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 crowdsourced 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] verb inside the forensic laboratory where scientist are trying to solve this mystery ARG, where ARG indicates the end point of event went. Efforts targeting 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] company headquarters in [New York] 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.

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/
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 ys); and 18,823 PERSONs, 1,080 FACs, 734 PRODUCTs, and 1,253 WORK_OF_ARTs (potential xs). Pairing all potential xs and ys 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 3.2% 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. 52.33% + 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 ≥years.
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).

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.
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 γ 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.

---

**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 $y_{verb}$.

| Sentence 1: [A {US\(GPE\) poll}$_{ARG0}$ shows verb [President {Clinton}$_{PERSON}$ and his wife, {First}$_{ORDINAL}$ Lady Rodham Clinton]$_{PERSON}$ are the man and woman most admired by {Americans}$_{NORP}$]$_{ARG1}$. |
|---|---|---|---|---|---|---|---|
| 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\(^{country}\)}$_{ARG0}$’s left President]$_{ARG0}$ has [awarded]$_{verb}$ [{Fidel Castro}$_{PERSON}$]$_{ARG2}$ the key to the city of {Caracas}$_{GPE}$]$_{ARG1}$. |
|---|---|---|---|---|---|---|---|
| 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\(^{place}\)}$_{ARG0}$ [today\(^{date}\)], senators were asking [the general who]$_{ARG0}$ has [sent]$_{verb}$ [the “USS Cole”\(^{work}$_{ART}$_{OF}\)}$_{ARG1}$ into the port of {Aden}$_{GPE}$ in {Yemen}$_{GPE}$]$_{ARG2}$, why he made that decision. |
|---|---|---|---|---|---|---|---|
| x: “USS Cole”\(^{352}\), y: Yemen, $y_{verb}$: sent | no | n/a | no | n/a | yes | weeks |
| x: “USS Cole”\(^{4}\), y: Aden, $y_{verb}$: sent | no | n/a | no | n/a | no | weeks |
| x: “USS Cole”\(^{9}\), y: Capitol Hill, $y_{verb}$: asking | no | n/a | no | n/a | no | n/a |

| Sentence 4: [{Yesterday\(^{date}\)}$_{ARG0}$ (Yemen\(^{country}\)}$_{ARG0}$’s ruling {Taliban}$_{ORG}$\(^{398}\) in {Afghanistan\(^{country}\)}$_{ARG0}$]$_{ARG1}$, [denied]$_{verb}$ [{Bin Laden’s}$_{PERSON}$]$_{ARG0}$ involvement in the {Yemen}$_{GPE}$]$_{ARG2}$, |
|---|---|---|---|---|---|---|---|
| x: Bin Laden, y: Afghanistan, $y_{verb}$: denied | unk | n/a | unk | n/a | unk | n/a |
experiments show that annotating this kind of temporally-located named entities are located or (Vempala and Blanco, 2016) the overlap between predicted set of 655 (vs. 1076, 60% overlap with gold test set). When an overall F1-score of 0.50 (vs. 0.54 and 0.51) with a test coarse-grained and clustered fine-grained results with mod-
using features extracted from predicted linguistic annotations.

| 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 |
|---------------------|--------|--------|-------|---------|
| Coarse-grained      |        |        |       |         |
| Yes                 | 0.62   | 0.71   | 0.66  | 0.66    |
| No                  | 0.55   | 0.53   | 0.54  | 0.50    |
| Avg                 | 0.54   | 0.56   | 0.55  | 0.53    |
| Clustered fine-grained |      |        |       |         |
| Yes                 | 0.56   | 0.46   | 0.60  | 0.58    |
| No                  | 0.44   | 0.36   | 0.63  | 0.58    |
| Avg                 | 0.51   | 0.52   | 0.51  | 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 |
|---------------------|--------|--------|-------|---------|
| Coarse-grained      |        |        |       |         |
| Yes                 | 0.57   | 0.83   | 0.67  | 0.67    |
| No                  | 0.52   | 0.44   | 0.47  | 0.47    |
| Avg                 | 0.55   | 0.55   | 0.49  | 0.54    |
| Clustered fine-grained |      |        |       |         |
| Yes                 | 0.57   | 0.57   | 0.57  | 0.59    |
| No                  | 0.43   | 0.18   | 0.25  | 0.34    |
| Avg                 | 0.50   | 0.50   | 0.49  | 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.

course-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). Glob-
ally 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 on 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,