SpatialML: Annotation Scheme, Corpora and Tools

SpatialML is an SGML-like annotation language for textual data. It includes syntax for marking up PLACEs mentioned in text and for linking them to data from gazetteers and other databases. LINKs are used to express relations between places, and PATHs to capture trajectories for relative locations. A set of language-independent annotation guidelines are included in the SpatialML specification.

LINK types are listed on page 2.

Tag extents are kept small to make annotation easier and less ambiguous. For example, only Thames would be annotated out of The River Thames.

SpatialML uses ISO 3166 for country and province labels; the annotation guidelines are compatible with standard ACE guidelines. The scheme can also be integrated with GML, and there are mappings from SpatialML to KML, as well as from MetaCarta to SpatialML.

Mitre used Callisto for annotation. Annotated data is used to train an entity tagger and a disambiguator. PATH and LINK taggers are then used to complete SpatialML annotation.

The disambiguator matches PLACEs against candidate gazetteer entries, based on the mention of a place, a 10-token window around the place mention, and whether this is the first mention of such a place. Features from the gazetteer are also used. Greedy learning is applied; the first resolution of a place name is applied to all other instances of that place name in the document.

LDC provide a SpatialML corpus, based on ACE documents, known as the ACE SpatialML Corpus (ASC). Inter-annotator agreement on place tags for this corpus had an F-measure of 77; this was possibly low due to erroneous application of guidelines, and technical difficulties with annotation tools. Transliteration and different ways of writing place names led to ambiguous annotation options. Only two annotators were present; one was not a trained as a linguist annotator. ProMED enjoyed much better IAA figures, as the guidelines and tools had been improved by this time, and both annotators were trained.

The SpatialML disambiguator had a high f-measure of 93. Previous study showed that larger gazetteers increase the degree of ambiguity. The current disambiguator worked well enough to automate pre-processing before human annotation. Two confidence measures are used for disambiguation; one shows how far the most-likely guess stands out, another compensates for the number of potential choices. A variety of baselines were chosen for disambiguation performance, the best of which reached only 57% accuracy compared to this system’s 93.4% accuracy.

Cross domain tagging is also investigated; performance dropped sharply. Document zoning of header/title/signature blocks contributed to this, as did the tagging of place names inside proper noun phrases (like West Nile Virus). Training the entity tagger on a mixed-domain corpus gave the best end results.

Future work will involve improving cross domain tagger accuracy, and evaluation of LINK and PATH taggers, as well as further IAA studies, and integration with TimeML.