SpatialML: Annotation Scheme, Corpora, and Tools

Inderjeet Mani, Janet Hitzeman, Justin Richer, Dave Harris, Rob Quimby, and Ben Wellner

The MITRE Corporation
202 Burlington Road
Bedford, MA 01730, USA

E-mail: {imani, hitz, jricher, drh, rquimby, wellner} at mitre dot org

Abstract

SpatialML is an annotation scheme for marking up references to places in natural language. It covers both named and
nominal references to places, grounding them where possible with geo-coordinates, including both relative and absolute
locations, and characterizes relationships among places in terms of a region calculus. A freely available annotation editor
has been developed for SpatialML, along with a corpus of annotated documents released by the Linguistic Data
Consortium. Inter-annotator agreement on SpatialML extents is 77.0 F-measure on that corpus, and 92.3 F-measure on a
ProMED corpus. Disambiguation agreement on geo-coordinates is 71.85 F-measure on the latter corpus. An automatic
tagger for SpatialML extents scores 78.5 F-measure. A disambiguator scores 93.0 F-measure. In adapting the extent
tagger to new domains, merging the training data from the above corpus with annotated data in the new domain provides
the best performance.

1. Introduction

While many areas of natural language processing have benefited from annotation schemes, tagged corpora, and
training and evaluation based on these, the problem of understanding spatial references in natural language has
been somewhat neglected in this regard. Such spatial references include both ‘absolute’ references (e.g.,
“Rome”, “Rochester, NY”, “southern Kerala district of Cudallah”), as well as relative references (“thirty miles
north of Boston”, “an underpass beneath Pushkin Square”,
“in the vicinity of Georgetown University”). We have
developed an annotation scheme called SpatialML that
attempts to address these concerns. This paper discusses
the scheme, the annotated corpora, resources, and tools
developed for it. There are two critical aspects that make
this approach especially attractive: (i) the annotation
scheme is compatible with a variety of different standards
(ii) most of the resources and tools used are freely
available.

While our focus is primarily on geography and culturally
relevant landmarks, we expect that these guidelines could
be adapted to other such domains with some extensions,
without changing the fundamental framework. The main
goal of SpatialML is to mark places mentioned in text
(indicated with PLACE tags) and map them to data from
gazetteers and other databases. Semantic attributes such
country abbreviations, country subdivision and
dependent area abbreviations, and  geo-coordinates are
used to help establish such a mapping. SpatialML uses
LINK tags to express relations between places, such as
inclusion between regions, and PATH tags to capture
spatial trajectories for relative locations, involving a
particular direction and/or distance. The SpatialML
guidelines indicate language-specific rules for marking up
SpatialML tags in English, as well as language-independent rules for marking up semantic
attributes of tags. The guidelines also provide a handful of multilingual examples.

In order to make SpatialML easy to annotate by people
without considerable training, the annotation scheme is
kept fairly simple, with straightforward rules for what to
mark and with a relatively “flat” annotation scheme. Here
is an example for the phrase “a building 5 miles east of
Fengshan”:

\[
\text{a <PLACE id="1" type="FAC" form="NOM">building</PLACE>}
\]
\[
\text{<SIGNAL id="2">5 miles</SIGNAL>}
\]
\[
\text{<SIGNAL id="3">east</SIGNAL> of}
\]
\[
\text{<PLACE id="4" type="PPL" country="TW" form="NAM" latLong="22\degree37'N 120\degree21'E">Fengshan</PLACE>}
\]
\[
\text{<PATH id="5" source="4" destination="1" distance="5:mi" direction="E" signals="2 3"/>}
\]

The idea here is that the relative location’s offsets as
described in the text are captured in the tags. The PATH
expresses a relation between a source PLACE and a target
PLACE, qualified by distance and direction attributes.
The framework accommodates both named and nominal
references to places. We have opportunistically drawn the
inventory of different PLACE types (20 in all) from the
much larger thesaurus (211 categories) of the Alexandria
Digital Library (ADL). Here is a more complex example, involving LINKs. The
set of LINK types is derived from the Region Connection
Calculus (RCC8) (Randell et al. 1992, Cohn et al. 1997).
Both English and Chinese versions are shown.

\[
a \text{[town] some [50 miles] [south] of [Salzburg] in}
\]

\[http://sourceforge.net/projects/spatialml\]

\[http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/top.htm\]
Table 1: Link Types

<table>
<thead>
<tr>
<th>LinkType</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN (tangential</td>
<td>[Paris], [Texas]</td>
</tr>
<tr>
<td>and non-tangential proper parts)</td>
<td></td>
</tr>
<tr>
<td>EC (extended connection)</td>
<td>the border between [Lebanon] and [Israel]</td>
</tr>
<tr>
<td>NR (near)</td>
<td>visited [Belmont], near [San Mateo]</td>
</tr>
<tr>
<td>DC (discrete connection)</td>
<td>the [well] outside the [house]</td>
</tr>
<tr>
<td>PO (partial overlap)</td>
<td>[Russia] and [Asia]</td>
</tr>
<tr>
<td>EQ (equality)</td>
<td>[Rochester] and [382044N 0874941W]</td>
</tr>
</tbody>
</table>

Syntactically, SpatialML tries to keep the tag extents as small as possible, to make annotation easier. Premodifiers such as adjectives, determiners, etc. are NOT included in the extent unless they are part of a proper name. For example, for “the river Thames,” only “Thames” is marked, but, for the proper names “River Thames” and “the Netherlands,” the entire phrase is marked. There is no need for tag embedding, since we have non-consuming tags (LINK and PATH) to express relationships between PLACEs. Adjectival forms of proper names (“U.S., “Brazilian”) are, however, tagged in order to allow one to link expressions such as “Georgian” to “capital” in the phrase “the Georgian capital.”

Deictic references such as “here” are not tagged. Non-referring expressions, such as “town” and “city” in “a small town is better to live in than a big city,” aren’t tagged. Also, “city” in “the city of Baton Rouge” is not tagged; the use of such a modifier is simply to indicate a property of the PLACE. In contrast, when “city” does refer, as in “John lives in the city” where “the city,” in context, must be interpreted as referring to Baton Rouge, it is tagged as a place and given the coordinates, etc., of Baton Rouge.

2. Standards Compatibility

SpatialML, it can be seen so far, can be used to ground PLACEs in terms of types and geo-coordinates (the sort of information found in gazetteers), as well as relate places by PATHs and LINKs. These capabilities together make it unique. While novel, it attempts to be compatible with other standards and proposals. It leverages ISO (ISO-3166-1 for countries and ISO-3166-2 for provinces), as well as various proposed standards towards the goal of making the scheme compatible with existing and future corpora.

The SpatialML guidelines are compatible with existing guidelines for spatial annotation and existing corpora within the Automatic Content Extraction (ACE) research program. In particular, we exploit the English Annotation Guidelines for Entities (Version 5.6.6 2006.08.01), specifically the GPE, Location, and Facility entity tags and the Physical relation tags, all of which are mapped to SpatialML tags. In comparison with ACE, SpatialML emphasizes the grounding of spatial locations in terms of geo-coordinates easier. Instead of grouping mentions into classes (called “entities” in ACE), SpatialML simply annotates mentions of place. SpatialML also doesn’t concern itself with referential subtleties like metonymy, including the GPE/non-GPE distinction; the latter has proven to be difficult for humans to annotate. Finally, SpatialML addresses relative locations involving distances and topological relations that ACE ignores.

We also borrow ideas from the Toponym Resolution Markup Language of Leidner (2006), the research of Schilder et al. (2004) and the annotation scheme in Garbin and Mani (2005). The SpatialML annotation scheme can be integrated with the Geography Markup Language (GML) defined by the Open Geospatial Consortium (OGC). Mappings have also been implemented from SpatialML to Google Earth’s Keyhole Markup Language (KML) , and from the output of a commercial

http://projects.ldc.upenn.edu/ace/annotation/2005Tasks.html
http://www.opengis.net/gml/
http://code.google.com/apis/kml/documentation/
geo-tagging tool, MetaCarta®, to SpatialML.

3. System Architecture

3.1 Overview

We have annotated documents in SpatialML using the freely available Callisto7 annotation editor (Figure 1), which includes the SpatialML task extension. The gazetteer used is the Integrated Gazetteer Database (IGDB) (Mardis and Burger 2005) (Sundheim et al. 2006). IGDB integrates together place name data from a number of different resources, including NGA GeoNames, USGS GNIS9, Tipster, WordNet, and a few others. It contains about 6.5 million entries. The ADL Gazetteer Protocol10 is used to access IGDB.

The annotated data is then used (Figure 2) to train a statistical entity tagger and a disambiguator. Both these tools are built on top of the freely available Carafe11 machine learning toolkit. The entity tagger uses a Conditional Random Field learner to mark up PLACE tags in the document, distinguishing between NAM, NOM, and other tags. A disambiguator (discussed below) looks up tagged PLACE mentions against the gazetteer, using a log linear learning model to rank the potential candidates from the gazetteer. Features associated with the PLACE mention as well as those associated with the gazetteer entry are weighted by the learner. PATH and LINK taggers (that recognize relations between PLACE tags, as well as direction, and distances), are then used. The overall pipeline can process any document (including Web documents in HTML, which are converted to XML using TagSoup12), generating SpatialML output. Finally, the SpatialML output can in turn be mapped to KML for display in Google Earth.

3.2 Disambiguator

For each training document, the disambiguator constructs the cross product of each PLACE tag occurrence (i.e., mention) and all applicable gazetteer candidates for that mention. Feature vectors are constructed for each combination, with the feature vector being labeled as positive for gazetteer candidates found in the training document. The features used are comprised of document features, gazetteer features, and joint features. The document features consist of the document id, the mention string, a window of 10 words on each side of the PLACE mention, and whether the mention is the first one in the document. The gazetteer features include the gazetteer id for the particular gazetteer candidate, the PLACE type, State, and Country, and its latitude and longitude. Joint features include the number of gazetteer candidates for the mention, and whether the parent (likewise, the sibling) of the gazetteer entry (e.g., the country if the gazetteer entry was a capital) is in the document features.

For disambiguation, a statistical ranking model is computed, so that for each gazetteer candidate $G_i$ for PLACE mention $M$, a weight vector for $G_i$ normalized against all other candidates for $M$. This is used to compute $Pr(G_i|M)$. More precisely, letting $w_k$ be the weight of feature $f_k$, and $Gaz(M)$ be the set of all candidate gazetteer entries for $M$, we have:

$$Pr(G_i | M) = \frac{\sum_{s_i} w_s^* f_s(G_i, M)}{\sum_{G_j \in Gaz(M)} \sum_{s_j} w_j^* f_j(G_j, M)}$$

At decode time, given a mention $M$ and a set of gazetteer entries for $M$, the decoder finds the $G_i$ that maximizes $Pr(G_i|M)$.

A threshold is used to filter the output candidates. For improved performance, the learned disambiguator is integrated with a postprocessor that enforces 1 sense per discourse. This involves a greedy learning strategy (e.g., if you find a sense (gazetteer entry) for a mention, commit to that sense through all mentions in that doc). Usually,

---

6http://www.metacarta.com/
7http://callisto.mitre.org
8http://gnswww.nga.mil/geonames/GNS/index.jsp
9http://geonames.usgs.gov/pls/gnispublic
10http://www.alexandria.ucsb.edu/downloads/gazprotocol/
11http://sourceforge.net/projects/carafe
12http://ccil.org/~cowan/XML/tagsoup/
just a few iterations suffice.

4. Accuracy

4.1 Corpus

A corpus of 428 ACE documents, originally from the University of Pennsylvania’s Linguistics Data Consortium (LDC), has been annotated in SpatialML. This corpus, drawn mainly from broadcast conversation, broadcast news, news magazine, newsgroups, and weblogs, contains 6338 PLACE tags, of which 4,783 are named PLACEs with geo-coordinates. This ACE SpatialML Corpus (ASC) has been re-released to the LDC, and is available to LDC members (LDC Catalog LDC2008T0313).

4.2 Inter-annotator Agreement

Inter-annotator agreement on SpatialML PLACE tags in the ASC corpus is 77.0 F-measure. Disagreements stemmed from two sources: application of guidelines and use of tools. The guideline application problems included an annotator failing to mark discourse-dependent references like “the state”, as well as specific references like “area” (to be marked as a REGION), incorrectly marking generic phrases like “areas” or “cities”, among others. The disagreement due to tool use has to do with one version of Callisto lacking the ability to carry out inexact string matches for text mentions of places against IGDB entries, including adjectival forms of names (e.g., “Rwandan”), different transliterations (e.g., “Nisarah” vs. “Nisara”), in addition to various alternative ways of looking up a name (“New York, State of” vs. “New York”). Computing agreement on disambiguation in the ASC is underway.

Table 2: Inter-annotator agreement on ProMED

Table 2 shows the agreement on SpatialML attributes for a corpus from ProMED\(^{14}\), an email reporting system for monitoring emerging diseases provided by the International Society for Infectious Diseases. A corpus of 100 documents was annotated by one annotator, of which 41 were re-annotated by another annotator. The agreement on extent is much higher than on the ASC, for two reasons. First, it was carried out much later in the project, with later versions of the tools as well as guidelines. Second, both annotators were expert linguistic annotators, whereas in the first study only one was (her annotations were used for the ASC).

The lower agreement on LatLong is due to different versions of Callisto being used in the study, giving rise to the tool use issues mentioned above. The higher agreement on LatLong compared to Gazref (i.e., IGDB gazetteer id) is a result of not being able to find an entry with a geo-coordinate in IGDB, using the Web instead, or else finding an alternative (redundant) entry in IGDB. These observations re-emphasize the need to take both guidelines and tool training into account during annotation.

4.3 Entity Tagger

The SpatialML entity tagger has an F-measure of 85.0 for tagging extents of names (i.e., form=“NAM”) and 72.0 for tagging extents of nominals (form=“NOM”), in five-fold cross-validation against 700 ACE documents that were auto-converted to SpatialML (i.e., auto-converted without geo-coordinates).

4.4 Disambiguator

The SpatialML trained disambiguator has an F-measure of 93.0, tested in five-fold cross-validation against 253 documents with geo-coordinates from the ASC. This is an impressive result, given the size of the IGDB gazetteer. Large gazetteers increase the degree of ambiguity; for example, there are 1420 matches for the name “La Esperanza” in IGDB. A study by (Garbin and Mani 2005) on 6.5 million words of news text found that two-thirds of the place name mentions that were ambiguous in the USGS GNIS gazetteer were ‘bare’ place names that lacked any disambiguating information in the containing text sentence. This accuracy is good enough for pre-processing. Human annotators using Callisto reported that post-editing tagger output was far more efficient than human annotation from scratch.

We now discuss the impact of different thresholds on disambiguator performance. Two “confidence” measures were computed for selecting a cutoff point between 0 and 1. For each measure, the top candidate would be selected provided that the measure was below the cutoff. That is, lower confidence measures were considered a good sign that the top choice was effectively separated from sibling choices. The measure One is 1 minus the probability Pr(top) for the top item, i.e. the portion of probability associated with the non-selected items. The measure Prop (for ‘Proportion’) is the reciprocal of the product of Pr(top) and the number of candidates, i.e., a low top probability with many choices should be counted the same as a high probability among few choices. The effect of these two confidence measures on the Precision and Recall of the disambiguator is shown in Figure 3.

\(^{13}\) http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T03

\(^{14}\) http://www.promedmail.org
FIGURE 3: Precision and Recall of Confidence Measures

It can be seen that precision increases slightly as the threshold is raised, but that recall drops off sharply as the threshold is raised beyond .9.

Figure 4 shows the Predictive Accuracy of the loglinear model (LogLin) in comparison to various baseline approaches. ParentInText gives a higher prior probability to a candidate with a ‘parent’ in the text, e.g., for a given mention, a candidate city whose country is mentioned nearby in the text. FirstCand selects the very first candidate (profiting from 37% of the mentions that have only one gazetteer candidate). Random randomly selects a candidate. TypePref prefers countries to capitals, or first-order administrative divisions to second-order. These baselines do not fare well, scoring no more than 57. In comparison, LogLin scores 93.4.

FIGURE 4: Disambiguator Predictive Accuracy

5. SpatialML Tagging across Domains

In order to investigate performance across domains, we annotated two other corpora: 100 documents from ProMED, and a corpus of 121 news releases spidered from the U.S. Immigration and Customs Enforcement (ICE) web site15. Our first observation was that results on the other corpora were lower than on ASC. One problem is the need to appropriately zone these other documents through domain-specific pre-processing, such as specialized handling of title, header and signature blocks in ProMED, for example. Another problem is the tendency of the entity tagger to tag place names inside disease-names or other names, e.g., West Nile Virus, Nashville Warbler. The system also did not fare well on disambiguating abbreviations.

The cost of annotating data in a new domain is generally high. We therefore investigated the extent to which taggers trained on the source ASC data could be adapted with varying doses of target domain data (ProMED or ICE) to improve performance. Information from source and target datasets might be aggregated by directly combining the data (Data Merge), or combining trained models (Model Combination), or else by preprocessing the data to generate “generic” and “domain-specific” features -- the latter based on the “Augment” method of Daume III (2007).

Table 3 shows the performance of the entity tagger trained and tested on different datasets and different combination methods. Here the Source data is ASC, and the Target data is either ICE or ProMED.

<table>
<thead>
<tr>
<th></th>
<th>ICE</th>
<th>ProMED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Data Only</td>
<td>85.60</td>
<td>67.54</td>
</tr>
<tr>
<td>Source Data Only</td>
<td>76.77</td>
<td>67.31</td>
</tr>
<tr>
<td>Data Merge</td>
<td>85.88</td>
<td>84.14</td>
</tr>
<tr>
<td>Model Combination</td>
<td>82.52</td>
<td>68.57</td>
</tr>
<tr>
<td>&quot;Augment&quot; Method</td>
<td>85.34</td>
<td>71.42</td>
</tr>
</tbody>
</table>

TABLE 3: F-Measure of Different Aggregation Methods

It can be seen that in both domains, training a single model over the combined data sets yielded strong results. In the ICE domain, which contained a total of 3,477 sample tags that were used for four-fold cross-validation, both the Augment model and the model trained only over ICE data performed comparably to the Data Merge model, while in the ProMED domain, with only 995 sample tags, Data Merge can be seen to clearly outperform all other techniques.

Figure 5 shows the effect of different amounts of target data in the ICE domain on F-Measure under various combination methods. The figure shows that the Data Merge model performs best with relatively low amounts of target data, but as increasing amounts of target data are included, the Augment and the model trained only over ICE data performed comparably to the Data Merge model, while in the ProMED domain, with only 995 sample tags, Data Merge can be seen to clearly outperform all other techniques.

Figure 6 is a similar chart for the ProMED domain. Here, the Data Merge technique is clearly superior to the others, however with the relatively small number of training tags, it’s possible that additional ProMED data would lead to improvement in the other techniques’ scores.

15http://www.ice.gov/
We have described an annotation scheme called SpatialML that focuses on geographical aspects of spatial language. A freely available annotation editor has been developed for SpatialML, along with a corpus (ASC) of annotated documents released by the Linguistic Data Consortium. Inter-annotator agreement on SpatialML extents is 77.0 F-measure on the ASC corpus, and 92.3 F-measure on a ProMED corpus. Disambiguation agreement on geo-coordinates is 71.85 F-measure on the latter corpus.

An automatic tagger for SpatialML extents scores 78.5 F-measure. A disambiguator scores 93.0 F-measure and 93.4 Predictive Accuracy. Training the extent tagger by merging the training data from the ASC corpus along with the target domain training data outperforms training from the target domain alone.

Future work will extend this porting across domains to the disambiguator, and will also evaluate the LINK and PATH taggers. We will also be conducting various inter-annotator studies on these other domains. In joint work with Brandeis University, we will also be integrating SpatialML with TimeML (Pustejovsky et al. 2005) and the Suggested Upper Merged Ontology\(^{16}\) (SUMO).

7. References


\(^{16}\)http://www.ontologyportal.org/