Resources for Place Name Analysis

Robert Irie[†], Beth Sundheim[‡]

SPAWAR Systems Center San Diego
53560 Hull Street
San Diego, CA 92152
[†]irier@spawar.navy.mil, [‡]beth.sundheim@navy.mil

Abstract

We present a new resource for annotating and visualizing the meaning of place names in natural language text, along with insights gained from analysis of manual annotations. The work addresses the issue of place name (toponym) meaning resolution, moving beyond simple named entity recognition to address the problem of grounding textual references, i.e., making a connection between the references and the real-world entities that they denote. The name, "San Francisco," for example, can be mapped to more than 900 distinct place entities that differ in terms of location and/or type, according to commonly available databases of named geographical entities, called gazetteers. Gazetteers serve as knowledge bases that can be exploited to support the analysis and disambiguation of named places, and the grounding of textual references in real-world entities. This process of grounding text in gazetteers offers a way of normalizing the meaning of the place name references that are found in the gazetteers, and essentially subsumes the text analysis process of determining when instances of a given name are being used to refer to the same or to different entities.

Introduction

There are large, publicly available databases of information about named places that contain useful information for natural language processing (NLP) about alternative names and/or spellings, place type (city, mountain, park, etc.) and the broader area that contains the place. Usually, they also give the precise physical location or centerpoint of the place in terms of map coordinates, which may be useful as a knowledge source when doing sophisticated inferencing about spatial relationships expressed in text, but is primarily important for applications such as plotting places of interest in a text on a map display. There is research and applications work ongoing at various sites that involve the use of gazetteer databases as knowledge resources; some of these efforts are reflected in papers presented at a recent workshop (Kornai and Sundheim, 2003). There is also a U.S. government-sponsored program on information extraction that has focused much attention on the problem of aggregating textual references (not just names, but also referential descriptions and pronouns) into entity-level representations, and normalizing aspects of their meaning (see http://www.nist.gov/speech/tests/ace/). Our work explores the potential usage of gazetteers for NLP purposes in some detail and addresses issues associated with normalizing the meaning of place name references using the gazetteers.

There are two phases in our planned work in place name reference (PNR) analysis. Phase I, which we present here, involves the manual annotation of PNRs in natural language text, and quantitative and qualitative analysis of the results. In Phase II we expect to invest a significant portion of our effort on incorporating standard machine learning techniques to perform automatic annotation.

The Phase I goals are to develop software and procedural resources to aid in the manual annotation task, and to analyze annotations to determine the nature and extent of the PNR ambiguity problem.

Resource Architecture

We required an easy-to-use tool that could quickly detect and index PNRs in corpora, visualize the possible candidates for resolving an instance of a PNR to a particular entity, and enable humans to create ground truth annotated data. We examined several freely available NLP and general purpose annotation tools, including GATE (Cunningham et al, 2002) and Amaya (Quint & Vatton, 1997), and concluded that none offered all of the above features in a convenient package. Therefore, we set out to develop a resource specifically designed for place name analysis and annotation.

The resource, the Testbed for Automatic Place-name Interpretation and Resolution (TAPIR), is a dynamic, modular web application that is primarily designed to aid in the manual annotation of PNRs. It organizes documents into corpora indices, runs a named entity tagger (Bikel et al., 1997) to identify PNRs, looks up geographical information in an integrated geospatial database, and presents a consistent and streamlined interface for manual annotation (reference resolution).

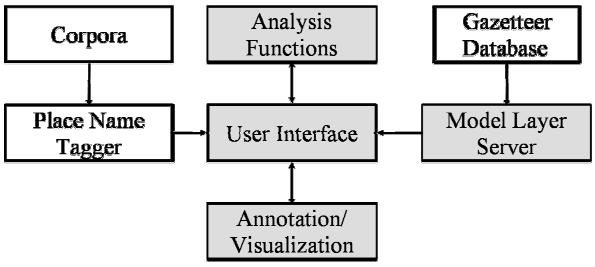

Figure 1: Block diagram of the TAPIR architecture. Shaded boxes correspond to in-house developed resources. The Model Layer Server interfaces to the gazetteer data using the standard web protocol XML-RPC.

Aside from the proprietary named entity tagger, all TAPIR components (Figure 1) are freely available or were developed internally, using standard web protocols (HTTP, XML-RPC). The two major components of TAPIR are the user interface and the integrated gazetteer database.

User Interface

The TAPIR user interface is based on an open source web application server called Zope (2004). The human annotator initiates TAPIR in a start-up window by selecting an indexed corpus to run. Thereafter, the annotator works exclusively within a single browser window, which is composed of five panels (Figure 2): an alphabetized list of identified place names, a keyword-in-context concordance of text segments containing a selected PNR instance, the document context for a selected instance, the results of a name lookup in the integrated gazetteer database, and the annotation data that corresponds to a particular instance.

The visualization interface (Figure 3) consists of a graphical latitude/longitude plot of a specific gazetteer entry, using coordinate data from the gazetteer entry attributes. The display is dynamically updated as the user moves the mouse over the gazetteer query results panel. In the case where there are several candidate entries in an unfamiliar region of the world, it establishes a quick and simple geographic context to aid the annotator in selecting the best matching gazetteer entry, especially when the document context provides some location clues about the particular place that is intended.

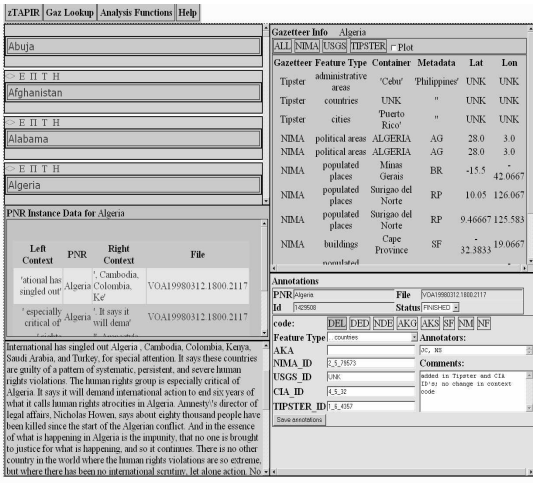

Figure 2: Screenshot of the main user interface of TAPIR, which is a web page composed of panels. Displayed are place names found in a corpus, instances of these names, sentence-level and document-level context, gazetteer entries, and annotation data.

Integrated Gazetteer Database

The integrated geospatial database was created from four publicly available gazetteers: the NIMA (National Imagery and Mapping Agency)[1] gazetteer (NIMA, 2003) for place names outside the United States, the USGS (United States Geological Survey) gazetteer (USGS, 2003) for those within the U.S, and general interest gazetteers from the CIA World Factbook (CIA, 2003) and TIPSTER (TIPSTER, 1992), a U.S. government-sponsored text research program. The NIMA and USGS gazetteers are relatively large databases, with 5.3 and 2.0 million entries, respectively, while the TIPSTER and CIA World Factbook gazetteers are much smaller (240,000 and 1,600, respectively). Most of these gazetteers are surveyed in (Sundheim, 2002).

A key step in the construction of the integrated gazetteer database was to define the mapping among the disparate entity type categorization schemes of each individual gazetteer. Rather than cross-referencing categories directly between each pairing of gazetteers, we mapped all of the categories to a uniform scheme developed by the University of California, Santa Barbara in project work for the Alexandria Digital Library (ADL) research program (Hill, 2000). The ADL scheme is a partially hierarchical thesaurus of over 200 "feature types" (categories of places).

After the gazetteers were loaded into the integrated database, duplicate and erroneous entries in the raw data were removed, and a normalized database schema was developed to allow flexible queries based on place name, location, feature type, etc. A dynamic update tool was implemented to allow changes to the component gazetteers to be incorporated into the integrated gazetteer.

An XML-RPC server forms the external interface to the integrated gazetteer database, allowing multiple distributed resources, including the TAPIR user interface, to access the database.

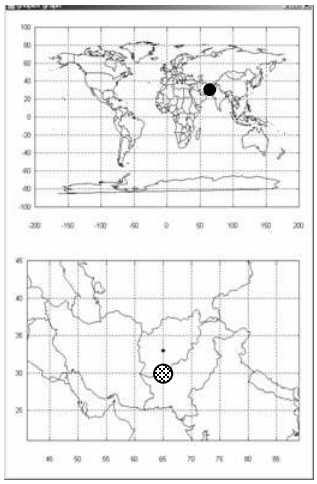

Figure 3: Screenshot of the visualization interface. Displayed is the geographic location of a particular entry that has been selected in the gazetteer query results panel of the TAPIR main interface. The plot is updated each time a new entry is examined. The plot points have been enlarged in the figure for better visibility.

Annotation

Over the course of the manual annotation effort in Phase I, instances of PNRs detected by the name tagger in 4,120 English documents were annotated. The text genre was news, including articles from multiple newswire services, transcriptions of broadcast news stories from multiple sources, news stories originating in various languages and media that were disseminated by the U.S. Foreign Broadcast Information Service, and stories in English from the Xinhua Chinese press. The corpora also included sets of domain-specific stories on contract killings and on

[1] The organization's name has recently changed to National Geospatial Intelligence Agency (NGA).

illicit material trafficking. The corpora were obtained from the Linguistic Data Consortium and from in-house projects conducted for the Defense Advanced Research Projects Agency and the Advanced Research and Development Activity. Several of the corpora emphasize a particular region of the world: Latin America, Russia, China, Eastern Europe, or the United States.

Four part-time individuals were involved in Phase I annotation. For each instance of each PNR in each of the indexed corpora, annotators selected zero or more entries from the gazetteer panel in the TAPIR user interface that were suitable matches. Match candidates included only those gazetteer entries whose name spelling matched that of the name as it appeared in the corpus. In total, they annotated 18,900 mentions of PNRs (over 3,000 unique names).

To ensure a level of consistency across the annotators, a set of guidelines was developed, providing general instructions on how to determine the correct match, and specific help on cases where there is an incomplete match between the PNR in the text and an entry in the gazetteers, in terms of feature type and/or location information.

The two decisions faced by the annotators that relate most directly to the expected output of NLP systems are:

1. "link-or-no-link": Is there at least one gazetteer entry that is a real candidate? 673 instances (out of 18,900) were judged not to have a match in the gazetteer, despite matching spelling of the name.

2. "which-link": Which candidate entry is the best match? There were just 27 instances (representing 20 distinct names) out of the 18,900 total instances, when the annotator was left with a number of candidate entries (at least six) after analysis and had insufficient world knowledge to be able to rank the likelihood of any of them. These will be difficult cases for NLP systems as well, but represent a very small portion of the data. When there were less than six good candidates, annotators were generally able to single one of them out as "first choice" and list the others as less likely alternatives. This situation was encountered in about 20% of all cases.

A careful study of interannotator agreement has been conducted using two well-trained annotators, a refined written set of annotation guidelines, and software to do the scoring.

1. Agreement on "link-or-no-link" decision: 95.3% on the F-measure.

2. Agreement on "which-link" decision: 87.0%-99.0% F, depending on which of the four gazetteers is used in the measurement.

Analysis

Once the manual annotation process was completed, we analyzed the annotations for any relevant information concerning the gazetteers, the feature type scheme, and the overall ambiguity resolution process.

Gazetteer Coverage

With all gazetteers taken together, 88.6% of all document mentions of PNRs have a corresponding gazetteer entry, 7.8% of the mentions are not found in the gazetteers at all, and 3.6% are found with only an incorrect sense. This means that even though the documents that were annotated are generally heavy on names of familiar political places such as countries and capital cities, and light on other sorts of place names, one can expect to be unable to link a PNR instance to a gazetteer entry over 10% of the time.

The importance of using multiple gazetteers with varying geographic scope and density of coverage was underscored by the fact that for some instances (approximately 7%), only the smaller gazetteers (TIPSTER and CIA) contained the best match.

During the annotation process, we identified several cases where an exact string lookup against our gazetteer database narrowly missed finding the correct entry. We therefore designed several preprocessing heuristics to improve the likelihood of finding appropriate matches. The integration of the heuristics was done after the manual annotation was completed, and involved implementing lookup tables and simple textual transformations. For example, postal codes of U.S. states were not represented in the gazetteers; we therefore implemented heuristics to replace codes such as ‘CA’ with their full name equivalents, like 'California.' Of the 3,636 PNR instances that were not initially matched to any gazetteer entry, 3,272 (90%) were resolved using the preprocessing heuristics.

Feature Type Correspondence

The annotations provide information on the closeness of fit between a place's type as revealed in a document and its type as given in a gazetteer. Almost 90% of the annotations show a difference between the two, and often annotators disagreed with each other as to the type that corresponded best to the document context. However, the differences between annotator decisions were generally semantically fine-grained, e.g., a difference between *capital* and *city*, and it did not prevent annotators from agreeing on the two basic decisions of "link-or-no-link" and "which-link". Similarly, the manually assigned feature type often differed from the feature type recorded in the closest-matching gazetteer entry only in terms of specificity, due to an inexact correspondence between categorization schemes, including differences in granularity between the schemes.

Ambiguity Resolution

Manual annotations provide some evidence to show that relatively simple NLP techniques may be powerful enough to make correct decisions on the "which-link" question in a large number of cases. The evidence for this from the manual annotation effort is that 65% of all links between a document mention and a gazetteer entry were determined on the basis of location evidence found within a narrow window of three sentences in the document, comprising the sentence containing the name, the sentence before, and the sentence after (Figure 4).

For example, a document might say, "Fighting erupted today in several cities in central Iraq. American personnel in Baghdad suffered a number of casualties." The NIMA gazetteer lists "Baghdad" as a populated place in five different Middle Eastern countries. The first sentence provides enough location evidence for "Baghdad" in the second sentence to permit the candidate gazetteer entries for "Baghdad" to be reduced to just one choice.

Another conclusion from the annotation analysis derives from the fact that the *potential* ambiguity of PNRs in the examined corpora was high, but the *actual* ambiguity was very low. By *potential* ambiguity, we mean the degree to which PNRs are ambiguous according to the gazetteers, and by *actual* ambiguity, we mean the degree to which PNRs are actually used in more than one sense within a corpus. For USGS/NIMA, the potential ambiguity averages 33, i.e., on average there are 33 different interpretations for a PNR in the combination of those two gazetteers. But the actual ambiguity of the names that appear in our annotated corpora is extremely low; almost all names are used in just one sense throughout the corpus, and those that are used in more than one sense are used on average in just two senses. Moreover, for common place names (such as country capitals, etc.), regardless of the degree of ambiguity, there was usually a unique "default" sense that took precedence. This suggests that a combination of heuristics and statistical algorithms can be developed to automatically disambiguate PNRs.

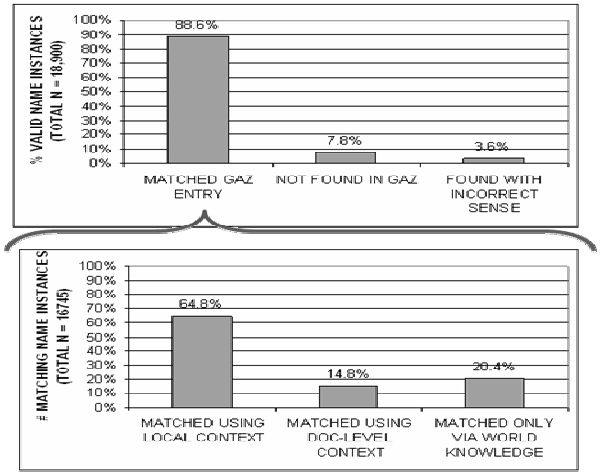

Figure 4: Matching valid name instances with gazetteer entries. A high percentage of name instances were matched to a gazetteer entry. Of the matching ones, most were matched using evidence only from local context.

Conclusion

We present a set of resources that aid in the analysis of PNRs, and some insight into the prospects for PNR processing in natural language systems. The software resources that have been developed include a modular web framework to manually (and eventually automatically) annotate previously-detected PNRs and an integrated geospatial database of place name information. Each component supports standard web protocols, allowing modularity and interoperability with other natural language systems. We have also extensively analyzed the manual annotations to gain a better understanding of the nature of the PNR ambiguity problem. Finally, a set of annotation guidelines has been established to promote uniformity across annotators. The annotations, in XML format, and the guidelines are being prepared for possible release to interested parties.

We believe we have sufficient software resources and annotated data to begin developing automatic PNR resolution algorithms. We are currently developing a combination of simple pruning heuristics (to reduce the gazetteer match search space) and clustering algorithms (to group together spatially and topically related geographic entities), using the annotations as ground truth data for training and validation.

Acknowledgements

We gratefully acknowledge the support of the Advanced Research and Development Activity (ARDA) program in Advanced Question Answering for Intelligence (AQUAINT). In addition, part of the research and development for this project was funded by an in-house independent research program overseen by the Office of Naval Research (ONR).

References

Bikel, D., Miller, S., Schwartz, R., and Weischedel, R. (1997). NYMBLE: A High-Performance Learning Name-finder. In Proceedings of the Fifth Conference on Applied Natural Language Processing (pp. 194-201). Association for Computational Linguistics.

CIA (2003). The World Factbook Appendix F: Cross-Reference List of Geographic Names. http://www.cia.gov/cia/publications/factbook/appendix/appendix-f.html.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. Philadelphia, PA.

Hill, L. (2000). Core elements of digital gazetteers: placenames, categories, and footprints. In J. Borbinha & T. Baker (Eds.), Research and Advanced Technology for Digital Libraries: Proceedings of the 4th European Conference, ECDL 2000 (pp. 280-290). Berlin: Springer.

Kornai, A. and Sundheim, B. (2003). Analysis of Geographic References (workshop proceedings), HLT-NAACL 2003, Edmonton, Canada, May 31, 2003. Association for Computational Linguistics.

NIMA (2003). GEOnet Names Server. http://earth-info.nima.mil/gns/html/index.html.

Quint, V. and Vatton, I. (1997). An Introduction to Amaya. W3C NOTE 20-February-1997. http://www.w3.org/TR/NOTE-amaya: World Wide Web Consortium.

Sundheim, B. (2002). Resources to Facilitate Progress in Place Name Identification and Reference Resolution. In Proceedings of the Second International Conference on Human Language Technology Research, San Diego, CA, March 2002 (pp. 319-324). San Francisco, CA: Morgan Kaufmann.

TIPSTER (1992). Gazetteer 4.0. ftp://crl.nmsu.edu/CLR/lexica/gazetteer.

USGS (2003). Geographic Names Information System. http://geonames.usgs.gov.

Zope (2004). Zope Application Server. http://www.zope.org.