Alakananda Vempala. 2015. Inferring temporally-anchored spatial knowledge from semantic roles. In *Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL*, pages 452–461.

Xavier Carreras and Lluís Màrquez. 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.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Matthew Gerber and Joyce Chai. 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.

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

Iris Hendrickx, Su Nam Kim, ́ Zornitsa Kozareva, Preslav Nakov, Sebastian Ó Séaghdha, Diarmuid andPadó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: the 90% Solution. In *NAACL '06: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers on XX*, pages 57–60, Morristown, NJ, USA. Association for Computational Linguistics.

Jena D. Hwang and Martha Palmer. 2015. Identification of caused motion construction. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 51–60, Denver, Colorado, June. Association for Computational Linguistics.

Oleksandr Kolomiyets, Parisa Kordjamshidi, Marie-Francine Moens, and Steven Bethard. 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.

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

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

Beth Levin. 1993. *English verb classes and alternations : a preliminary investigation*.

Mike Lewis, Luheng He, and Luke Zettlemoyer. 2015. Joint a* ccg parsing and semantic role labelling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1444–1454, Lisbon, Portugal, September. Association for Computational Linguistics.

Lluís Màrquez, Xavier Carreras, Kenneth C Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: an introduction to the special issue. *Computational linguistics*, 34(2):145–159.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. Annotating noun argument structure for nombank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004)*.

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

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

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

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL ’06, pages 192–199, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 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.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 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.

Karin Kipper Schuler. 2005. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, Philadelphia, PA, USA. AAI3179808.

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

Carina Silberer and Anette Frank. 2012. Casting implicit role linking as an anaphora resolution task. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval ’12, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Alakananda Vempala and Eduardo Blanco. 2016. Complementing semantic roles with temporally anchored spatial knowledge: Crowdsourced annotations and experiments. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2652–2658.

Ralph Weischedel and Ada Brunstein. 2005. BBN pronoun coreference and entity type corpus. Technical report, Linguistic Data Consortium, Philadelphia.