

## **A Framework For Inferring Spatial Locations And Relationships From Text**

**Inderjeet Mani, Dave Anderson, and Janet Hitzeman**

Systems that interpret spatial information in natural language text need to deal not only with ‘absolute’ references (e.g., “*Rome*”, “*Rochester, NY*”), but also relative references (“*thirty miles north of Boston*”, “*an underpass beneath Pushkin Square*”). Current approaches to extracting information from text have made excellent progress using a methodology of first developing an annotation scheme for marking up expressions of interest with various features, and then training machine learning algorithms to reproduce the annotation. Earlier research along these lines has yielded success in resolving absolute and vague time expressions in different languages using the TIMEX2, annotation scheme ([timex2.mitre.org](http://timex2.mitre.org), Mani et al. 2005), and temporally situating text mentions of events using the TimeML annotation scheme ([www.timeml.org](http://www.timeml.org), Mani et al. 2006). We have recently begun a 3-year project to apply such a methodology, for the first time, to the automatic interpretation of spatial expressions in natural language texts in English and Chinese. Here we describe aspects of our project, building on our work to date, that are relevant to the themes of the workshop.

### ***SpatialML Markup Language***

We are currently developing a markup language for spatial expressions called SpatialML that provides a semantically-based scheme for marking up spatial expressions. It is being applied to a variety of different types of texts (including news, weather forecasts, route descriptions, geographical descriptions, etc.), with a corpus with this markup (currently already marked up with place names disambiguated with gazetteer-related features) being distributed and used as training data by various machine learning algorithms. For example, in the case of “*an underpass beneath Pushkin Square*”, “*underpass*” would be tagged in SpatialML as a feature of a particular type based on an existing feature ontology, “*beneath*” would be tagged as a *signal* with a value for a *topological relation* feature, and “*Pushkin Square*” as a particular *place* with a value for a *geo-coordinate* feature.

### ***Place Name Disambiguation***

A common way of referring to space is of course in terms of proper names. Accurate disambiguation of place names in text in terms of points and regions on a map are dependent on gazetteers with geo-coordinates and geographic inclusion information. Large gazetteers increase the degree of ambiguity; for example, there are 1420 matches for the name “*La Esperanza*”, according to the GeoNames Database from the National Geospatial-Intelligence Agency (NGA). A recent study (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 U.S. Geological Survey’s GNIS gazetteer were ‘bare’ place names that lacked any disambiguating information in the containing text sentence.

Information extraction systems can use disambiguation rules based on human intuition as well as rules discovered by programs trained from disambiguated examples. Since it is expensive to generate adequate samples of training data, research has tried to trade off

quality of training data against quantity. The Garbin and Mani study showed that nearly four out of five place names were accurately disambiguated when a machine learning program was trained on 11.7 million words of English news text that had been automatically disambiguated using hand-coded heuristics. The success of such heuristics is dependent in part on the gazetteer used; a larger gazetteer will lead to more ambiguity. MetaCarta ([www.metacarta.org](http://www.metacarta.org)), one of the well-known systems for place name tagging and disambiguation, for example, has a gazetteer of roughly 10 million entries.

### ***Gazetteer Integration***

Gazetteers are fundamental to geographical information extraction. Earlier work has explored harvesting and semi-automatic integration of multiple gazetteers to support place-name identification and disambiguation from text. We have experimented with spatial distance and geographical feature-based entries in matching entries across gazetteers. In related work, we have developed algorithms for matching transliterated variants of person names based on both sound and spelling, which are used to generate training data for machine learning approaches that learn costs of string edit operations. We will explore the impact of such name comparison approaches in terms of merging gazetteers as well as gazetteer lookup.

### ***Spatial and Temporal Reasoning***

Today's information extraction systems today reason very little, if at all, about space and time, e.g., systems cannot represent the fact that the same entity cannot be in two places at the same time. This results in extracted data that is highly incomplete, requiring considerable interpolation and extrapolation by a user. For the problem of temporally situating events, we have discovered that temporal reasoning, in the form of transitive closure over annotated qualitative temporal relations (precedence, inclusion, etc.), can be used to dramatically expand the amount of training data (Mani et al. 2006), and we expect similar benefit from spatial closure, e.g., over relations such as connection and inclusion of spatial regions. Information from domain databases will also be used to constrain information extraction results, so that in the case of "*an underpass beneath Pushkin Square*", the potential candidate underpasses can be identified and displayed on a map.

### ***References***

- Inderjeet Mani, Marc Verhagen, Ben Wellner and James Pustejovsky. (2006). Machine Learning of Temporal Relations. Proceedings of the Association for Computational Linguistics (ACL'2006), Sydney, Australia.
- Eric Garbin and Inderjeet Mani. (2005). Disambiguating Toponyms in News. In Proceedings of the Human Language Technology Conference (HLT-EMNLP'05).
- Inderjeet Mani, James Pustejovsky, and Rob Gaizauskas. (2005). The Language of Time: A Reader. Oxford University Press.