# SemEval-2015 Task 8: SpaceEval

**James Pustejovsky**[1], **Parisa Kordjamshidi**[2,3], **Marie-Francine Moens**[2],
**Aaron Levine**[1], **Seth Dworman**[1], **Zachary Yocum**[1]
[1]Brandeis University, Waltham, MA
[2]Katholieke Universiteit Leuven, Belgium
[3]University of Illinois, Urbana/Champaign, IL
{jamesp,zyocum,aclevine,sdworman}@brandeis.edu,
Sien.Moens@cs.kuleuven.be, kordjam@illinois.edu

## Abstract

Human languages exhibit a variety of strategies for communicating spatial information, including toponyms, spatial nominals, locations that are described in relation to other locations, and movements along paths. SpaceEval is a combined information extraction and classification task with the goal of identifying and categorizing such spatial information. In this paper, we describe the SpaceEval task, annotation schema, and corpora, and evaluate the performance of several supervised and semi-supervised machine learning systems developed with the goal of automating this task.

## 1 Introduction

SpaceEval builds on the Spatial Role Labeling (SpRL) task introduced in SemEval 2012 (Kordjamshidi et al., 2012) and used in SemEval 2013 (Kolomiyets et al., 2013). The base annotation scheme of the previous tasks was introduced in (Kordjamshidi et al., 2010), with empirical practices in (Kordjamshidi et al., 2011; Kordjamshidi and Moens, 2015). While those previous tasks are similar in their goal, SpacEval adopts the annotation specification from ISOspace (Pustejovsky et al., 2011a; Moszkowicz and Pustejovsky, 2010; ISO/TC 37/SC 4/WG 2, 2014), a new standard for capturing spatial information. The SpRL in SemEval 2012 had a focus on the main roles of *trajectors*, *landmarks*, *spatial indicators*, and the links between these roles which form *spatial relations*. The formal semantics of the relations were considered at a course-grained level, consisting of three types: directional, regional (topological), and distal. The related annotated data, CLEF IAPR TC-12 Image Benchmark (Grubinger et al., 2006), contained mostly static spatial relations. In SemEval 2013, the SpRL task was extended to the recognition of *motion indicators* and *paths*, which are applied to the more dynamic spatial relations. Accordingly, the data set was expanded and the text from the Degree Confluence Project (Jarrett, 2013) webpages were annotated.

SpaceEval extends the task in several dimensions, first by enriching the granularity of the semantics in both static and dynamic spatial configurations, and secondly by broadening the variety of annotated data and the domains considered. In SpaceEval the concept of *place* is distinguished from the concept of *spatial entity* as a fundamental typing distinction. That is, the roles of trajector (figure) and landmark (ground) are roles that are assigned to spatial entities and places when occurring in spatial relations. Places, however, are inherently typed as such, and remain places, regardless of what spatial roles they may occupy. Obviously, an individual may assume multiple role assignments, and in both ISOspace and SpRL this is assumed to be the case. However, because SpRL focuses on role assignment, it does not introduce the general concept of spatial entity.

There are other differences in the relational schemas of SpRL and SpaceEval which can be easily mapped to each other. For example, in SpRL the general concept of *spatial relation* is defined and the semantics of the relationship (e.g., directional, regional) is added as an attribute of the relation while in SpceEval these semantics introduce new types of relations (e.g., QSLINK and OLINK). In addition to the variations in relational schemas, there are some additional extensions in the SpaceEval annotation. These include augmenting the main elements with more fine-grained attributes. These



*Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 884–894,
Denver, Colorado, June 4-5, 2015. 

attributes, in turn, impact the way the spatial semantics are interpreted. For example, the spatial entities are described with their *dimensionality*, *form*, etc. SpaceEval, also strongly highlights the concepts involved in dynamic spatial relations by introducing *movelink* relations and *motion* tags for annotating motion verbs or nominal motion events and their category from the perspective of spatial semantics. These fine-grained annotations of all the relevant concepts that contribute to grasping spatial semantics makes this scheme and the accompanying corpus unique. The details of the task, including the annotation schema, evaluation configurations, breakdown of the sub-tasks, data set, participant systems, and evaluation results are described in the rest of the paper.

## 2 The Task

The goals of SpaceEval include identifying and classifying items from an inventory of spatial concepts:

- Places: toponyms, geographic and geopolitical regions, locations.
- Spatial Entities: entities participating in spatial relations.
- Paths: routes, lines, turns, arcs.
- Topological relations: *in*, *connected*, *disconnected*.
- Orientational relations: *North*, *left*, *down*, *behind*.
- Object properties: intrinsic orientation, dimensionality.
- Frames of reference: absolute, intrinsic, relative.
- Motion: tracking objects through space over time.

Participants were offered three test configurations for this task.

**Configuration 1** Only unannotated test data was provided.

**Configuration 2** Manually annotated spatial elements, without attributes, were provided.

**Configuration 3** Manually annotated spatial elements, with attributes, were provided.

The SpacEval task is broken down into the following sub-tasks:

**Spatial Elements (SE)**

a. Identify spans of spatial elements including locations, paths, events and other spatial entities.

b. Classify spatial elements according to type: PATH (road, river, highway), PLACE (mountain, village), MOTION (walk, fly), NONMOTION_EVENT (sit, read), SPATIAL_ENTITY (any entity in a spatial relation).

c. Identify their attributes according to type.

**Spatial Signal Identification (SS)**

a. Identify spans of spatial signals (in, on, above).

b. Identify their attributes.

**Motion Signal Identification (MI)**

a. Identify spans of path-of-motion and manner-of-motion signals (arrive, leave, drive, walk).

b. Identify their attributes.

**Motion Relation Identification (MoveLink)**

a. Identify relations between motion-event triggers, motion signals, and motion-event participants (source, goal, landmark, path).

b. Identify their attributes.

**Spatial Configuration Identification (QSLink)**

a. Identify qualitative spatial relations between spatial signals and spatial elements (connected, unconnected, part-of, etc.).

b. Identify their attributes.

**Spatial Orientation Identification (OLink)**

a. Identify orientational relations between spatial signals and spatial elements (above, under, in front of, etc.).

b. Identify their attributes.

## 3 The SpaceBank Corpus

The data for this task are comprised of annotated textual descriptions of spatial entities, places, paths, motions, localized non-motion events, and spatial relations. The data set selected for this task, a subset of the SpaceBank corpus first described in (Pustejovsky and Yocum, 2013), consists of submissions retrieved from the Degree Confluence Project (DCP) (Jarrett, 2013), Berlitz Travel Guides retrieved from

the American National Corpus (ANC) (Reppen et al., 2005), and entries retrieved from a travel weblog, Ride for Climate (RFC) (Kroosma, 2012). The DCP documents are the same set as those annotated with Spatial Role Labeling (SpRL) for SemEval-2013 Task 3 (Kolomiyets et al., 2013), however, for this task, the DCP texts were re-annotated according to ISO-Space.

### 3.1 Annotation Schema

The annotation of spatial information in text involves at least the following: a PLACE tag (for locations and regions participating in spatial relations); a PATH tag (for paths and boundaries between regions); a SPATIAL_ENTITY tag (for spatial objects whose location changes over time); link tags (for topological relations, direction and orientation, frames of reference, and motion event participants); and signal tags (for spatial prepositions)[1]. ISO-Space has been designed to capture both spatial and spatio-temporal information as expressed in natural language texts (Pustejovsky et al., 2012). We have followed a strict methodology of specification development, as adopted by ISO TC37/SC4 and outlined in (Bunt, 2010) and (Ide and Romary, 2004), and as implemented with the development of ISO-TimeML (Pustejovsky et al., 2005) and others in the family of SemAF standards.

SpaceEval's three link tags are as follows:

1. MOVELINK – for movement relations;
2. OLINK – orientation relations;
3. QSLINK – qualitative spatial relations;

QSLINKs are used in ISO-Space to capture topological relationships between tagged elements. The `relType` attribute values come from an extension to the RCC8 set of relations that was first used by SpatialML (Mani et al., 2010). The possible RCC8+ values include the RCC8 values (Randell et al., 1992), in addition to IN, a disjunction of TPP and NTPP.

Orientation links describe non-topological relationships. A SPATIAL_SIGNAL with a DIRECTIONAL `semantic_type` triggers such a link. In contrast to topological spatial relations, OLINK relations are built around a specific frame of reference type and a reference point. The `referencePt` value depends on the `frame_type` of the link. The ABSOLUTE frame type stipulates that the `referencePt` is a cardinal direction. For INTRINSIC OLINKs, the `referencePt` is the same identifier that is given in the `landmark` attribute. For OLINKs with a RELATIVE frame of reference, the identifier for the viewer should be provided as to the `referencePt`.

The following samples from the RFC and ANC sub-corpora have been annotated with a subset of ISO-Space for the SpaceEval task[2]:

1. [**Arriving**$_{m1}$] [**in**$_{ms1}$] the [**town of Juanjui**$_{pl1}$], near the [**park**$_{pl2}$], [**I**$_{se1}$] learned that my map had lied to me.

```
<MOTION id=m1 extent=``Arriving''
motion_type=PATH motion_class=REACH
motion_sense=LITERAL>
<MOTION_SIGNAL id=ms1 extent=``in''
motion_signal_type=PATH>
<PLACE id=pl1 extent=``town of
Juanjui'' form=NAM countable=TRUE
dimensionality=AREA>
<PLACE id=pl2 extent=``park'' form=NAM
countable=TRUE dimensionality=AREA>
<SPATIAL_ENTITY id=se1 extent=``I''
form=NOM countable=TRUE
dimensionality=VOLUME>
<MOVELINK id=mvl1 trigger=m1
goal=pl1 mover=se1 goal_reached=TRUE
motion_signalID=ms1>
```

2. Just [**south of**$_{s1}$] [**Ginza**$_{pl3}$] itself, as [**you**$_{se2}$] [**walk**$_{m2}$] [**toward**$_{ms2}$] the [**bay**$_{pl4}$], you see [**on**$_{s2}$] your [**left**$_{pl5}$] the red [**lanterns**$_{se4}$] and long [**banners**$_{se5}$] of the [**Kabuki-za**$_{pl6}$]...

```
<SPATIAL_SIGNAL id=s1 extent=``south
of'' semantic_type=DIRECTIONAL>
<PLACE id=pl3 extent=``Ginza''
form=NAM countable=TRUE
dimensionality=AREA>
<SPATIAL_ENTITY id=se2 extent=``you''
form=NOM countable=TRUE
dimensionality=VOLUME>
<MOTION id=m2 extent=``walk''
motion_type=COMPOUND
motion_class=REACH
motion_sense=LITERAL>
<MOTION_SIGNAL id=ms2
extent=``toward''
motion_signal_type=PATH>
<PLACE id=pl4 extent=``bay'' form=NAM
countable=TRUE dimensionality=AREA>
<PLACE id=pl5 extent=``left'' form=NAM
countable=TRUE dimensionality=AREA>
<SPATIAL_ENTITY id=se4
```

[1] For more information, cf. (Pustejovsky et al., 2012).

[2] The MEASURE and MLINK tags were not a part of this task.

```
extent=''lanterns'' form=NAM
countable=TRUE dimensionality=VOLUME>
<SPATIAL_ENTITY id=se5
extent=''banners'' form=NAM
countable=TRUE mod=''long''
dimensionality=VOLUME>
<PLACE id=pl6 extent=''Kabuki-za''
form=NAM countable=TRUE
dimensionality=VOLUME>
<OLINK id=ol1 trajector=m2
landmark=pl3 trigger=s1
frame_type=ABSOLUTE referencePt=SOUTH
projective=FALSE>
<MOVELINK id=mvl2 trigger=m2
mover=se2 goal=pl4 goal_reached=NO
motion_signalID=ms2>
<QSLINK id=qsl1 trigger=s2
trajector=se5 landmark=pl5 relType=IN>
<QSLINK id=qsl2 trigger=s2
trajector=se6 landmark=pl5 relType=IN>
```

Since SpaceEval is building on the SpRL shared tasks, we opted to retain the `trajector` and `landmark` attributes for labeling the participants in QSLINK and OLINK relations. This is a deviation from the ISO-Space (Pustejovsky et al., 2011b) standard, which specifies `figure` and `ground` labels based on cognitive-semantic categories explored in the semantics of motion and location by Leonard Talmy (Talmy, 1978; Talmy, 2000) and others. ISO-Space adopted the `figure`/`ground` terminology to identify the potentially asymmetric roles played by participants within spatial relations. For MOVELINKs, however, we distinguish the notion of a `figure`/`trajector` with the ISO-Space `mover` attribute label.

### 3.2 Corpus Statistics

Table 1 includes corpus statistics broken down into the ANC, DCP, and RFC sub-corpora in addition to the train:test partition (∼3:1). The counts of document, sentence, and lexical tokens are tabulated as well as counts of each annotation tag type.

### 3.3 Annotation and Adjudication

All annotations for this task were of English language texts and all annotations were created and adjudicated by native English speakers. Due to dependencies of link tag elements on extent tag elements, the annotation and adjudication tasks were broken down into the following phases:

**Phase 1** Extent tag span and attribute annotation.

| | Sub-corpus | | | Partition | | |
|---|---|---|---|---|---|---|
| | **ANC** | **DCP** | **RFC** | **Train** | **Test** | **Total** |
| **words** | 1577 | 7673 | 21048 | 24150 | 6148 | 30298 |
| **sents** | 61 | 369 | 821 | 1001 | 250 | 1251 |
| **docs** | 3 | 22 | 44 | 55 | 14 | 69 |
| **pl** | 148 | 691 | 1250 | 1661 | 428 | 2089 |
| **se** | 34 | 461 | 1175 | 1347 | 323 | 1670 |
| **qsl** | 69 | 348 | 693 | 886 | 224 | 1110 |
| **mvl** | 15 | 345 | 614 | 779 | 195 | 974 |
| **m** | 16 | 330 | 588 | 751 | 183 | 934 |
| **s** | 39 | 216 | 550 | 653 | 152 | 805 |
| **ms** | 17 | 260 | 365 | 508 | 134 | 642 |
| **p** | 19 | 246 | 278 | 415 | 128 | 543 |
| **e** | 14 | 66 | 301 | 321 | 60 | 381 |
| **ol** | 14 | 82 | 191 | 225 | 62 | 287 |

pl=PLACE; se=SPATIAL_ENTITY; qsl=QSLINK; mvl=MOVELINK; m=MOTION; s=SPATIAL_SIGNAL; ms=MOTION_SIGNAL; p=PATH; e=NONMOTION_EVENT; ol=OLINK

Table 1: Corpus Statistics

**Phase 2** Extent tag adjudication.

**Phase 3** Link tag argument and attribute annotation.

**Phase 4** Link tag adjudication.

Phases 2 and 4 produced gold standards from annotations in the preceding annotation phases. This annotation strategy ensured that the intermediate gold standard extent tag set was adjudicated before any link tag annotations were performed.

The annotation and adjudication effort was conducted at Brandeis University using Multi-document Annotation Environment (MAE) and Multi-annotator Adjudication Interface (MAI) (Stubbs, 2011). We used MAE to perform each phase of the annotation procedure and MAI to adjudicate and produce gold standard standoff annotations in XML format. In addition to the ISO-Space annotation tags and attributes, as a post-process, we also provided sentence and lexical tokenization as a separate standoff annotation layer in the XML data for the training and test sets.

Each document was covered by a minimum of three annotators for each annotation phase (though not necessarily the same annotators per phase). As such, we report inter-annotator agreement (IAA) as a mean Fleiss's $\kappa$ coefficient for all extent tag types annotated in Phase 1, and individual kappa scores for each of the three link tag types annotated in

Phase 3 in Table 2. The scores for extent tags and MOVELINK indicate high agreement, however link tag annotation was less consistent for the remaining link tags. Though the OLINK and QSLINK tag agreement is better than chance, it is not high. We believe the lower agreement for these link tags reflects the complexity of the annotation task.

| Extent Tags | Link Tags | | |
|---|---|---|---|
| **All Types** | **MOVELINK** | **OLINK** | **QSLINK** |
| 0.85 | 0.91 | 0.39 | 0.33 |

Table 2: Overall Fleiss's $\kappa$ Scores

## 4 Evaluation

Participant systems were evaluated for each enumerated configuration as follows:

1 a. SE.a precision, recall, and F1.
   b. SE.b precision, recall, and F1 for each type, and an overall precision, recall, and F1.
   c. SE.c precision, recall, and F1 for each attribute, and an overall precision, recall, and F1.
   d. MoveLink.a, QSLink.a, OLink.a precision, recall, and F1.
   e. MoveLink.b, QSLink.b, OLink.b precision, recall, and F1 for each attribute, and an overall precision, recall, and F1.

2 a. SE.b and SE.c precision, recall, and F1 for each type and its attributes, and an overall precision, recall, and F1.
   b. MoveLink.a, QSLink.a, OLink.a precision, recall, and F1.
   c. MoveLink.b, QSLink.b, OLink.b precision, recall, and F1 for each attribute, and an overall precision, recall, and F1.

3 a. MoveLink.a, QSLink.a, OLink.a precision, recall, and F1.
   b. MoveLink.b, QSLink.b, OLink.b precision, recall, and F1 for each attribute, and an overall precision, recall, and F1.

## 5 Submissions and Results

In this section we evaluate results from runs of five systems. Three systems were submitted by outside groups including Honda Research Institute Japan (HRIJP-CRF-VW), Ixa Group in the University of the Basque Country (IXA), and University of Texas, Dallas (UTD)[3]. We also present results for two systems developed internally at Brandeis University: a suite of logistic regression classifiers with minimal feature engineering intended as a performance baseline covering all sub-tasks in addition to a CRF system with more advanced features, but limited to subtasks 1a and 1b for Configuration 1.

**BASELINE** A suite of logistic regression models using Scikit-learn (Pedregosa et al., 2011) with simple bag-of-words and n-gram features.[4]

**BRANDEIS-CRF** A system using a conditional random field (CRF) model (Okazaki, 2007) with features including Stanford POS and NER tags (Toutanova et al., 2003) (Finkel et al., 2005) in combination with Sparser (McDonald, 1996) tags.[5]

**HRIJP-CRF-VW** A system using a CRF model using CoreNLP, (Manning et al., 2014), CRFSuite (Okazaki, 2007) and Vowpal Wabbit (Langford et al., 2007) with lemmatization, POS, NER, GloVe word vector (Pennington et al., 2014) and dependency parse features.

**IXA** *X-Space*: A system using a binary support vector machine model from SVM-light (Joachims, 1999) and a pipeline architecture using ClearNLP (Choi and Adviser-Palmer, 2012), OpenNLP (OpenNLP, 2014), and leveraging computational linguistic resources including WordNet (Fellbaum, 1998), PropBank (Palmer et al., 2003) and the Predicate Matrix (de la Calle et al., 2014).

**UTD** A suite of 13 classifiers for classifying spatial roles and relations including classifiers for stationary spatial relations and their participants in addition to classification of participants of motion events and their attributes.

[3]UTD submitted three runs, however, after evaluating all the data, all three runs achieved similar scores; the results reported here are for their third and final submitted run.

[4]These baseline classifiers were developed at Brandeis University by Aaron Levine and Zachary Yocum. Cf. Section 5.1 for full description.

[5]This system was developed at Brandeis University by Seth Dworman. Cf. Section 5.2 for full description.

### 5.1 Baseline

Our baseline classification system (BASELINE) consists of a suite of 47 classifiers built from Scikit-learn's (Pedregosa et al., 2011) `sklearn.linear_model` logistic regression package. The system builds a collection of extent objects from the annotation and lexical tokenizations provided in the SpaceEval XML distribution data. Each extent instance has attributes for further feature and label extraction: the target chunk used to form the extent instance; any annotation tag associated with the chunk; lists of all surrounding tokens in the sentence, split between tokens preceding the target and those following, and a pointer to the original annotation XML for the purposes of global feature extraction and generating new XML tags based on the eventual model predictions.

Some extent attributes are optional, depending on the sub-task. E.g., in sub-task 1a, no attributes are required since this sub-task is a simple classification task. For link tags, extent objects are instantiated using the text chunks associated with the extent tags that serve as the link trigger. After pre-processing, the system has a complete collection of extent instances for the corpus.

Subsequent to pre-processing, the extent data are further processed for label and feature extraction. The label and feature extractors were hand-tweaked for each sub-task:

- For extent tag identification, the label extractor checks if a given token occurs at the end of a chunk, and the feature extractors include capitalization and POS tags.
- For classifying extent tag types, the feature extractors include the target chunk string, POS tag, and a seven-token context window (bounded by the sentence) centered on the target token.
- For extent tag attribute classification, the only feature extracted was the text of the chunk associated with the target tag.
- For link tag identification, a heuristic system was developed to select candidate extent tags for the trigger argument. The remaining arguments in the relation were identified by their distance and direction from the trigger. Feature extractors for this process included the text of the trigger chunk, a count of the tags in local context (the same sentence) before and after the trigger, and the types of the extent tags that occur in the context.
- For open-class link tag attributes, feature extractors included the count of extent tags before and after the trigger tag in the sentence. For closed-class link tag attributes feature extractors were limited to the text of the trigger chunk and the trigger tag type.[6]
- For link tag arguments that take an `IDREF` as a value, a unique label function was created that extracts the offsets of the candidate extent tags in the same sentence as the trigger.

The label and feature vectors were maintained using the `DictVectorizer` from Scikit-learn's `feature_extraction` module. To train the system, the vectors were used to fit the model to the training data. For decoding, the tag labels and attributes from the test data were discarded and the remaining feature vectors were transformed into a hypothesis index based on the model, which was translated to a final value using a codebook. The hypotheses were then written out to XML in accordance to the task DTD.

### 5.2 Brandeis CRF

In addition to the BASELINE system, we also developed a more advanced pipeline (BRANDEIS-CRF) to automate the SpaceEval sub-tasks 1a and 1b using a linear-chain conditional random field model using lexical, part-of-speech (POS), named-entity-recognition (NER), and semantic labels. We report overall F1 measures of $0.83$ and $0.77$ for tasks 1a and 1b, respectively, which are comparable to other top results (cf. Section 5.3). Our implementation used the CRFSuite (Okazaki, 2007) open source package, which facilitated rapid training and model inspection. The hypotheses were written out to XML in accordance to the task DTD.

We used a small set of 9 core features, augmented with bigram contexts, resulting in a total of 27 features. These features consist of lexical, syntactic, and semantic information, many of which have

[6] We experimented with additional features for attribute classification, such as counting tags and their types in the local context of the trigger, however additional features all resulted in performance decreases.

been applied successfully in a variety of information extraction tasks (Fei Huang et al., 2014), such as named entity recognition (Vilain et al., 2009b) or coreference resolution (Fernandes et al., 2014). The complete set of features are outlined in Table 3.

| Type | Id | Value |
|---|---|---|
| **Lexical** | word[-1,0,1] | string |
| | isupper[-1,0,1] | binary |
| | wordlen[-1,0,1] | ternary[7] |
| **Syntactic** | pos[-1,0,1] | POS tag |
| **Semantic** | ner[-1,0,1] | NER tag |
| **Sparser** | CATEGORY[-1,0,1] | Sparser category |
| | FORM[-1,0,1] | Sparser form |
| | LCATEGORY[-1,0,1] | Sparser category |
| | LFORM[-1,0,1] | Sparser form |

Table 3: BRANDEIS-CRF Features

For part-of-speech (POS) and named entity (NE) tags, we used the Stanford Log-linear Part-of-Speech Tagger (Toutanova et al., 2003) and the Stanford Named Entity Recognizer (Finkel et al., 2005). Additionally, we made use of Sparser (McDonald, 1996), a rule-based natural language parser in order to provide rich semantic features. Sparser parses unstructured text in cycles, where a variety of handwritten rules apply given the applications of previous rules or the current parse of the text. After parsing, Sparser provides a set of edges, which provide both semantic and syntactic information. For our purposes, we used the `CATEGORY` and `FORM` attributes of the resulting edges. Table 4 shows that the Sparser features can be informative for this task, as five of the top ten positive weights are from Sparser. As a disclaimer, we acknowledge that model weights are not always sufficient for determining the most informative features (Vilain et al., 2009a).

However, there were several problems using Sparser. One issue is that Sparser performs its own internal tokenization and chunking, as it expects unstructured text as input, i.e. a string. To align the already tokenized sentences with a Sparser parse, we used a matching algorithm that aligned a token with its corresponding Sparser edge. A second problem was that Sparser frequently fails on inputs, and the points of failure can be difficult to identify due to the interaction of its various phases and context based rules. Thus, we were not able to get `CATEGORY` and `FORM` for all tokens. As a remedy, we included *local* forms of these Sparser features (prefixed with *L*), which were collected by inputting tokens by themselves to Sparser. This suggests that word lists could be very informative for this task.

[7]Token character length is $\leq 5$, $(5..10]$, or $> 10$.

| Weight | Feature | State |
|---|---|---|
| 3.45 | LCATEGORY=PATH-TYPE | p |
| 2.95 | LCATEGORY=REGION-TYPE | pl |
| 2.66 | word=( | ∅ |
| 2.66 | LCATEGORY=BE | ∅ |
| 2.47 | word=) | ∅ |
| 2.33 | word=near | me |
| 2.28 | word=border | p |
| 2.21 | LCATEGORY=TIME-UNIT | ∅ |
| 2.17 | LCATEGORY=NEAR | me |
| 2.16 | pos=PRP | se |

p=PATH; pl=PLACE; me=MEASURE; se=SPATIAL_ENTITY

Table 4: Top Ten Positive Feature Weights

### 5.3 Evaluation Results

Table 5 shows mean precision (P), recall (R), F1, and accuracy (ACC) scores for each group for each evaluation configuration and sub-task that was attempted. The overall precision and recall measures we report are the arithmetic means of the precision and recall for each tag label or attribute in the corresponding sub-task. The overall, macro-average F1 measures we report are the harmonic mean of the overall P and R. Accuracy is computed as the number of correctly classified labels or attributes divided by the total number of labels or attributes in the gold standard. Overall accuracy and F1 are plotted in Appendix A.

Not all groups attempted all of the evaluation configurations[8]. The HRIJP-CRF-VW system was evaluated only for Configuration 1 tasks 1a, 1b, 1d, and 1e (not 1c), and Configuration 3 sub-tasks 3a and 3b. HRIJP-CRF-VW was not evaluated for Configuration 2 since those sub-tasks were not attempted. The UTD submission only covered Configuration 3, thus was only evaluated for sub-tasks 3a and 3b.

[8]The IXA system was the only one to complete all evaluation configurations.

| System | Task | | P | R | F1 | ACC |
|---|---|---|---|---|---|---|
| **BASELINE** | 1 | a | 0.55 | 0.52 | 0.53 | 0.75 |
| | | b | 0.55 | 0.51 | 0.53 | 0.86 |
| | | c | 0.10 | 0.02 | 0.04 | 0.05 |
| | | d | 0.50 | 0.50 | 0.50 | 0.50 |
| | | e | 0.05 | 0.02 | 0.02 | 0.06 |
| | 2 | a | 0.27 | 0.28 | 0.27 | 0.76 |
| | | b | 0.79 | 0.58 | 0.67 | 0.90 |
| | | c | 0.19 | 0.20 | 0.19 | 0.66 |
| | 3 | a | 0.86 | 0.84 | 0.85 | 0.98 |
| | | b | 0.26 | 0.26 | 0.26 | 0.79 |
| **BRANDEIS-CRF** | 1 | a | 0.85 | 0.80 | 0.83 | 0.89 |
| | | b | 0.78 | 0.76 | 0.77 | 0.92 |
| **HRIJP-CRF-VW** | 1 | a | 0.84 | 0.83 | 0.83 | 0.89 |
| | | b | 0.77 | 0.76 | 0.76 | 0.91 |
| | | d | 0.56 | 0.51 | 0.53 | 0.57 |
| | | e | 0.03 | 0.04 | 0.03 | 0.25 |
| | 3 | a | 0.78 | 0.57 | 0.66 | 0.86 |
| | | b | 0.05 | 0.06 | 0.05 | 0.48 |
| **IXA** | 1 | a | 0.81 | 0.72 | 0.76 | 0.88 |
| | | b | 0.75 | 0.72 | 0.74 | 0.90 |
| | | c | 0.18 | 0.15 | 0.16 | 0.30 |
| | | d | 0.54 | 0.51 | 0.53 | 0.55 |
| | | e | 0.06 | 0.05 | 0.05 | 0.25 |
| | 2 | a | 0.26 | 0.33 | 0.29 | 0.63 |
| | | b | 0.55 | 0.51 | 0.53 | 0.89 |
| | | c | 0.06 | 0.08 | 0.07 | 0.46 |
| | 3 | a | 0.63 | 0.51 | 0.56 | 0.89 |
| | | b | 0.07 | 0.09 | 0.08 | 0.48 |
| **UTD** | 3 | a | 0.87 | 0.82 | 0.85 | 0.98 |
| | | b | 0.05 | 0.09 | 0.07 | 0.51 |

Table 5: Overall Performance

## 6 Conclusion

It is clear from the participating system results that recognizing spatial entities as a sub-task is a fairly well-understood area, with reasonable performance. All systems using CRF models for recognizing places, paths, motion and non-motion events, and spatial entities performed well. Furthermore, MOVELINK recognition results were extremely promising, due to the general tendency for movement to be accompanied by recognizable clues. The overall poor performance for recognition of spatial relations between entities, on the other hand (QSLINKs and OLINKs) indicates that these are difficult relational identification tasks, reflected in the lower IAA scores for these relations as well.

For the next SpaceEval evaluation, we believe that a more focused task, possibly embedded within an application, would lower the barrier to entry in the competition. It would also permit us to use an extrinsic evaluation for performance of the systems. We also hope to release the SpaceBank corpus through LDC later this year. This would enable the community to become more familiar with the dataset and specification.

## Acknowledgements

This research was supported by grants from NSF's IIS-1017765 and NGA's NURI HM1582-08-1-0018. We would like to acknowledge all the annotators and adjudicators who have contributed to the SpaceBank annotation effort, including Adam Berger, Alexander Elias, Alison Marqusee, April Dobkin, Benjamin Beaudett, Eric Benzschawel, Heather Friedman, Katie Glanbock, Keelan Armstrong, Kenyon Branan, Kiera Sarill, Lauren Weber, Martha Schwarz, Meital Singer, Rebecca Loewenstein-Harting, Sean Bethard, and Stephanie Grinley.

## References

Harry Bunt. 2010. A methodology for designing semantic annotation languages exploiting syntactic-semantic iso-morphisms. In *Proceedings of ICGL 2010, Second International Conference on Global Interoperability for Language Resources*.

Jinho D Choi and Martha Adviser-Palmer. 2012. Optimization of natural language processing components for robustness and scalability.

Maddalen López de la Calle, Egoitz Laparra, and German Rigau. 2014. First steps towards a predicate matrix. In *Proceedings of the Global WordNet Conference (GWC 2014), Tartu, Estonia, January. GWA*.

Arun Ahuja Fei Huang, Doug Downey, Yi Yang, Yuhong Guo, and Alexander Yates. 2014. Learning representations for weakly supervised natural language processing tasks. *Computational Linguistics*, 40:1(85-120).

Christiane Fellbaum, editor. 1998. *Wordnet: an electronic lexical database*. MIT Press.

Eraldo Rezende Fernandes, Cícero Nogueira dos Santos, and Ruy Luiz Milidiú. 2014. Latent trees for coreference resolution. *Computational Linguistics*, 40:4(801-835).

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR benchmark: A new evaluation resource for visual information systems. In *Int. Conf. on Language Resources and Evaluation, LREC'06*.

Nancy Ide and Laurent Romary. 2004. International standard for a linguistic annotation framework. *Natural Language Engineering*, 10(3-4):211–225.

Kiyong Lee ISO/TC 37/SC 4/WG 2, Project leaders: James Pustejovsky. 2014. Iso 24617-7:2014 language resource management - part 7: Spatial information (isospace). ISO/TC 37/SC 4/WG 2.

Alex Jarrett. 2013. The degree confluence project. *Retrieved August, 2013, http://www.confluence.org*.

Thorsten Joachims. 1999. Svmlight: Support vector machine. *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, 19(4).

Oleksandr Kolomiyets, Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 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 255–266.

Parisa Kordjamshidi and Marie-Francine Moens. 2015. Global machine learning for spatial ontology population. *Web Semantics: Science, Services and Agents on the World Wide Web*, 30(0):3 – 21. Semantic Search.

Parisa Kordjamshidi, Marie-Francine Moens, and Martijn van Otterlo. 2010. Spatial role labeling: task definition and annotation scheme. In *Proceedings of LREC 2010 - The sevent international conference on langauge resources and evaluation*.

Parisa Kordjamshidi, Martijn van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: towards extraction of spatial relations from natural language. *ACM - Transactions on Speech and Language Processing*, 8:1–36.

Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 2012. Semeval-2012 task 3: Spatial role labeling. 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*, pages 365–373. Association for Computational Linguistics.

David Kroosma. 2012. Ride for climate. *Retrieved September, 2012, http://rideforclimate.com/blog/*.

John Langford, L Li, and A Strehl. 2007. Vowpal wabbit. *URL https://github. com/JohnLangford/vowpal_wabbit/wiki*.

Inderjeet Mani, Christy Doran, Dave Harris, Janet Hitzeman, Rob Quimby, Justin Richer, Ben Wellner, Scott Mardis, and Seamus Clancy. 2010. Spatialml: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3):263–280, September.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 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.

David McDonald. 1996. Internal and external evidence in the identification and semantic categorization of proper names. *Corpus processing for lexical acquisition*, pages 21–39.

Jessica L. Moszkowicz and James Pustejovsky. 2010. Iso-space: towards a spatial annotation framework for natural language. *Processing Romanian in Multilingual, Interoperational and Scalable Environments*.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

Apache OpenNLP. 2014. Apache software foundation. *URL http://opennlp. apache. org*.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2003. The proposition bank: an annotated corpus of semantic roles. *Computational Linguistics*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12.

James Pustejovsky and Zachary Yocum. 2013. Capturing motion in iso-spacebank. In *Workshop on Interoperable Semantic Annotation*, page 25.

James Pustejovsky, Robert Knippen, Jessica Littman, and Roser Saurí. 2005. Temporal and event information in natural language text. *Language Resources and Evaluation*, 39:123–164, May.

James Pustejovsky, Jessica L. Moszkowicz, and Marc Verhagen. 2011a. Iso-space: the annotation of spatial information in language. In *Proceedings of ISA-6: ACL-ISO International Workshop on Semantic Annotation*, Oxford, England, January.

James Pustejovsky, Jessica L. Moszkowicz, and Marc Verhagen. 2011b. Iso-space: The annotation of spatial information in language. In *Proceedings of ISA-6: ACL-ISO International Workshop on Semantic Annotation*, Oxford, England, January.

James Pustejovsky, Jessica Moszkowicz, and Marc Verhagen. 2012. A linguistically grounded annotation language for spatial information. *TAL*, 53(2).

David Randell, Zhan Cui, and Anthony Cohn. 1992. A spatial logic based on regions and connections. In Morgan Kaufmann, editor, *Proceedings of the 3rd Internation Conference on Knowledge Representation and REasoning*, pages 165–176, San Mateo.

Randi Reppen, Nancy Ide, and Keith Suderman. 2005. American national corpus (anc). *Linguistic Data Consortium, Philadelphia. Second release*.

Amber Stubbs. 2011. Mae and mai: Lightweight annotation and adjudication tools. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 129–133. Association for Computational Linguistics.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Marc Vilain, Jonathan Huggins, and Ben Wellner, 2009a. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, chapter A simple feature-copying approach for long-distance dependencies, pages 192–200. Association for Computational Linguistics.

Marc Vilain, Jonathan Huggins, and Ben Wellner. 2009b. Sources of performance in crf transfer training: a business name-tagging case study. In *Proceedings of the International Conference RANLP-2009*, pages 465–470. Association for Computational Linguistics.

## A. Performance Plots

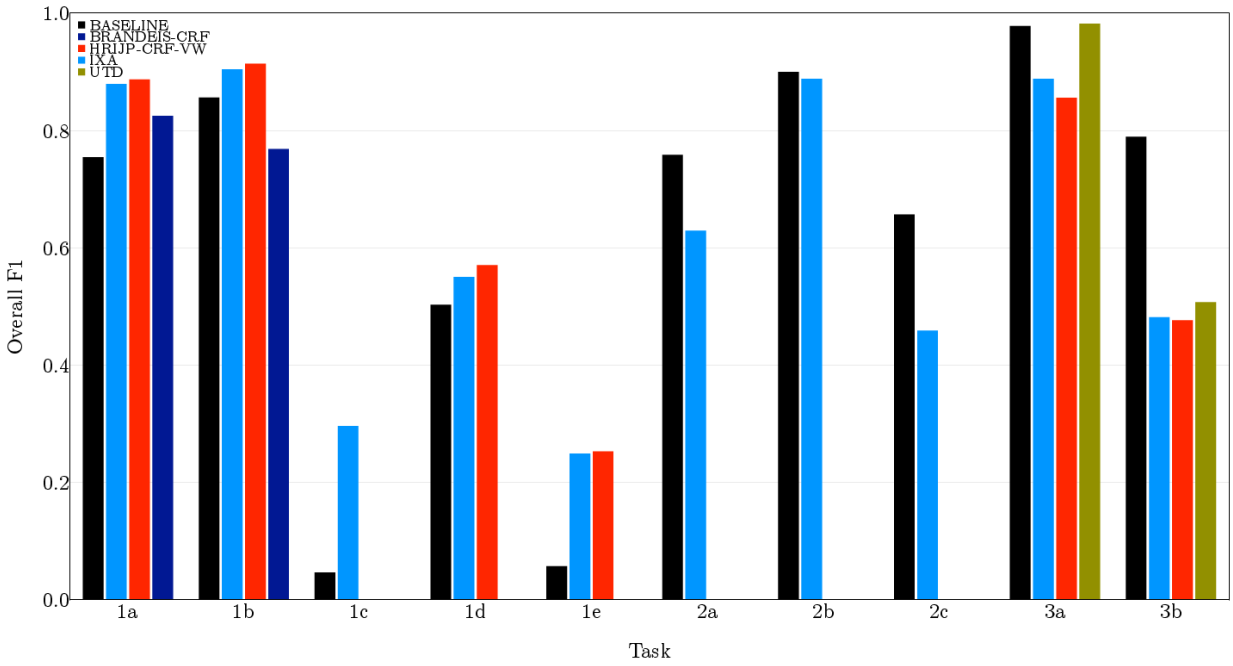


Figure 1: Overall Accuracy for All Sub-tasks

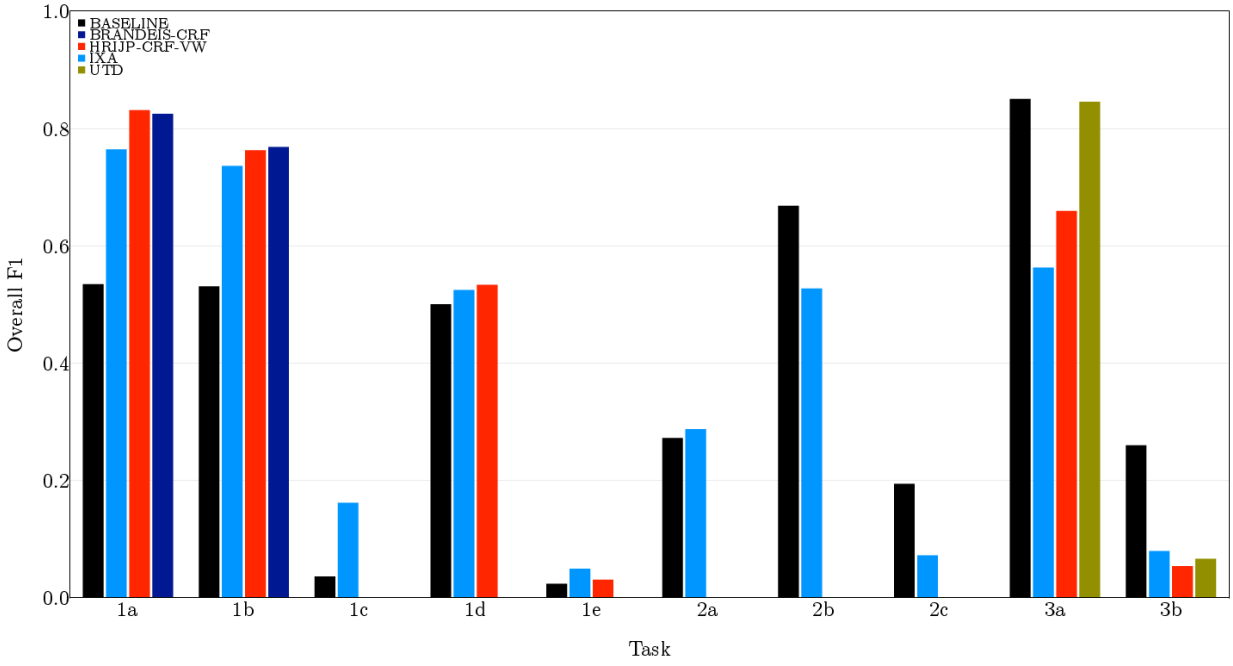


Figure 2: Overall F1 for All Sub-tasks