ATLIS: Identifying Locational Information in Text Automatically

John Vogel, Marc Verhagen, James Pustejovsky
Computer Science Department, Brandeis University
Waltham, Massachusetts, USA
E-mail: jvogel@cs.brandeis.edu, marc@cs.brandeis.edu, jamesp@cs.brandeis.edu

Abstract

ATLIS (short for " ATLIS Tags Locations in Strings") is a tool being developed using a maximum-entropy machine learning model for automatically identifying information relating to spatial and locational information in natural language text. It is being developed in parallel with the ISO-Space standard for annotation of spatial information (Pustejovsky, Moszkowicz & Verhagen 2011). The goal of ATLIS is to be able to take in a document as raw text and mark it up with ISO-Space annotation data, so that another program could use the information in a standardized format to reason about the semantics of the spatial information in the document. The tool (as well as ISO-Space itself) is still in the early stages of development. At present it implements a subset of the proposed ISO-Space annotation standard: it identifies expressions that refer to specific places, as well as identifying prepositional constructions that indicate a spatial relationship between two objects. In this paper, the structure of the ATLIS tool is presented, along with preliminary evaluations of its performance.

Keywords: ISO-Space, location tagging, spatial processing

1. Motivation

There is a process underway to build a standard for marking up information in text related to space and locational propositions – the ISO-Space standard (Pustejovsky, Moszkowicz & Verhagen 2011). This standard is proposed to take sentences like "A Chinese merchant ship has run aground off the coast of Australia and began leaking oil into the ocean near the Great Barrier Reef"[1] and provide annotations with the spatial relationships among the Chinese merchant ship, the coast of Australia, the oil, and the Great Barrier Reef. For example, the oil is in the process of moving from the merchant ship to the ocean; the merchant ship is near both Australia and the Great Barrier Reef; and the merchant ship is contained by the ocean.

ATLIS is being developed as a tool for automatically extracting this locational information in order to pass it on to other programs that could work with this standardized data for any application that needed to construct a model of the spatial relationships in a document, such as a story understanding system (Mueller 2003). It will provide an alternative to the MITRE SpatialML location tagger MIPLACE (Mani et. al., 2010), using the ISO-Space annotation system instead of MITRE's SpatialML annotation system, thereby eventually providing a richer annotation.

2. Functionality

ATLIS's design goal is to automatically tag all the spatial information in a piece of text according to the ISO-Space standard. ATLIS currently implements a subset of this design goal. At present, it identifies expressions in text that contain locational information, tagging their extent so that a second pass (to be developed) can analyze the nature of the locations and their relations to one another and tag the expressions for other important spatial information, such as absolute coordinates, relative locations, orientations, and topological relationships. ATLIS currently focuses on two types of expressions

[1] Example taken from Wikinews at http://en.wikinews.org/wiki/Chinese_ship_leaking_oil_near_Great_Barrier_Reef

containing locational information: standalone locations and prepositional locational expressions. Systems for identifying other locationally informative expressions, such as locational verbs, will be added later. The systems for identifying standalone locations and for identifying prepositional locational expressions stand apart from each other.

2.1 Standalone Locations

"Standalone locations" are simple expressions that identify specific locations in space. For example, "Boston", "the plaza", and "the Great Barrier Reef" are all standalone locations. Compositional expressions like "the salt flats in the middle of the desert" would be considered prepositional locational expressions instead.

ATLIS uses two passes to tag standalone locations. The first pass identifies standalone locations represented by multiple words, and the second pass identifies locations represented by a single word that are not already contained within a multi-word location expresssion.

2.1.1 Multiword Expressions

Multiword expressions of standalone location are essentially always proper nouns naming particular locations (like "Great Barrier Reef"). To identify them in a given text, ATLIS looks up sequences of words in the GeoNames database of geographical placenames[2] . Sequences that match entries in the GeoNames database are then filtered to reduce the incidence of false positives. The system filters three types of multiword expression out. The first is expressions whose first word is a common first name, to eliminate false positives based on locations named after people. For example, "George Washington" is the name of a famous U.S. historical person, but it appears in the GeoNames database because it is also the name of a U.S. University. The second filter filters out two-word expressions whose first word is an article. This reduces false positives because a place name with only one content word in it is more likely to also be used in a non-locational sense. If a true positive ends up filtered out in this stage it can still be caught in the second stage for single word expressions, only without its article. The third filter filters out multi-word expressions with line breaks between the words. This filters out collocations that only appear because of an accident of the document's structure.

2.1.2 Single-word Expressions

To identify single-word locational expressions in a text, ATLIS uses a maximum-entropy classifier on each word in the text. Before running the classifier over the text, the text is tagged using the open-source HunPOS tagger (Halácsy, Kornai & Oravecz, 2007). This allows ATLIS to skip evaluation of any words that are of the wrong part of speech to represent locations, and also provides features for the classifier.

The feature set the classifier uses can be seen in Table 1 below.

Feature Name	Feature Value
PreviousWord, Word, NextWord, PreviousWordBigram	The word and its two neighbors, and the previous bigram of words.
PreviousPOS, POS, NextPOS, PreviousPOSBigram, NextPOSBigram	The POS tags of the word and its two neighbors and the neighboring POS bigrams.
WordnetLoc	True if the word falls under "location" in Wordnet (see below).
IsName, PreviousIsName	True if the word (or the previous word) is a common first name.
LastThreeLetters	The last three letters of the word.
ManyPeople	True if the word has an entry in GeoNames with a listed population of at least 45,000 (see below)
FirstIsDigit	True if the first character in the word is a digit.
LongerThan5	True if the word is longer than 5 characters.

Table 1: Single-word Classifier Features

The "WordnetLoc" feature in the table above looks at the location of the word in the WordNet lexical database (Miller 1995). If it finds the word in WordNet, it checks to see if the most frequent sense of the word is a hyponym of "location" or "continent". If it is, the value

[2] Accessible at www.geonames.org.

of the feature is True, other wise it is False.

The ManyPeople feature in the table brings in data from the GeoNames database, but unlike in the case of multi-word locational expressions, locations with a population of less than 45,000 are ignored. Population is here being used as a proxy for the likelihood that the geographic sense is the actual sense that is meant by a particular usage of the word.

Below is a diagram illustrating the structure of the standalone location tagging system.

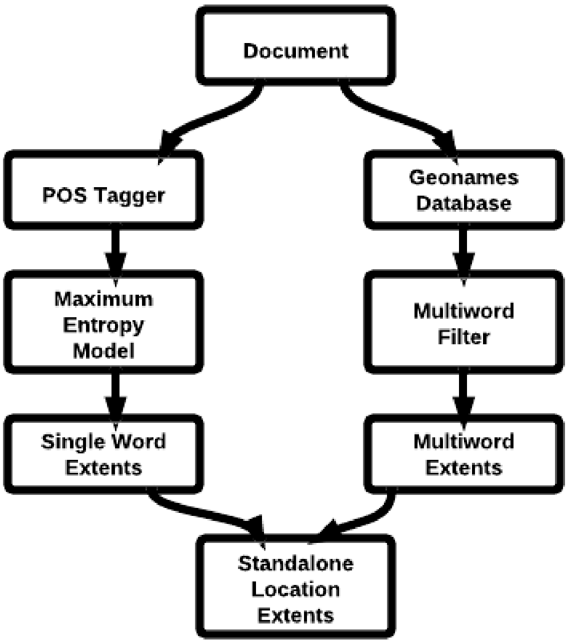

Figure 1: ATLIS Standalone Location Tagger

2.2 Prepositional Locational Expressions

Not all elements in a sentence that carry locational information are inherently locational *per se*. For example, in the sentence "'Shut the door behind you', he whispered, as he placed the lamp on the table"[3] , "the table" is being used as a location in space, and we want to capture the spatial relationship between the lamp and the table, even though "the table" does not really name a place. In this case, "the table" is being coerced to act as a location because it is in a prepositional locational expression. A prepositional locational expression consists of two expressions, which may or may not be standalone locations themselves, syntactically connected by a preposition which establishes a relationship in space between them.

To tag prepositional location expressions, ATLIS generates a dependency parse of the system using the MaltParser dependency parser (Nivre et. al., 2007). It identifies possible candidates for prepositions being used to generate a spatial relationship by picking out those prepositions in the parse that link two nouns or a noun and a verb. It then uses a maximum entropy machine learning model, as in the single-word standalone expression classifier, to classify the candidate preposition/ expression/ expression triples as locational or non-location in nature.

The feature set the classifier uses can be seen in Table 2 below.

Feature Name	**Feature Value**
PreviousWord, Word, NextWord	The preposition and the two words it connects in the dependency parse.
PreviousPOS, NextPOS	The POS tags of the words the preposition connects (always some sort of noun or verb).
NextIsAbstract	True if the second connected word is a hyponym of "abstraction" in WordNet.
PreviousIsLoc, NextIsLoc	True if the connected words are hyponyms of "location" as described above.
LocInSentence	True if any word in the sentence is a hyponym of "location" as described above.
PrevIsVision	True if the first connected word is a hyponym of one of a list of vision verbs (see below)

Table 2: Preposition Classifier Features

The verbs that are considered to be the basic "vision" verbs for the purposes of ATLIS's features are "look",

[3] Example taken from *The Picture of Dorian Gray* by Oscar Wilde.

"stare", and "watch". Any verb whose most common sense is a hyponym of those verbs will be classified as a vision verb.

Below is a diagram illustrating the structure of the standalone location tagging system.

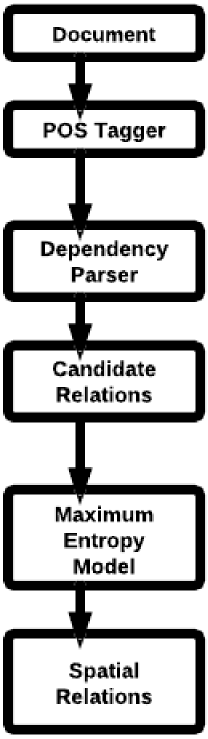

Figure 2: ATLIS Preposition Tagger

3. Example of Use

Below is an example of using ATLIS's standalone location recognizer on a passage from a randomly-selected Google News document.[4] Strings identified as locations are marked in **bold**.

"A mayor-elect from the Mexican state of **Veracruz** was kidnapped and killed along with two companions on Monday, local media reported.

The reports did not link the killings to the violence sweeping Mexico as the government fights powerful drug cartels. Several mayors and other elected officials have been targeted by drug gangs in recent months.

Gregorio Barradas Miravete, the mayor-elect from the municipality of Juan Rodriguez Clara, was a member of President Felipe Calderon's National Action Party.

The three men were forced into a Hummer truck in the afternoon in the south of **Veracruz** and were then taken to the neighboring state of **Oaxaca**, local prosecutors said, according to several online newspapers."

ATLIS correctly identified "Veracruz" and "Oaxaca" in the document, but missed "Mexico" and "Juan Rodriguez Clara".

4. Evaluation

Currently, training and evaluation of ATLIS's recognizers is difficult due to the lack of a specially-built corpus with ISO-Space annotations for locational expressions. Such corpora are under development. In the meantime ATLIS is using corpora built for other purposes to stand in for a proper ISO-Space corpus.

To evaluate the standalone location tagger, the MITRE SpatialML corpus is being used, with the extents tagged as PLACE being used as the extents of standalone locations (Mani et. al., 2010). Five-fold cross-validation was used to separate training and test data. When evaluated on the data, ATLIS's standalone location tagger achieved a precision of 0.938 with a recall of 0.688. This is a substantial improvement over the baseline method of tagging as a standalone location anything which appears in the GeoNames database, which achieves a precision of 0.176 with a recall of 0.467. It is worse than, but comparable to, the reported performance of the MITRE MI-PLACE tagger on the data, which achieved a precision of 0.973 with a recall of 0.785. However, it is hoped that when a full ISO-Space corpus becomes available ATLIS's performance will improve as it targets that corpus.

To evaluate the prepositional location tagger, an artificial corpus was created from the corpus used for the 2007 SemEval preposition disambiguation shared task (Litkowski & Hargraves, 2007). Each preposition sense in the SemEval annotation was labeled as being either a spatial sense or a non-spatial sense, and the prepositions that were labeled in the SemEval corpus with a spatial sense were then considered to participate in a prepositional locational expression. This created a

[4] "Incoming small town Mexico mayor kidnapped, killed". 2010. Reuters.

derivative corpus, which is noisy but usable for development purposes. When evaluated on this noisy corpus, ATLIS distinguishes spatial prepositions from non-spatial ones with an accuracy of 0.736.

5. Conclusions and Future Work

ATLIS is still in an early stage of its development. Much of the future work depends on the development of a corpus specifically annotated to the ISO-Space standard. Once a corpus that is directly targeted at this ATLIS's task becomes available, more specifically-targeted development can be done which is likely to improve accuracy. Also, this will enable ATLIS's tagging to contain more information; rather than simply determining that there is a locational expression at a given extent, or determining spatial link between two expressions, ATLIS should be able to determine the nature of that location or link in terms of ISO-Space information like relative position, orientation, direction of movement, and so on.

6. Acknowledgements

Part of this research was funded under the NGA-NURI grant "Spatiotemporal Tracking of Entities: Determining Object Location from Image and Text Data", award number HM1582-08-1-0018.

7. References

Pustejovsky, J., Moszkowicz J., Verhagen M. (2011). ISO-Space: The Annotation of Spatial Information in Language. In *Proceedings of the Sixth Joint ISO ACL SIGSEM Workshop on Interoperable Semantic Annotation.* pp. 1-9.

Mueller, E. (2003). Story Understanding Through Multi-Representation Model Construction. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning.* Vol. 9, pp. 46-53

Mani I., Doran C., Harris D., Hitzeman J., Wellner B., Mardis S., Clancy S. (2010). SpatialML: Annotation Scheme, Resources, and Evaluation. In *Language Resources and Evaluation,* Vol. 44, No. 3, pp. 263-280.

Halácsy P., Kornai A. Oravecz C.(2007). HunPos – an Open Source Trigram Tagger. In *Computational Linguistics* (June), pp. 209-212.

Miller G.A. (1995). WordNet: A Lexical Database for English. In *Communications of the ACM,* Vol. 38, No.11, pp. 39-41.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E. (2007). MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. In *Natural Language Engineering,* Vol. 13, No. 2, pp. 95-135

Litkowski, K. C., Hargraves, O. (2007). SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions. In *Proceedings of the Fourth International Workshop on Semantic Evaluations SemEval2007,* p. 24-29