# Annotating Temporally-Anchored Spatial Knowledge by Leveraging Syntactic Dependencies

**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 temporally-anchored spatial knowledge on top of OntoNotes. We first generate potential knowledge using syntactic dependencies, and then crowdsource annotations to validate the potential knowledge. The resulting annotations indicate how long entities are or are not located somewhere, and temporally anchor this information. Crowdsourcing experiments show that spatial inferences are ubiquitous and intuitive, and experimental results show that they can be done automatically.

**Keywords:** Semantics, Information extraction, Spatial knowledge

## 1. Introduction

Extracting spatial meaning from text is of utmost importance in natural language understanding. Efforts focused on spatial meaning—both corpora development and automatic tools—have become popular. Existing approaches to extract spatial knowledge usually focus on extracting locations of events, someone or something. For example, semantic role labeling (Palmer et al., 2005) determines who did what to whom, when and where, e.g., *Thelma Gutierrez [went]*$_{\text{verb}}$ *[inside the forensic laboratory where scientist are trying to solve this mystery]*$_{\text{ARG}_4}$, where ARG$_4$ indicates the END POINT of event *went*. Efforts targetting locations of entities include geo-locating Twitter users (Liu and Inkpen, 2015), and pairing companies with the location of their headquarters (Mintz et al., 2009) e.g., *[IBM's]*$_{\text{company}}$ *headquarters in [New York]*$_{\text{location}}$.

Determining the temporal span where the spatial knowledge holds is not extensively researched. From the sentence *John parked Jamie's car at the Highland Garage*, we can infer that *John* and the *car* are certainly located at *the Highland Garage* minutes before and during parking, and that *John* will leave shortly after *parking* whereas the *car* will be at the garage for a few days but not months. We can also infer that *Jamie* will probably be at *the Highland Garage* at some point after *parking* to pick up his car.

This paper presents (1) a two-step methodology to extract temporally-anchored spatial knowledge by manipulating syntactic dependencies, and a (2) crowdsourced corpus annotated with temporally-anchored spatial knowledge.

The work presented here extends our previous work (Vempala and Blanco, 2016), which only manipulated semantic roles. We show that additional temporally-anchored spatial knowledge can be extracted by leveraging syntactic dependencies. We release a new corpus that annotates how long entities are and are *not* located somewhere, and temporally anchor this spatial information.[1]

## 2. Background

We work on top of OntoNotes (Hovy et al., 2006) as it is a well known corpus with text from various domains. Ontonotes contains over 64,000 sentences. It annotates, among other linguistic information, part-of-speech tags, parse trees, named entities and co-reference chains. We use the CoNLL- 2011 Shared Task distribution (Pradhan et al., 2011), and transform the gold parse trees into syntactic dependencies using Stanford CoreNLP (Manning et al., 2014). De Marneffe and Manning (2008) present and exemplify the Stanford dependencies, and Weischedel and Brunstein (2005) the named entity types used in OntoNotes.

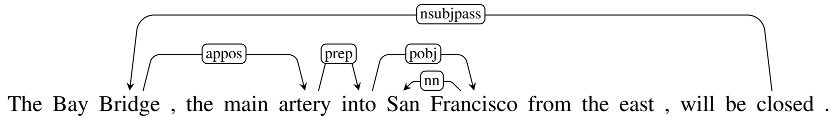
We use the term *temporally-anchored spatial knowledge* to refer to information regarding whether a given $x$ is or is not located at some location $y$, and for how long with respect to an event. We use the notation LOCATION($x$, $y$) to indicate the spatial relation between $x$ and $y$. We use the term potential spatial knowledge to refer to spatial relations LOCATION($x$, $y$) that are yet to be validated.

There are 2 types of 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. In the sentence *John called Google's office for Bill's appointment*, the relation LOCATION(*John*, *Google's office*) is of type (1) and LOCATION(*Bill*, *Google's office*) is of type (2). Also, in the example *Officer Jack found the missing diamond at a warehouse owned by Mr. Walker*, LOCATION(*Jack*, *warehouse*) is of type (1a) and LOCATION(*Mr. Walker*, *warehouse*) is of the type (1b). Our previous work (Vempala and Blanco, 2016) extracts spatial knowledge of type (1). In Section 3, we detail the approach that leverages syntactic dependencies to extract spatial knowledge of type (1) and (2).

## 3. Corpus Creation

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

[1] Available at https://alakanandav.bitbucket.io/




| $x$ | $y$ | $y_{verb}$ |
|---|---|---|
| The Bay Bridge | San Francisco | closed |

Figure 1: Sample sentence and potential spatial knowledge generated using syntactic dependencies.

### 3.1. Generating Potential Additional Spatial Knowledge

We enforce the restrictions below to generate potential spatial relations LOCATION($x$, $y$):

1. $y$ is a GPE or LOC named entity;
2. $x$ is a PERSON, FAC, PRODUCT or WORK_OF_ART named entity from the same sentence than $y$;
3. $y$ is reachable from $y_{verb}$, where $y_{verb}$ is the closest verb going up the dependency tree from $y$; and
4. $y_{verb}$ is not the verb *to be* or *to have*.

We defined Restrictions 1–2 because we are interested in locations of named entities and assigning spatial information to other entities (e.g., DATE) is nonsensical. Restriction 3 reduces the annotation effort. Restriction 4, however, was designed after pilot annotations revealed that no temporally-anchored spatial knowledge could be inferred from them.

OntoNotes annotates 19,478 GPEs and 1,858 LOCs (potential $y$s); and 18,823 PERSONs, 1,080 FACs, 734 PRODUCTs, and 1,253 WORK_OF_ARTs (potential $x$s). Pairing all potential $x$s and $y$s within a sentence results in 10,136 pairs (Restrictions 1–2). Enforcing Restriction 3 reduces the number of pairs to 9,351, and enforcing Restriction 4 further reduces the number to 8,775. Out of these 8,775 pairs, 7,029 have a PERSON as $x$, 951 a FAC, 411 a PRODUCT, and 384 a WORK_OF_ART.

Figure 1 presents sample sentence and potential spatial relations generated using dependencies.

### 3.2. Crowdsourcing Annotations

After generating potential spatial knowledge, it must be validated manually. To do so, we crowdsourced annotations using CrowdFlower and asking questions in plain English. More specifically, for each potential pair ($x$, $y$), annotators were asked "After reading the sentence above, is $x$ located at $y$ before / during / after $y_{verb}$?" The annotation interface showed the original sentence with $x$ and $y$ highlighted, and no further information. Annotators were instructed to answer questions based on the sentence provided, and to not use prior knowledge about $x$ or $y$.

After pilot annotations, it became clear that answering the above question with `yes` or `no` is suboptimal. First, the question is sometimes nonsensical because (a) $x$ cannot be literally located anywhere, or (b) $y_{verb}$ is a state, thus the meaning of before / during / after $y_{verb}$ is unclear.

Second, sometimes there is not enough information in the sentence to unambiguously determine whether $x$ is or is not located at $y$ with respect to $y_{verb}$. Recall that potential pairs are generated automatically, so some will inevitably be spurious. The final interface forces annotators to choose from one of the following coarse-grained labels for each temporal anchor (before / during / after $y_{verb}$):

- `yes` or `no` if $x$ is (or is *not*) located at $y$ before / during / after $y_{verb}$;
- `inv` if asking the question for $x$ is nonsensical; and
- `unk` if the question is intelligible but the answer is unknown, i.e., neither `yes` nor `no`.

Additionally, if the coarse-grained label is `yes`, annotators had to choose a fine-grained label:

- Before and after: `secs`, `mins`, `days`, `weeks`, `months`, `years`, or `inf` for infinite. They were instructed to choose the longest unit of time possible (e.g., `days` means for a few days but less than a week).
- During: `entire` if $x$ is located at $y$ for the entire duration of $y_{verb}$, `some` otherwise.

Out of the 8,775 ($x$, $y$) pairs automatically generated, we collected three annotations for 25% of pairs per $y_{verb}$ (total: 1,689). Among these relations, 478 belong to type (1) (a: 227, b: 251) and 1,211 belong to type (2) i.e., $x$ and $y$ belong to the same semantic role or are not the heads of any semantic role.

## 4. Corpus Analysis

Figure 2 shows percentages of coarse-grained labels per temporal anchor (before, during, after and all). Overall (bottom right sub figure), only 3.20% questions are invalid, and annotators answered with `yes` or `no` 74.28% of questions (`yes`: 51.77%, `no`: 22.51%), i.e., almost 75% of potential spatial knowledge is deemed correct by annotators. Percentages per named entity type of $x$ follow similar trends overall, but WORK_OF_ART has more `inv` labels (18.05%) than the rest (0.85%–2.87%), and PRODUCT has more `yes` labels than the rest (62.65% vs. 45.83%–53.23%). The percentages per temporal anchor indicate that more temporally-anchored spatial knowledge can be extracted for *before* than *after* (52.87% + 28.89% = 81.76% vs. 52.34 + 22.68% = 75.02%). Also, more potential spatial knowledge can be extracted for *during* than *before* and *after* (63.47% + 21.79% = 85.26%).

Percentages for fine-grained labels are shown in Table 1. For *during*, the vast majority of labels (91.22%) are `entire`, and only 8.77% are `some`. For *before* and *after*, most labels are either `years` (49.28% and 36.72%) or `inf` (34.42% and 45.19%). Other labels (`secs`, `mins`, . . . , `months`) are uncommon (1.10%–10.55%). Because of the unbalanced distribution, we experiment with clustered fine-grained labels $<$`years` and $\geq$`years`.

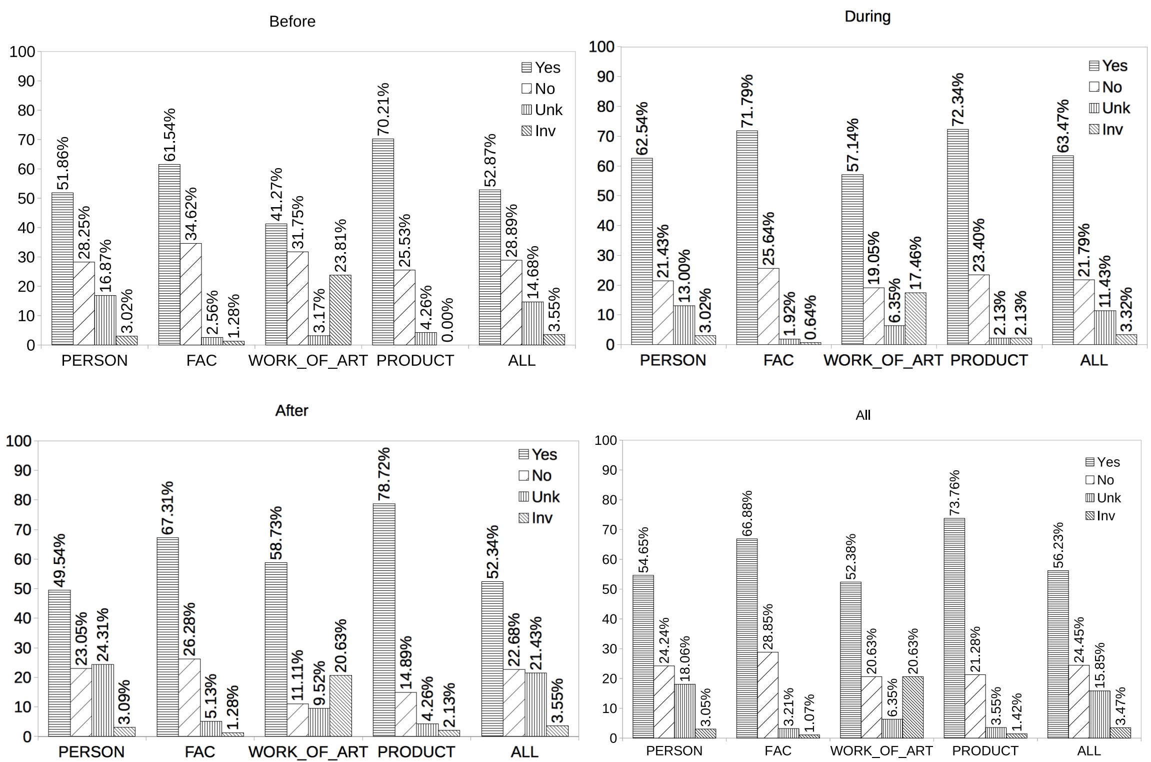



Figure 2: Percentages of coarse-grained labels per temporal anchor (top left: before, top right: during, bottom left: after, and bottom right: all). Percentages are divided by the named entity type of *x*.

| | `some` | `entire` | `secs` | `mins` | `hours` | `days` | `weeks` | `months` | `years` | `inf` |
|---|---|---|---|---|---|---|---|---|---|---|
| Before | n/a | n/a | 2.48 | 1.83 | 6.39 | 2.09 | 2.09 | 1.43 | 49.28 | 34.42 |
| During | 8.77 | 91.22 | n/a | n/a | n/a | n/a | n/a | n/a | n/a | n/a |
| After | n/a | n/a | 1.30 | 2.34 | 10.55 | 2.73 | 1.69 | 2.47 | 36.72 | 42.19 |
| All | 3.65 | 37.94 | 1.10 | 1.22 | 4.95 | 1.41 | 1.10 | 1.13 | 25.11 | 22.37 |

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

| | $r$ | % instances such that | | |
|---|---|---|---|---|
| | | 3/3 | 2/3 | 0/3 |
| Before | 0.62 | 31.20 | 56.42 | 12.37 |
| During | 0.63 | 36.64 | 51.56 | 11.78 |
| After | 0.59 | 26.34 | 60.86 | 12.78 |
| All | 0.59 | 31.39 | 56.28 | 12.31 |

Table 2: Annotation quality for coarse-grained labels. We show weighted Pearson correlations($r$) between annotators and the majority label, and percentage of pairs (*x*, *y*) for which 3, 2 and none of the annotators agree (out of 3).

### 4.1. Annotation Quality

A majority coarse-grained label exists in over 87% of pairs (Table 2, 3/3 or 2/3 agreed). We calculated weighted Pearson correlation between annotators and the majority label following this mapping: `yes`:1, `no`:-1, `unk` and `inv`:0; correlations range between 0.59 and 0.62. Fleiss Kappa agreements (not shown in Table 2) range between 0.51 and 0.55, which are considered moderate (Landis and Koch, 1977). We believe Pearson is better suited than Kappa, as not all disagreement are equally bad (e.g., `yes` vs. `no` and `unk` vs. `inv`).

| | $r$ | % instances such that | | | | |
|---|---|---|---|---|---|---|
| | | 3/3 | 2/3 | 0/3 | 2/2 | 0/2 |
| Before | 0.62 | 39.43 | 46.69 | 13.88 | 88.53 | 11.47 |
| During | 0.60 | 38.57 | 48.62 | 12.81 | 84.94 | 15.06 |
| After | 0.58 | 29.67 | 54.64 | 15.68 | 87.78 | 12.22 |
| All | 0.62 | 36.19 | 49.8 | 14.01 | 87.04 | 12.96 |

Table 3: Annotation quality for fine-grained labels. We show weighted Pearson correlations($r$) between annotators and the majority label and percentages of pairs (*x*, *y*) for which annotators agree. We divide the percentages into pairs in which 3 or 2 annotators agreed on the coarse-grained label *yes* (3/3, 2/3 or 0/3; and 2/2 or 0/2).

Table 3 presents a similar analysis for fine-grained labels. A majority label exists in 86% of pairs when 3/3 annotators agreed on the coarse-grained label `yes`, and 87.04% when 2/3 annotators agreed. Pearson correlations range from 0.58 to 0.60, and overall Kappa is 0.56 (not shown).

### 4.2. Annotation Examples

In table 4, we present real examples from the annotated corpus. In Sentence 1, the annotators chose the label `years`

| Sentence | Before | | During | | After | |
|---|---|---|---|---|---|---|
| | C | F | C | F | C | F |
| Sentence 1: [A {US}$^{\text{GPE}}$ poll]$_{\text{ARG}_0}$ [shows]$_{verb}$ [President {Clinton}$^{\text{PERSON}}$ and his wife, {First}$^{\text{ORDINAL}}$ Lady {Hillary Rodham Clinton}$^{\text{PERSON}}$ are the man and woman most admired by {Americans}$^{\text{NORP}}$]$_{\text{ARG}_1}$. | | | | | | |
| *x*: Hillary Rodham Clinton, *y*: US, $y_{verb}$: *shows* | yes | years | yes | entire | yes | years |
| *x*: President Clinton, *y*: US, $y_{verb}$: *shows* | yes | years | yes | entire | yes | years |
| Sentence 2: [{Venezuela}$^{\text{GPE}}$'s leftist President]$_{\text{ARG}_0}$ has [awarded]$_{verb}$ [{Fidel Castro}$^{\text{PERSON}}$]$_{\text{ARG}_2}$ [the key to the city of {Caracas}$^{\text{GPE}}$]$_{\text{ARG}_1}$. | | | | | | |
| *x*: Fidel Castro, *y*: Caracas, $y_{verb}$: awarded | yes | days | yes | entire | yes | days |
| *x*: Fidel Castro, *y*: Venezuela, $y_{verb}$: awarded | no | n/a | no | n/a | no | n/a |
| Sentence 3: On {Capitol Hill}$^{\text{LOC}}$ {today}$^{\text{DATE}}$, senators were asking [the general who]$_{\text{ARG}_0}$ [sent]$_{verb}$ [{the "USS Cole"}$^{\text{WORK\_OF\_ART}}$]$_{\text{ARG}_1}$ [into the port of {Aden}$^{\text{GPE}}$ in {Yemen}$^{\text{GPE}}$]$_{\text{ARG}_2}$, why he made that decision. | | | | | | |
| *x*: "USS Cole", *y*: Yemen, $y_{verb}$: sent | no | n/a | no | n/a | yes | weeks |
| *x*: "USS Cole", *y*: Aden, $y_{verb}$: sent | no | n/a | no | n/a | yes | weeks |
| *x*: "USS Cole", *y*: Capitol Hill, $y_{verb}$: asking | no | n/a | no | n/a | no | n/a |
| Sentence 4: [{Yesterday}$^{\text{DATE}}$]$_{\text{ARGM-TMP}}$, [{Afghanistan}$^{\text{GPE}}$'s ruling {Taliban}$^{\text{ORG}}$]$_{\text{ARG}_0}$ [denied]$_{verb}$ [{Bin Laden's}$^{\text{PERSON}}$ involvement in the {Yemeni}$^{\text{NORP}}$ attack]$_{\text{ARG}_1}$. | | | | | | |
| *x*: Bin Laden, *y*: Afghanistan, $y_{verb}$: denied | unk | n/a | unk | n/a | unk | n/a |

Table 4: Annotation examples for the generated pairs. We show coarse- and fine-grained annotations (C and F respectively); $y_{verb}$ denotes the first verb going up the dependency tree from y, curly brackets and superindices indicate named entities, and square brackets and subindices indicate semantic roles of $y_{verb}$

for before, after and `entire` for during because *Hillary Rodham Clinton* (*x* in pair 1) and President *Clinton* (*x* in pair 2) will be located in *US* (*y*) with respect to the $y_{verb}$ *shows* for all the three temporal anchors. The label `years` can also be justified because the sentence states that *Hillary Rodham Clinton* is the wife of *US President Clinton*. Also, this is a type 2 spatial relation since *x* and *y* are not semantic roles of any verb.

From Sentence 2, we can say that *Fidel Castro* (*x*) is located in the *Caracas* city at least for a few `days` before, after and for the `entire` time during $y_{verb}$ *awarded* took place. Also, the annotators correctly interpreted that *Fidel Castro* is not located in *Venezuela* before, during and after *awarded* took place.

From Sentence 3, the annotators could infer that the *USS Cole* (*x*) is not located in *Yemen* (*y* in pair 1) or *Aden* (*y* in pair 2) before and during $y_{verb}$ *sent* took place. They also inferred that it will be located in *Yemen* and *Aden* at least for a few `weeks` after *sent* took place. Also, it can be justified that *USS Cole* is not present in *Capitol Hill* before, during and after $y_{verb}$ *asking* took place.

In Sentence 4, we cannot say anything about the location of *Bin Laden* (*x*) from the sentence, so the annotators choose the label `unk` for all the three temporal anchors.

## 5. Experiments

In this section, we present learning experiments with the corpus. Each LOCATION(*x*, *y*) relation has three labels corresponding to temporal anchors before, during and after, thus we generate 3 instances per relation. We perform two classification tasks: (1) coarse-grained classification to predict `yes`, `no` or `unk`, and (2) clustered fine-grained classification to predict $\geq$ `years`, `<years`, `no` or `unk`. This classification is inspired by the previous work by Pan et al. (2006) who predict event durations.

We discard all the instances with `inv` label. We found it advantageous to train one classifier per temporal tag. We divide the instances into 80% train and 20% test by ensuring all the LOCATION(*x*, *y*) relations generated from a particular sentence belong to either the train or test sets. We use scikit-learn (Pedregosa et al., 2011) to train one SVM per temporal anchor and tune the C and $\gamma$ parameters using 10-fold cross-validation and grid search over the train set. Results are reported on the corresponding test set.

Table 5 lists the feature set we experiment with. *Basic* features encode basic information regarding argument *x*, location *y*, x_*verb* and y_*verb*. NE features categorize the argument *x* and location *y* based on their named entity types. Syntax features capture dependency structure of *x* and *y*. Semantic features add information regarding spatial and temporal roles.

## 6. Results

We performed experiments using gold-standard linguistic annotations as well as predicted linguistic annotations. The gold POS tags, parse-trees, semantic roles, dependencies and named entities are taken from the CoNLL release and the predicted linguistic information is obtained using SyntaxNet (Andor et al., 2016). The baseline systems predict the most frequent label per temporal anchor and obtain an overall F-score of 0.46 (0.29 for before, 0.49 for during and 0.29 for after).

Results with coarse-grained and clustered fine-grained labels obtained with all features per temporal anchor using gold standard linguistic information are presented in Table 6. Models trained with all features perform best with respect to all temporal anchors. In general, results with before and during are better than results with after.

Results with coarse-grained and clustered fine-grained labels obtained with all features per temporal anchor using predicted linguistic information is presented in Table 7. The

| Type | Feature | Description |
|---|---|---|
| Basic | 1 | whether $x$ occurs *before* or *after* $y$ |
| | 2 | number of tokens between $x$ and $y$ |
| | 3–4 | number of tokens in $x$ and $y$ |
| | 5–8 | words and POS tags of the heads of $x$ and $y$ |
| | 9–16 | words and POS tags of the first and last tokens in $x$ and $y$ |
| | 17–24 | words and POS tags of the previous and next tokens to $x$ and $y$ |
| | 25–26 | word and POS tag of the closest verb to y going up the dependency tree |
| NE | 27–28 | named entity types of x and y |
| Syntax | 29–30 | outgoing dependencies from heads of x and y |
| | 31–34 | outgoing dependencies from first and last tokens of x and y |
| | 35–38 | outgoing dependencies from previous and next tokens to x and |
| Semantic | 39–40 | counts of ARGM-LOC and ARGM-TMP roles in the sentence |

Table 5: Feature set to determine whether $x$ is (or is not) located at $y$, and for how long.

| | | Before | | | During | | | After | | | Overall | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | P | R | F | P | R | F | P | R | F | P | R | F |
| Coarse-grained | Yes | 0.62 | 0.71 | 0.66 | 0.66 | 0.84 | 0.74 | 0.58 | 0.63 | 0.61 | 0.62 | 0.73 | 0.67 |
| | No | 0.55 | 0.52 | 0.54 | 0.33 | 0.21 | 0.26 | 0.46 | 0.47 | 0.47 | 0.45 | 0.40 | 0.42 |
| | Unk | 0.32 | 0.23 | 0.27 | 0.35 | 0.12 | 0.18 | 0.38 | 0.32 | 0.35 | 0.35 | 0.22 | 0.27 |
| | Avg | 0.54 | 0.56 | 0.55 | 0.54 | 0.60 | 0.56 | 0.50 | 0.51 | 0.50 | 0.53 | 0.56 | 0.54 |
| Clustered fine-grained | ≥years | 0.56 | 0.64 | 0.60 | n/a | n/a | n/a | 0.60 | 0.57 | 0.58 | 0.58 | 0.61 | 0.59 |
| | <years | 0.44 | 0.15 | 0.22 | n/a | n/a | n/a | 0.33 | 0.12 | 0.18 | 0.39 | 0.14 | 0.20 |
| | entire | n/a | n/a | n/a | 0.67 | 0.82 | 0.74 | n/a | n/a | n/a | 0.67 | 0.82 | 0.74 |
| | some | n/a | n/a | n/a | 0.00 | 0.00 | 0.00 | n/a | n/a | n/a | 0.00 | 0.00 | 0.00 |
| | No | 0.55 | 0.60 | 0.58 | 0.30 | 0.28 | 0.29 | 0.48 | 0.66 | 0.55 | 0.44 | 0.51 | 0.47 |
| | Unk | 0.33 | 0.28 | 0.31 | 0.31 | 0.10 | 0.16 | 0.44 | 0.40 | 0.42 | 0.36 | 0.26 | 0.30 |
| | Avg. | 0.51 | 0.52 | 0.51 | 0.51 | 0.57 | 0.52 | 0.50 | 0.51 | 0.50 | 0.51 | 0.53 | 0.51 |

Table 6: Precision recall and F1-score with coarse-grained labels and clustered fine-grained labels using features extracted from gold standard linguistic annotations for the best system per temporal anchor

| | | Before | | | During | | | After | | | Overall | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | P | R | F | P | R | F | P | R | F | P | R | F |
| Coarse-grained | Yes | 0.57 | 0.83 | 0.67 | 0.67 | 0.88 | 0.76 | 0.56 | 0.56 | 0.56 | 0.60 | 0.76 | 0.66 |
| | No | 0.52 | 0.44 | 0.47 | 0.32 | 0.17 | 0.22 | 0.41 | 0.45 | 0.43 | 0.42 | 0.35 | 0.37 |
| | Unk | 0.00 | 0.00 | 0.00 | 0.67 | 0.08 | 0.14 | 0.24 | 0.21 | 0.23 | 0.30 | 0.10 | 0.12 |
| | Avg. | 0.45 | 0.55 | 0.49 | 0.58 | 0.62 | 0.56 | 0.44 | 0.44 | 0.44 | 0.49 | 0.54 | 0.50 |
| Clustered fine-grained | ≥years | 0.57 | 0.57 | 0.57 | n/a | n/a | n/a | 0.60 | 0.39 | 0.47 | 0.59 | 0.48 | 0.52 |
| | <years | 0.43 | 0.18 | 0.25 | n/a | n/a | n/a | 0.25 | 0.17 | 0.20 | 0.34 | 0.18 | 0.23 |
| | entire | n/a | n/a | n/a | 0.64 | 0.88 | 0.74 | n/a | n/a | n/a | 0.67 | 0.82 | 0.74 |
| | some | n/a | n/a | n/a | 0.00 | 0.00 | 0.00 | n/a | n/a | n/a | 0.00 | 0.00 | 0.00 |
| | No | 0.52 | 0.61 | 0.56 | 0.33 | 0.17 | 0.22 | 0.40 | 0.62 | 0.49 | 0.42 | 0.47 | 0.42 |
| | Unk | 0.24 | 0.22 | 0.23 | 0.38 | 0.12 | 0.18 | 0.32 | 0.36 | 0.34 | 0.31 | 0.23 | 0.25 |
| | Avg. | 0.49 | 0.50 | 0.49 | 0.51 | 0.59 | 0.52 | 0.50 | 0.51 | 0.50 | 0.50 | 0.53 | 0.50 |

Table 7: Precision recall and F1-score values for best systems with coarse-grained labels and clustered fine-grained labels using features extracted from predicted linguistic annotations.

coarse-grained and clustered fine-grained results with models trained using predicted linguistic information obtained an overall F1-score of 0.50 (vs. 0.54 and 0.51) with a test set of 655 (vs. 1076, 60% overlap with gold test set). When semantic roles are used extract potential spatial knowledge (Vempala and Blanco, 2016) the overlap between predicted and gold test set is only 30%.

## 7. Conclusions

We have presented an approach to determine whether selected named entities are located or *not* located somewhere, and specify *when* with respect to an event. Crowdsourcing experiments show that annotating this kind of temporally-anchored spatial knowledge can be done by non-experts. Most of the pairs (74.28%, Figure 2) automatically generated are validated by annotators (coarse-grained labels `yes` and `no`). Importantly, working with named entities and syntactic dependencies instead of semantic roles allows us to generate more potential spatial knowledge and obtain better results in a realistic scenario, i.e., with predicted linguistic annotations.

## 8. Bibliographical References

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally normalized transition-based neural networks. *CoRR*, abs/1603.06042.

De Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

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 on XX*, pages 57–60, Morristown, NJ, USA. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Liu, J. and Inkpen, D. (2015). Estimating user location in social media with stacked denoising auto-encoders. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 201–210, Denver, Colorado, June. Association for Computational Linguistics.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August. 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.

Pan, F., Mulkar, R., and Hobbs, J. R. (2006). An annotated corpus of typical durations of events. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 77–82. Citeseer.

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.

Vempala, A. and Blanco, E. (2016). Beyond plain spatial knowledge: Determining where entities are and are not located, and for how long. In *ACL (1)*.

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