Lang Resources & Evaluation (2010) 44:263–280
DOI 10.1007/s10579-010-9121-0

SpatialML: annotation scheme, resources, and evaluation

Inderjeet Mani · Christy Doran · Dave Harris · Janet Hitzeman · Rob Quimby · Justin Richer · Ben Wellner · Scott Mardis · Seamus Clancy

Published online: 5 May 2010

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, and characterizes relationships among places in terms of a region calculus. A freely available annotation editor has been developed for SpatialML, along with several annotated corpora. Inter-annotator agreement on SpatialML extents is 91.3 *F*-measure on a corpus of SpatialML-annotated ACE documents released by the Linguistic Data Consortium. Disambiguation agreement on geo-coordinates on ACE is 87.93 *F*-measure. An automatic tagger for SpatialML extents scores 86.9 F on ACE, while a disambiguator scores 93.0 F on it. Results are also presented for two other corpora. In

I. Mani (✉) · C. Doran · D. Harris · J. Hitzeman · R. Quimby · J. Richer · B. Wellner · S. Mardis · S. Clancy
The MITRE Corporation, 202 Burlington Road, Bedford, MA 01730, USA
e-mail: imani@mitre.org

C. Doran
e-mail: cdoran@mitre.org

D. Harris
e-mail: drh@mitre.org

R. Quimby
e-mail: rquimby@mitre.org

J. Richer
e-mail: jricher@mitre.org

B. Wellner
e-mail: wellner@mitre.org

S. Mardis
e-mail: mardis@mitre.org

S. Clancy
e-mail: sclancy@mitre.org

adapting the extent tagger to new domains, merging the training data from the ACE corpus with annotated data in the new domain provides the best performance.

Keywords Annotation · Guidelines · Spatial language · Geography · Information extraction · Evaluation · Adaptation

1 Introduction

The problem of understanding spatial references in natural language poses many interesting opportunities and representational challenges. In this paper, we address the problem of extracting information about places, including 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"). Our focus is on grounding such references, where possible, to precise positions that can be characterized in terms of geo-coordinates. We have developed an annotation scheme called SpatialML[1] that addresses this problem, along with tools to automatically construct such an annotation.

Prior research in natural language semantics and artificial intelligence has attempted to map the various ways in which languages conceptualize space to formal, computable representations. The conceptualizations found in natural language descriptions of places involve aspects such as the sizes and shapes of places, as well as their positions and spatial relationships. These relationships can involve topological relations that connect objects to each other, including one being included in the other. Relationships such as "beneath", "to the left", etc., involve orientation relations that hold between a primary object and a reference object (Clementini et al. 1997; Levinson 2006); the relations are based on frames of reference, either centered at the reference object, the viewer, or an external frame of reference (such as a coordinate system based on a geodetic model).

While there has been a great deal of interest in geographical language from the perspective of both Natural Language Processing and Geographical Information Systems (see Bateman (2008) for an overview), there has been a dearth of annotation schemes and guidelines for annotating linguistic data with precise quantitative representations such as geo-coordinates or qualitative representations involving topological and orientation relations. 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. SpatialML also models topological relations between places, such as inclusion or contact between regions. 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 scheme has been applied to annotated corpora in English as well as Mandarin Chinese.

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

From a theoretical standpoint, the advantages of taking an annotation-based approach are that the spatial representational challenges are put to an empirical test, and the performance of annotators can be measured. The creation of SpatialML-annotated corpora allows one to explore in great detail the mapping of individual natural language examples to the particular set of precise spatial representations used in SpatialML, allowing for assessments of existing theories. Further, such annotated corpora can eventually be integrated with formal reasoning tools, testing how well these tools scale up to problem sets derived from natural language. The recording of topological and orientation relations by the annotator provides a first step to support such further inference. In addition to these potential theoretical advantages, there are two practical benefits offered by SpatialML: (1) the annotation scheme is compatible with a variety of different annotation standards, and (2) most of the resources and tools used are freely available. For pragmatic reasons, our focus is on geography and culturally-relevant landmarks, rather than other domains of spatial language.

We discuss the annotation scheme in Sect. 2, followed, in Sect. 3, by an account of the expressiveness of the scheme. In Sect. 4, we illustrate the annotation editor, and describe the annotated corpora. In Sect. 5, we describe our overall system architecture. Section 6 discusses the accuracy of our tools along with inter-annotator agreement. Section 7 concludes.

2 SpatialML annotation scheme

2.1 Annotation model

The SpatialML annotation model consists of locations, marked by PLACE tags around each location mention, and links between them. Locations can have geo-coordinates; these are recorded in a *latLong* attribute of the PLACE tag. Locations can also be restricted by orientation relations; accordingly, the PLACE tag has a *mod* attribute whose value is also drawn from a small inventory of placeholders for orientation. The form of reference in the location mention is also recorded in the PLACE tag: either a proper name (a *form* attribute of type NAM in the PLACE tag) or a nominal (a *form* attribute of type NOM).

Links come in two varieties: the first are relative links (implemented by non-consuming RLINK tags) that relate relative locations to absolute ones, recording any orientation and distance relations stated between them (via *direction* and *distance* attributes on the RLINK). The *direction* attributes have values drawn from the inventory of placeholders for orientation. The frame of reference for the orientation relation is also captured, via the *frame* attribute on the PLACE tag, whose value can be VIEWER, INTRINSIC, or EXTRINSIC. The other type of link relates locations to each other while recording the type of topological relation involved, using a set drawn from the Region Connection Calculus (Randell et al. 1992; Cohn et al. 1997), or RCC. This is implemented using non-consuming LINK tags. Finally, the portions of the text that license a link are marked in a SIGNALS tag; these have no formal status, and can correspond in the case of an RLINK to a

phrase expressing a distance or direction, or in the case of a LINK to a preposition indicating a relation such as inclusion.

2.2 XML examples

The following example has the place marked as being a named place, and in addition, latitude and longitude are filled in, along with the country code for Taiwan.

<PLACE id = "4" country = "TW" form = "NAM" latLong = "22°37′N 120°21′E"> ***Fengshan****</PLACE>*

In the next example, we see a mention that has been tagged as a nominal reference.

a*<PLACE id = "1" form = "NOM">****building****</PLACE>*

Here is an example of the use of the *mod* attribute:

the southern*<PLACE mod = "S" country = "US" form = "NAM">****United States****</PLACE>*.

Consider an example of an RLINK tag, which expresses a relation between a source PLACE and a destination PLACE, qualified by *distance* and *direction* attributes.

a*<PLACE id = "1" form = "NOM">****building****</PLACE>*
*<SIGNAL id = "2" type = "DISTANCE" >****5 miles****</SIGNAL>*
*<SIGNAL id = "3" type = "DIRECTION">****east****</SIGNAL>* ***of***
<PLACE id = "4" country = "TW" form = "NAM" latLong = "22°37′N 120°21′E">
Fengshan*</PLACE>*
<RLINK id = "5" source = "4" destination = "1" distance = "2" direction = "E" frame = "VIEWER" signals = "2 3"/>

Here is an example which illustrates the use of LINK tags. The SIGNAL licensing the LINK is indicated.

an*<PLACE id = "1" form = "NOM">****escarpment****</PLACE>*
*<SIGNAL id = "2">****in****</SIGNAL>*
*<PLACE country = "ZA" id = "3" form = "NAM">****South Africa****</PLACE>*
<LINK source = "1" target = "3" signals = "2" linkType = "IN"/>.

The set of LINK types are shown in Table 1. We will discuss these in more detail in Sect. 3.2.

2.3 Annotation guidelines

In order for humans to carry out SpatialML annotation without considerable training, the annotation scheme is kept fairly simple, with straightforward rules for what to mark and with a relatively "flat" annotation scheme. The tag extents are kept as small as possible. As we have just seen, the extents of modifiers, as

Table 1 Link types

Link type	Example
DC (disconnection)	The [well] outside the [house]
EC (external connection)	The border between [Lebanon] and [Israel]
EQ (equality)	[Rochester] and [382044 N 0874941 W]
PO (partial overlap)	[Russia] and [Asia]
IN (tangential and non-tangential proper parts)	[Paris], [Texas]

expressed by the *mod* attribute, are not tagged. Pre-modifiers 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", "the Netherlands," or "South Africa", the entire phrase is marked. There is no need for tag embedding, since we have non-consuming link tags.

In annotating spatial information, the guidelines allow for the annotator to consult gazetteers to decide, for example, on a place's country and its geo-coordinates,[2] where applicable. This is necessary since humans do not typically know geographical details for most places. However, the annotator is not to use specialized knowledge, such as personal experience from having lived at that location. The annotator must rely solely on the information in the text and in the gazetteer in order to keep the annotation more representative of general geospatial knowledge, and therefore more consistent with the work of other annotators.

Non-referring expressions, such as "town" and "city" in "a small town is better to live in than a big city." are not tagged. In contrast, when "city" does refer, as in "John lives in the city" where "the city" in context, must be interpreted as referring, for example, to Baton Rouge, it is tagged as a place and given the coordinates of Baton Rouge. Deictic references such as "there" and pronouns are not tagged. The reason we do not tag "there" in "John lives there" or "it" in "It has a great climate" is because it would take far too long for an annotator to consider every occurrence of "it" to decide if it is pleonastic or referential, and whether the referential meaning is geographic or not. While each individual mention of a location is tagged, without any coreference, the latter can be determined post hoc in cases where geo-coordinate or gazetteer information is recorded in the PLACE tag.

Natural language allows facilities (schools, ice-cream parlors, buildings, etc.) and vehicles to be coerced into places, as in "I arrived at the *station*." SpatialML, unlike ACE, is ontologically permissive, annotating such entities as PLACEs even when there is no coercing environment. SpatialML also does not concern itself with referential subtleties like metonymy. These simplifications are introduced for practical reasons: (1) determining when these extended senses are intended can be difficult for the annotator (2) even if these extended, non-locative senses of PLACEs

[2] Note that even in situations where it is acceptable for a place to be construed as a point, its punctuality is only an abstraction at some level of resolution.

are intended, location information about the basic locative sense (e.g., where "Paris" is) is still relevant.

3 Expressiveness

SpatialML is not by any means intended to address all of spatial language; the focus here is on references to places and semantic relations between them in the context of geographic language. We first examine how SpatialML compares to other proposed annotation schemes, and then discuss the adequacy of SpatialML to support formal reasoning.

3.1 Comparison with other annotation schemes

SpatialML 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. 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 guidelines are also compatible with existing guidelines for spatial annotation and existing corpora within the Automatic Content Extraction[3] (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 uses a classification scheme that is closer to information represented in gazetteers, thereby making the grounding of spatial locations in terms of geo-coordinates easier. SpatialML also addresses relative locations involving distances and topological relations that ACE ignores. The ACE 'GPE', 'Location', and 'Facility' Entity types can be represented in SpatialML, as are ACE 'Near' Relations. However, SpatialML, unlike ACE, is a 'flat' annotation scheme. Instead of grouping mentions into classes (called "entities" in ACE), SpatialML simply annotates mentions of places. As we have seen, the possibility (considered by ACE) that a location mention is being used metonymically is ignored.

SpatialML can be in addition mapped to the Guidelines for Electronic Text Encoding and Interchange described by the Text Encoding Initiative (TEI).[4] While there are numerous points of similarity between the TEI scheme and SpatialML, there are three main differences: (1) TEI has two variety of tags (placeName and geogName), with the former classified into four types; in contrast, SpatialML is generally more fine-grained, and has (although not discussed here) 20 PLACE types. While settlements, regions, and countries are distinguished by both, the TEI further distinguishes districts (subdivisions of settlements), which SpatialML ignores, and blocs (regions with more than one country), which SpatialML treats as a region. (2) TEI allows for embedded tags, so that a relative place name such as "an X y miles

[3] http://projects.ldc.upenn.edu/ace/annotation/2005Tasks.html.

[4] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ND.html#NDGEOG.

from Z" is marked with a single tag; as we have seen, SpatialML uses only atomic tags but exploits non-consuming linking tags such as LINK and RLINK. (3) the TEI scheme allows for the representation of spatial relations, but does not annotate any particular ones, whereas, as we have seen, SpatialML represents certain topological and orientation relations.

The SpatialML annotation scheme can also be integrated with the Geography Markup Language[5] (GML) defined by the Open Geospatial Consortium (OGC). This integration, which includes the mapping, in an XML layer, of PLACE tag ids with GML tag ids, allows SpatialML to defer details of geographical representation to GML, which is a very rich specification language for structured geographic data. For example, SpatialML allows the latLong feature in a PLACE tag to be any string, including strings with or without decimals that can be parsed into GML coordinates along with appropriate coordinate systems. Likewise, a distance expression tagged with a SIGNAL tag (of type DISTANCE) can be mapped to units of measure specified in GML. Mappings have also been implemented from SpatialML to Google Earth's Keyhole Markup Language (KML), and from the output of a commercial geo-tagging tool, MetaCarta, to SpatialML.

Finally, SpatialML has also been mapped to the Generalized Upper Model[6] (GUM) ontology developed at the University of Bremen. The GUM concepts that are relevant to SpatialML fall under SpatialDistanceModality (both QualitativeDistance, namely 'near' and 'far', and QuantitativeDistance, including CardinalDirectional), as well as RelativeSpatialModality (including Connection, Parthood, and ProjectionRelation). For example, the GUM ProjectionRelation called AboveProjectionExternal maps to the RLINK attribute *direction* with value ABOVE ("above the house"), while AboveProjectionInternal maps to the PLACE tag's *mod* attribute TOP ("the top of the mountain", discussed earlier).

3.2 Adequacy for formal reasoning

SpatialML represents topological relations between places which are viewed as topological regions. These relations are captured in the LINK types, which we now discuss further. DC, EC, EQ, and PO are from the RCC calculus version known as RCC8. IN is not RCC8, but collapses two RCC8 relations, TPP and NTPP (tangential proper part and non-tangential proper part, respectively). IN is in fact the PP "proper part" relation in RCC5. The reason for the collapsing is that it is often difficult for annotators to reliably determine whether the "part" touches or does not touch the including region's border. We also do not include the remaining RCC8 inverse links TPPi and NTPPi, since these can be represented in annotation by swapping arguments, and are in addition likely to confuse annotators. These annotation efficiency considerations leave one with a hybrid calculus involving relations drawn from RCC5 and RCC8. For this reason, the annotation does not use existing XML markup schemes for RCC8 such as SpaceML from Cristiani and Cohn (2002). Nevertheless, RCC was preferred to more expressive representations

[5] http://www.opengis.net/gml/.

[6] http://www.ontospace.uni-bremen.de/linguisticOntology.html.

such as the 9-intersection calculus of Egenhofer and Herring (1990), which has separately been mapped to natural language in the experiments of Rashid et al. (1998). RCC is simpler in terms of the primitive elements of the representation and choices facing an annotator.

The collapsing of TPP and NTPP means that in SpatialML, whether a proper part shares a boundary with its whole or not cannot be distinguished. Further, SpatialML is also limited in the general representation of borders. While SpatialML can represent the fact that Lebanon has a border (via the PLACE tag for Lebanon having a *mod* attribute with value BORDER), it cannot support reasoning about the border, since *mod*s aren't connected via LINKs; the border itself cannot be tagged as a place, as we are treating places as regions in RCC. However, SpatialML can represent the fact that Lebanon and Israel have a common border, i.e., are EC.

From the standpoint of reasoning, further research can determine whether this hybrid calculus has any interesting formal properties. Even though SpatialML represents both regions and 'points' (the latter with geo-coordinates), how best to reason with mixed-granularity representations involving both remains to be explored. Similarly, while geometric approximations of places in terms of minimum bounding rectangles or polygons are common, integrating them with RCC representations (Papadias et al. 1995) can pose problems.

SpatialML also represents a limited number of orientation relations commonly encountered in the geographical language corpora, captured in the values of the RLINK *direction* attribute and in the PLACE tag's *mod* attribute. The coverage here is uneven, but adding a more substantial set of orientation relations to fully cover the distinctions in GUM is clearly possible, for example, mappings for GUM classes such as LateralProjectionExternal, as in "to the left of the sofa", as well as GeneralDirectional, such as "inside the house". However, these orientation relations are not tied to any formal calculi, and when captured by the *mod* attribute, the relations are treated as unary predicates. Using other link tags would be preferable to the use of *mod*, so that binary relations can be expressed.

There are other major lacunae in terms of expressiveness; for example, sets of places ("the Americas") and complex modification ("subtropical and temperate regions of …") are not handled as yet in SpatialML. Addressing the former is feasible, while the latter requires a substantial revision to the way *mod* attributes are represented.

4 Annotation environment and corpora

We have annotated documents in SpatialML using the freely available Callisto[7] annotation editor (Fig. 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,[8] USGS

[7] http://callisto.mitre.org.

[8] http://gnswww.nga.mil/geonames/GNS/index.jsp.

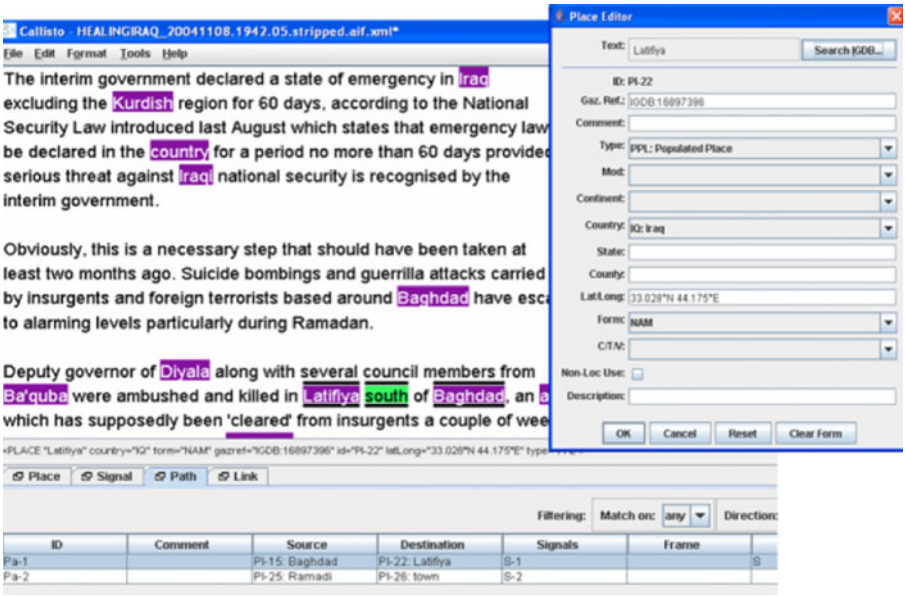

Fig. 1 Callisto editing session

GNIS,[9] Tipster, WordNet, and a few others. It contains about 6.5 million entries. The Alexandria Digital Library (ADL) Gazetteer Protocol[10] is used to access IGDB.

Four corpora have been annotated in SpatialML, chosen because they can either be shared freely, or are sharable under a license from the Linguistic Data Consortium (LDC).The first corpus consists of 428 ACE English documents from the LDC, 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 as LDC2008T03.[11] The second corpus consists of 100 documents from ProMED,[12] an email reporting system for monitoring emerging diseases provided by the International Society for Infectious Diseases. This corpus yielded 995 PLACE tags. The third is a corpus of 121 news releases from the U.S. Immigration and Customs Enforcement (ICE) web site.[13] This corpus provides 3,477 PLACE tags. The fourth corpus is a collection drawn from the ACE 2005 Mandarin Chinese collection (LDC2006T06). So far, 298 documents have been annotated, with 4194 PLACE tags; they will be available through LDC in 2010. The lack of multilingual gazetteers makes the annotation task challenging, given that the annotator tries to lookup a place name in Mandarin Chinese native script. So far, the main language-specific

[9] http://geonames.usgs.gov/pls/gnispublic.

[10] http://www.alexandria.ucsb.edu/downloads/gazprotocol/.

[11] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008T03.

[12] http://www.promedmail.org (we are investigating the possibility of sharing this corpus).

[13] http://www.ice.gov/ (this data can be shared).

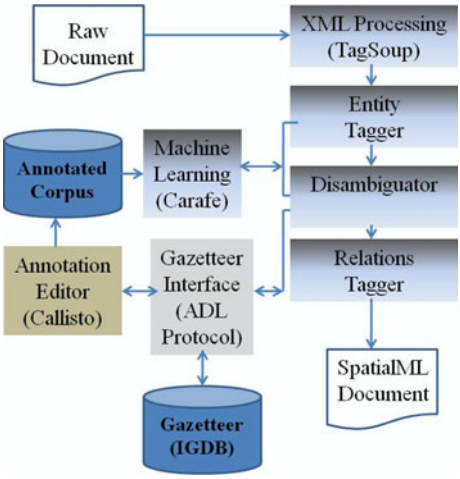

Fig. 2 System architecture

annotation issue that has come up is the need to have sub-word tags in SpatialML, as postpositions can be incorporated into the verb; e.g., "飞往" is a single word meaning "flies towards", calling for a sub-word SIGNAL tag.

5 System architecture

5.1 Overview

The annotated data is used (Fig. 2) to train a statistical Entity Tagger and a disambiguator. Both these tools are built on top of the freely available Carafe[14] machine learning toolkit. The Entity Tagger uses a Conditional Random Field learner to mark up PLACE tags in the document, distinguishing between PLACES and non-PLACEs. The features weighted by the learner include contextual features as well as a feature which checks for membership in a list of place names. For PLACEs, the Entity Tagger also indicates the syntactic form of the mention, i.e., named or nominal (NAM or NOM, respectively). Entity Tagging is a crucial first step, given that distinguishing place names from other kinds of names such as person or organization names can severely reduce the ambiguity early in the pipeline. At this stage, the PLACE tag, even it is a NAM, does not contain geo-coordinates, and nor is it known whether that PLACE is in the gazetteer; these latter decisions are left to the next stage involving the disambiguator.

The 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. The Relations

[14] http://sourceforge.net/projects/carafe.

Tagger then computes LINKs and RLINKs between PLACEs. The overall pipeline can process any document (including Web documents in HTML, which are converted to XML using TagSoup[15]), generating SpatialML output. Finally, the SpatialML output can in turn be mapped to KML for display in Google Earth.

5.2 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 in the cross-product, with the feature vector being labeled as positive if the gazetteer candidate is found in the annotated training document. The features consist 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 an indicator for 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, a 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 is 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 derived by automatic lookup, we have:

$$\Pr(G_i|M) = \frac{e^{\sum_k w_k^* f_k(G_i,M)}}{\sum_{Gj \in Gaz(M)} e^{\sum_k w_k^* f_k(G_j,M)}}$$

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 determine if we have sufficient confidence in the best-ranked sense to provide it as output.

5.3 Relations tagger

The Relations Tagger tags RLINKs and LINKs. It uses a rule-based component that takes a document with SpatialML PLACE tags and locates signals, directions, and distances based on a pre-determined list of common syntactic patterns. It then tags RLINKs and LINKs based on finite-state patterns based on lexical items as well as syntactic chunks. One problem here is data sparseness—we have relatively few RLINKs in our current collection of annotated data.

[15] http://ccil.org/~cowan/XML/tagsoup/.

Table 2 *F*-measure [Precision, Recall] of MIPLACE and human annotators

	ASC		ProMED	
	MIPLACE	HUMAN	MIPLACE	HUMAN
Extent	86.9 [97.3, 78.5]	91.3 [91.86, 90.73]	67.54 [90.35, 53.93]	92.3 [89.32, 95.4]
LatLong	93.0	87.93 [87.76, 88.06]	85.0	71.85 [96.51, 57.22]

As explained in Sect. 6.2, MIPLACE and the HUMAN are evaluated somewhat differently on LatLong, so the comparison here is not direct

6 Evaluation

6.1 Entities

We first discuss the accuracy of human annotation of entities in the ASC and ProMED corpora. These are shown in Table 2, in the row marked 'Extent' and the columns marked 'Human'. In evaluating extents, the start and end offsets of the tag must match exactly across annotators. While the study of agreement on ProMED was carried out early in the project with a single pair of annotators, the agreement study on ASC was carried out much later in the project by three annotators, with the results being an average of their agreement. The ASC results thus reflect improved training as well as guideline maturity. Disagreements stemmed mainly from the guidelines and performance errors; they included (for ProMED) cases such as an annotator failing to mark discourse-dependent references like "the state", as well as specific references like "area", and incorrectly marking generic phrases like "areas" or "cities", among others.

Table 2 also shows the *F*-measure of the MIPLACE Entity Tagger in the row marked 'Extent' and the columns marked 'MIPLACE', along with Precision and Recall shown in brackets. The Entity tagger is trained using the perfect tag extents in the training data. The much poorer performance on ProMED is due to the lack of domain-specific pre-processing, such as specialized handling of title, header and signature blocks, as well as the failure to expand abbreviations. Another problem on ProMED is the tendency of the Entity Tagger to tag place names inside disease-names or other names, e.g., "West Nile Virus", "Nashville Warbler".

6.2 Disambiguation

Table 2 (the row marked 'LatLong' and the columns marked 'Human') indicates the inter-annotator agreement in disambiguation on ASC and ProMED. LatLongs have to agree within two meters of distance along the earth's surface, and discrepancies in units (decimal degrees versus degree-minute-seconds) are treated as errors.[16]

One source of disagreement between annotators is the granularity of the gazetteer. Gazetteers often include multiple entries for a place, with slightly

[16] In the ProMED study, which was conducted early in the project, LatLongs had to agree exactly as strings, with leading or trailing zeros treated as errors. This scoring accounts for some of the lower performance on ProMED.

different geo-coordinates depending on whether the place is viewed, say, as a town versus an administrative region. Even at a given precision, there can be a degree of arbitrariness in a gazetteer's choice of a particular geo-coordinate for a place. These problems are exacerbated in IGDB, which integrates several gazetteers; annotators differed in which entry their picked. Another source of error involves mistyping a gazetteer reference. In addition, Callisto lacks the ability to carry out inexact string matches for text mentions of places against IGDB entries, including adjectival forms of names (e.g., "Rwandan") and different transliterations (e.g., "Nisarah" vs. "Nisara"). The annotator also has to be creative in trying out various alternative ways of looking up a name ("New York, State of" vs. "New York"). There was no evidence of disagreements arising due to an annotator making use of specialized knowledge.

It is worth pointing out that the level of agreement on disambiguation depends on the size of the gazetteer. Large gazetteers increase the degree of ambiguity; for example, there are 1,420 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.

Let us turn to the MIPLACE Disambiguator. The Disambiguator is trained based on perfect extents using the disambiguated information in the training data. It is evaluated as follows: for each (perfect extent) mention M, given a gold standard gazetteer entry $G_r(M)$ in the human-annotated data for M, the disambiguator ranks the gazetteer entries in Gaz(M). The top-ranked entry $G_i(M)$ in Gaz(M) is compared against $G_r(M)$. This evaluation guarantees that $G_r(M)$, if it exists, is always ranked. It is possible to instead evaluate without such a guarantee; for example, the lookup may fail to retrieve $G_r(M)$ due to problems with transliteration, qualified names, adjectival forms, etc. However, such an evaluation, while more end-to-end, is less insightful, as it would not distinguish classifier performance from the performance of the database interface for gazetteer lookup.

The Disambiguator performance is shown in Table 2 (the row marked 'LatLong', the columns marked 'MIPLACE'). The precision and recall are discussed below in Fig. 3. The better performance of MIPLACE compared to the human is due in part to the difference in tasks: in the case of MIPLACE, the ranking of gazetteer candidates, including the correct one, from the automatic lookup in Gaz(M), versus the larger search space for the human selecting the right place, if any, in IGDB. The poorer MIPLACE disambiguation performance on ProMED compared to ASC is due to the smaller quantity of training data as well as the aforementioned errors such as text zoning and abbreviations affecting the Disambiguator.

We now discuss the impact of different thresholds on Disambiguator performance on the ASC corpus. Two "confidence" measures were computed for selecting a cutoff point between 0 and 1. For each measure, the top gazetteer 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

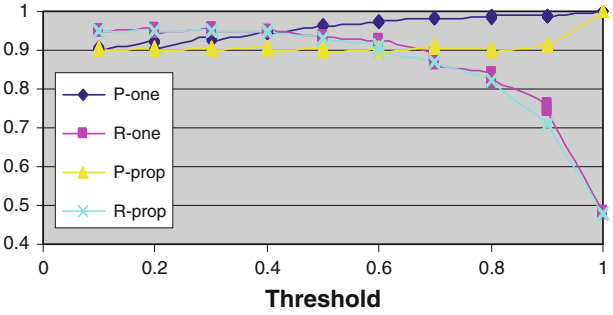

Fig. 3 Precision and recall of confidence measures on ASC

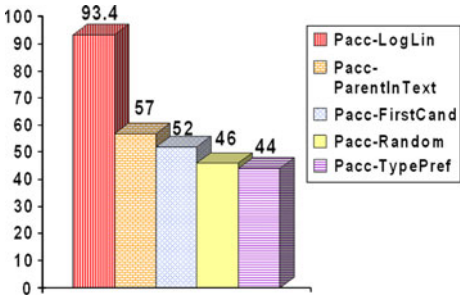

Fig. 4 Disambiguator predictive accuracy on ASC

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 Fig. 3. 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 0.9.

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 gazetteer 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.

6.3 Entity tagger across domains

When we applied the MIPLACE tool to other domains, our first observation was that results on the other corpora were lower than on ASC. We have already

Table 3 Entity tagging *F*-measure of different data aggregation methods

	ICE	ProMED
Target data only	85.60	67.54
Source data only	76.77	67.31
Data merge	85.88	84.14
Model combination	82.52	68.57
"Augment" method	85.34	71.42

mentioned some problems with MIPLACE on ProMED. Overall, 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 (i.e., measuring exact match on extents) 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.

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 fourfold 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*

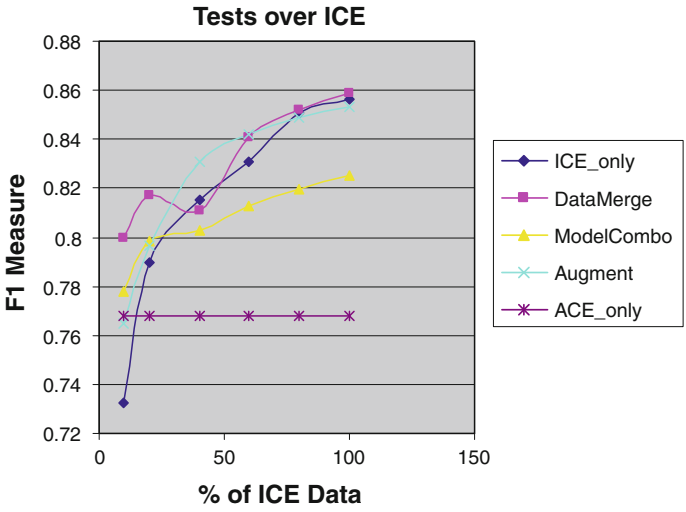

Fig. 5 Learning curves over ICE

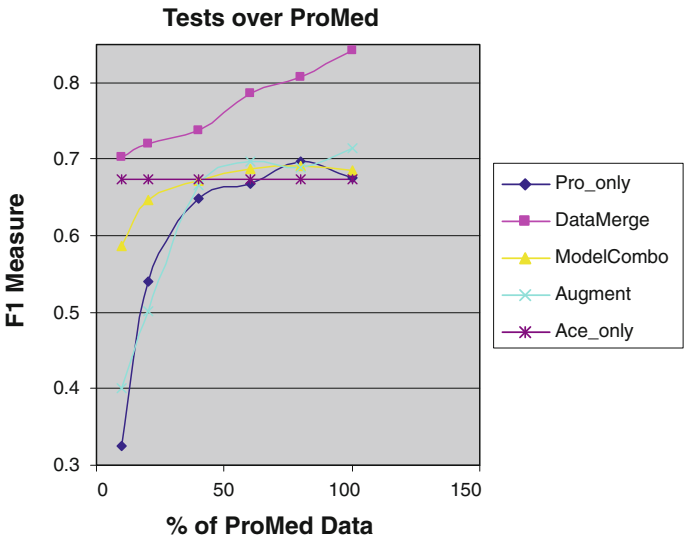

Fig. 6 Learning curves over ProMED

Merge model performs best with relatively low amounts of target data, but as increasing amounts of target data are included, the *Data Merge*, *Augment*, and target-only curves converge, implying that there is enough target data that the relatively poorly-performing source data is no longer useful.

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 is possible that additional ProMED data would lead to improvement in the other techniques' scores.

7 Conclusion

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 corpora of annotated documents with geo-coordinates, in English and Mandarin Chinese. The agreement on annotation is acceptable: inter-annotator agreement on SpatialML extents is 91.3 *F*-measure on the ASC corpus, while disambiguation agreement on geo-coordinates is 87.93 *F*-measure on it. Automatic tagging is also reasonable, though improvements are desirable in other domains. An automatic tagger for SpatialML extents scores 86.9 *F*-measure on ASC, while a disambiguator scores 93.0 *F*-measure on it. In terms of porting the extent tagger across domains, 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. When there is less target domain training data, mixing in general purpose data which is similar in content is shown to be a good strategy.

SpatialML has also gained some currency among other research groups. Pustejovsky and Moszkowicz (2008) have worked on integrating SpatialML with TimeML (Pustejovsky et al. 2005) for interpreting narratives involving travel events, using on-line sources such as travel blogs. In addition, we have collaborated with the University of Bremen in mapping SpatialML to GUM. Barker and Purves (2008) have used SpatialML in the TRIPOD image search system. SpatialML is also the inspiration for a Cross-Language Evaluation Forum (CLEF) information retrieval task aimed at search engine log analysis (Mandl et al. 2009).[17] Finally, SpatialML forms part of the initial framework for the proposed ISO-Space standard, currently a Work Item under ISO Working Group TC 37 SC4 (Language Resource Management).

Future work will extend the porting across domains to the disambiguator, and will also evaluate the system on Mandarin.[18] Our larger push is towards extending our multilingual capabilities, by bootstrapping lexical resources such as multilingual gazetteers. We also expect to do more with relative locations; currently, locations such as "a building five miles east of Fengshan" can be displayed in KML-based maps where lines are drawn between the source and target PLACEs from the RLINK. Research is underway to determine appropriate fudge factors to compute the actual orientation and length of such lines from their natural language descriptions. Finally, since we are in position to extract certain semantic relationships involving topology and orientation, we expect to enhance and then use these capabilities for formal reasoning.

Acknowledgments This research has been funded by the MITRE Innovation Program (Public Release Case Number 09-3827). We would like to thank three anonymous reviewers for their comments. We fondly and gratefully remember our late co-author Janet Hitzeman (1962–2009), without whom this work would not have been possible.

References

Barker, E., & Purves, R. (2008). A caption annotation system for georeferencing images. In *Fifth workshop on geographic information retrieval (GIR'08). ACM 17th Conference on Information and Knowledge Management, Napa, CA*, October 30, 2008.

Bateman, J. (2008). The long road from spatial language to geospatial information, and the even longer road back: the role of ontological heterogeneity. *Invited talk, LREC workshop on methodologies and resources for processing spatial language.* http://www.sfbtr8.spatial-cognition.de/SpatialLREC/.

Clementini, E., Di Felice, P., & Hernández, D. (1997). Qualitative representation of positional information. *Artificial Intelligence, 95*(2), 317–356.

Cohn, A. G., Bennett, B., Gooday, J., & Gotts, N. M. (1997). Qualitative spatial representation and reasoning with the region connection calculus. *GeoInformatica, 1*, 275–316.

Cristiani, M., & Cohn, A. G. (2002). SpaceML: A mark-up language for spatial knowledge. *Journal of Visual Languages and Computing, 13*, 97–116.

Daume III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of ACL'2007.*

Egenhofer, M., & Herring, J. (1990). Categorizing binary topological relations between regions, lines, and points in geographic databases/technical report. Department of Surveying Engineering, University of Maine, 1990.

[17] http://www.uni-hildesheim.de/logclef/LAGI_TaskGuidelines.html.

[18] On the ACE Mandarin corpus, as a baseline, the entity tagger scores 61.8 *F*-measure without the benefit of a Chinese place name list feature.

Garbin, E., & Mani, I. (2005). Disambiguating toponyms in news. In *Proceedings of the human language technology conference and conference on empirical methods in natural language processing* (pp. 363–370).

Leidner, J. L. (2006). Toponym resolution: A first large-scale comparative evaluation. Research Report EDI-INF-RR-0839.

Levinson, S. C. (2006). *Space in language and cognition: Explorations in cognitive diversity*. Cambridge: Cambridge University Press.

Mandl, T., Agosti, M., Di Nunzio, G. M., Yeh, A., Mani, I., Doran, C. et al. (2009). LogCLEF 2009: The CLEF 2009 multilingual logfile analysis track overview. *Working notes for the CLEF 2009 workshop, Corfu, Greece*. http://clef.isti.cnr.it/2009/working_notes/LogCLEF-2009-Overview-Working-Notes-2009-09-14.pdf.

Mardis, S., & Burger, J. (2005). Design for an integrated gazetteer database: Technical description and user guide for a gazetteer to support natural language processing applications. Mitre technical report, MTR 05B0000085. http://www.mitre.org/work/tech_papers/tech_papers_06/06_0375/index.html.

Papadias, D., Theodoridis, Y., Sellis, T. K., & Egenhofer, M. J. (1995). Topological relations in the world of minimum bounding rectangles: A study with R-trees. In *Proceedings of the 1995 ACM SIGMOD international conference on management of data* (pp. 92–103). San Jose, California. May 22–25, 1995.

Pustejovsky, J., Ingria, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., et al. (2005). The specification language timeML. In I. Mani, J. Pustejovsky, & R. Gaizauskas (Eds.), *The language of time: A reader* (pp. 545–557). Oxford: Oxford University Press.

Pustejovsky, J., & Moszkowicz, J. L. (2008). Integrating motion predicate classes with spatial and temporal annotations. In *Proceedings of COLING 2008: Companion volume—posters and demonstrations* (pp. 95–98).

Randell, D. A., Cui, Z., & Cohn, A. G. (1992). A spatial logic based on regions and connection. In *Proceedings of 3rd international conference on knowledge representation and reasoning*, Morgan Kaufmann, San Mateo (pp. 165–176).

Rashid, A., Shariff, B. M., Egenhofer, M. J., & Mark, D. M. (1998). Natural-language spatial relations between linear and area objects: The topology and metric of english-language terms. *International Journal of Geographic Information Science, 12*(3), 215–246.

Schilder, F., Versley, Y., & Habel, C. (2004). Extracting spatial information: Grounding, classifying and linking spatial expressions. *Workshop on geographic information*. Retrieval at the 27th ACM SIGIR conference, Sheffield, England, UK.

Sundheim, B., Mardis, S., & Burger, J. (2006). Gazetteer linkage to WordNet. In *The Third International WordNet Conference, South Jeju Island, Korea*. http://nlpweb.kaist.ac.kr/gwc/pdf2006/7.pdf.