Reuse of a Proper Noun Recognition System in Commercial and Operational NLP Applications

Chinatsu Aone and **John Maloney**
SRA International
4300 Fair Lakes Court
Fairfax, VA 22033
aonec@sra.com, maloneyj@sra.com

Abstract

SRA's proprietary product, NameTag^{TM}, which provides fast and accurate name recognition, has been reused in many applications in recent and ongoing efforts, including multilingual information retrieval and browsing, text clustering, and assistance to manual text indexing. This paper reports on SRA's experience in embedding name recognition in these three specific applications, and the mutual impacts that occur, both on the algorithmic level and in the role that name recognition plays in user interaction with a system. In the course of this, we touch upon various interactions between proper name recognition and machine translation (MT), as well as the role of accurate name recognition in improving the performance of word segmentation algorithms needed for languages whose writing systems do not segment words.

1 Introduction

Fast and accurate name recognition products are only now coming onto the market. SRA's proprietary product, NameTag, has been reused in many applications in recent and ongoing efforts, including multilingual information retrieval and browsing, text clustering, and assistance to manual text indexing. In the following paper, we report on our experience in embedding name recognition in these three specific applications, as well as the mutual impacts that occur, both on the algorithmic level and in the role that name recognition plays in user interaction with a system. In the course of this, we touch upon various interactions between proper name recognition and machine translation (MT), as well as the role of accurate name recognition in improving the performance of word segmentation algorithms needed for languages such as Japanese. Name recognition clearly offers added value when integrated with other algorithms and systems, but the latter also affect the way in which name recognition is performed, specifically the choice of high-recall or high-precision strategies. But first, we discuss the relevant features of NameTag.

2 Description of NameTag

NameTag is a multilingual name recognition system. It finds and disambiguates in texts the names of people, organizations, and places, as well as time and numeric expressions with very high accuracy. The design of the system makes possible the dynamic recognition of names: NameTag does not rely on long lists of known names. Instead, NameTag makes use of a flexible pattern specification language to identify novel names that have not been encountered previously. In addition, NameTag can recognize and link variants of names in the same document automatically. For instance, it can link "IBM" to "International Business Machines" and "President Clinton" to "Bill Clinton."

NameTag incorporates a language-independent C++ pattern-matching engine along with the language-specific lexicons, patterns, and other resources necessary for each language. In addition, the Japanese, Chinese, and Thai versions integrate word segmenters to deal with the orthographic challenges of these languages. (NameTag currently has these language versions available plus ones for English, Spanish, and French.)

NameTag is an extremely fast and robust system that can be easily integrated with other applications through its API. It has been our experience that NameTag has lent itself to so many successful integrations in diverse applications not just due to its accuracy, but to its speed. (Its NT version is currently benchmarked at 300 megabytes/hour on a Pentium Pro.) It is an attractive package to embed in an application, as it does not cause significant retardation of performance.

In the following discussion, we refer to various versions of NameTag, most prominently systems for English and Japanese. Their extraction accuracy varies. For example, in the Sixth Message Understanding Conference (MUC-6), the English sys-

tem was benchmarked against the Wall Street Journal blind test set for the name tagging task, and achieved a 96% F-measure, which is a combination of recall and precision measures. Our internal testing of the Japanese system against blind test sets of various Japanese newspaper articles indicates that it achieves from high-80 to low-90% accuracy, depending on the types of corpora. Indexing names in Japanese texts is usually more challenging than English for two main reasons. First, there is no case distinction in Japanese, whereas English names in newspapers are capitalized, and capitalization is a very strong clue for English name tagging. Second, Japanese words are not separated by spaces and therefore must be segmented into separate words before the name tagging process. As segmentation is not 100% accurate, segmentation errors can sometimes can use name tagging rules not to fire or to misfire.

3 Proper Name Recognition Integrated With a Browsing & Retrieval System

We have recently developed a system incorporating NameTag that allows monolingual users to access information on the World Wide Web in languages that they do not know (Aone, Charocopos, and Gorlinsky, 1997). For example, previously it was not easy for a monolingual English speaker to locate necessary information written in Japanese. The user would not know the query terms in Japanese even if the search engine accepted Japanese queries. In addition, even when the users located a possibly relevant text in Japanese, they would have little idea about what was in the text. Output of off-the-shelf machine translation (MT) systems are often of low quality, and even "high-end" MT systems have problems particularly in translating proper names and specialized domain terms, which often contain the most critical information to the users.

Now these users have available our multilingual (or cross-linguistic) information browsing and retrieval system, which is aimed at monolingual users who are interested in information from multiple language sources. The system takes advantage of name-recognition software as embodied in NameTag to improve the accuracy of cross-linguistic retrieval and to provide innovative methods to browse and explore multilingual document collections. The system indexes texts in different languages (currently English and Japanese) and allows the users to retrieve relevant texts in their native language (currently English). The retrieved text is then presented to the users with proper names and specialized domain terms translated and hyperlinked. Among the innovations in our system is the stress placed upon proper names and their role as indices for document content.

The system consists of an Indexing Module, a Client Module, and a Term Translation Module. The Indexing Module creates and inserts indices into a database while the Client Module allows browsing and retrieval of information in the database through a Web-browser-based graphical user interface (GUI). The Term Translation Module dynamically translates English user queries into Japanese and the indexed terms in retrieved Japanese documents into English.

The Indexing Module

For the present application, the system indexes names of people, entities, and locations, as well as scientific and technical (S&T) terms in both English and Japanese texts, and allows the user to query and browse the indexed database in English. As NameTag processes texts, the indexed terms are stored in a relational database with their semantic type information (person, entity, place, S&T term) and alias information along with such meta data as source, date, language, and frequency information.

The Client Module

The Client Module lets the user both retrieve and browse information in the database through the Web-browser-based GUI. In the query mode, a form-based Boolean query issued by a user is automatically translated into an SQL query, and the English terms in the query are sent to the Term Translation Module. The Client Module then retrieves documents which match either the original English query or the translated Japanese query. As the indices are names and terms which may consist of multiple words (e.g., "Warren Christopher," "memory chip"), the query terms are delimited in separate boxes in the form, making sure no ambiguity occurs in both translation and retrieval.

In its browsing mode, the Client Module allows the user to browse the information in the database in various ways. For example, once the user selects a particular document for viewing, the client sends it to an appropriate (i.e., English or Japanese) indexing server for creating hyperlinks for the indexed terms, and, in the case of a Japanese document, sends the indexed terms to the Term Translation Module to translate the Japanese terms into English. The result that the user browses is a document each of whose indexed terms are hyperlinked to other documents containing the same indexed terms. Since hyperlinking is based on the original or translated *English* terms, the monolingual English speaker can follow the links to both English and Japanese documents transparently. In addition, the Client Module is integrated with a commercial MT system for a rough translation of the whole text.

The Term Translation Module

The Term Translation Module is used by the Client Module bi-directionally in two different modes. That is, it translates English query terms into Japanese in the query mode and, in reverse, translates Japanese indexed terms into English for viewing of a retrieved Japanese text in the browsing mode.

3.1 Issues Concerning Proper Name Recognition for Browsing and Retrieval

Based on the system description above in the preceding sections, we describe in more detail in the following the impacts of name recognition on multilingual browsing and retrieval.

3.1.1 Indexing Accuracy

To index, the system uses two different configurations of NameTag for English and Japanese. Indexing of names is particularly significant in the Japanese case, where the accuracy of indexing depends on the accuracy of segmentation of a sentence. In English, since words are separated by spaces, there is no issue of indexing accuracy for individual words. However, in languages such as Japanese, where word boundaries are not explicitly marked by spaces, word segmentation is necessary to index terms. However, most segmentation algorithms are more likely to make errors on names, as these are less likely to be in the lexicons. Thus, use of name indexing can improve overall segmentation and indexing accuracy.

3.1.2 Query Disambiguation

As described above, the Indexing Module not only identifies names of people, entities and locations, but also disambiguates types among themselves and between names and non-names. Thus, if the user is searching for documents with the location "Washington" (not a person or company named "Washington") or a person "Clinton" (not a location), the system allows the user to specify, through the GUI, the type of each query term. This ability to disambiguate types of queries not only constrains the search and hence improves retrieval precision, but also speeds up the search time considerably, especially when the size of the database is very large.

3.1.3 Translation Disambiguation

In developing this system, we have intentionally avoided an approach where we first translate foreign-language documents into English and index the translated English texts (Fluhr, 1995; Kay, 1995; Oard and Dorr, 1996). In (Aone et al., 1994), we have shown that, in an application of extracting information from foreign language texts and presenting the results in English, the "MT first, Information Extraction second" approach was less accurate than the approach in the reverse order, i.e., "Information Extraction first, MT second." In particular, translation quality of names by even the best MT systems was poor. In an indexing and retrieval application such as the one under discussion, the proper identification and translation of names are critical.

There are two cases where an MT system fails to translate names. First, it fails to recognize where a name starts and ends in a text string. This is a non-trivial problem in languages such as Japanese where words are not segmented by spaces and there is no capitalization convention. Often, an MT system "chops up" names into words and translates each word individually. For example, among the errors we have encountered, an MT system failed to recognize a person name "Mori Hanae" in kanji characters, segmented it into three words "mori," "hana," and "e" and translated them into "forest," "England," and "blessing," respectively.

Another common MT system error is where the system fails to make a distinction between names and non-names. This distinction is very important in getting correct translations as names are usually translated very differently from non-names. For example, a person named "Dole" in katakana was translated into a common noun "doll." Abbreviated country names for Japan and the United States in single kanji characters, which often occurs in newspapers, were sometimes translated by an MT system into their literal kanji meanings, "day" and "rice," respectively.

The proper name recognition capability provided by NameTag solves both of these problems. NameTag's ability to identify names prevents chopping names into pieces. NameTag's ability to assign semantic types to names makes possible greater precision in translating names.

4 Proper Name Recognition Integrated with Text Clustering

Multimedia Fusion (MMF) is a system SRA developed to provide a tool to help people deal with large incoming streams of multimedia data (Aone, Bennett, and Gorlinsky, 1996). MMF clusters texts automatically into a hierarchical concept tree, and, unlike a typical message routing system, the users do not need to specify beforehand what topics that the incoming texts cluster into. It employs Cobweb-based conceptual clustering (Fisher, 1987), with the feature vectors required for that algorithm supplied by keywords picked from the body of a text based upon their worth as determined by the Inverse Document Frequency (IDF) metric (Church and Gale, 1995). In addition, NameTag is run over the incoming texts (CNN closed-captions and ClariNet news feeds) to identify the proper names in the document (persons, companies, locations).

One of the novel features in this system was the important role of proper name recognition. It

is important to recognize that using white-space-delimited tokens in a text as keywords provides significantly less information than is actually available. The proper name information (for persons, organizations, and locations) adds considerable information to what otherwise would be a meaningless string of tokens.

4.1 Issues Concerning Proper Name Recognition for Text Clustering

4.1.1 Proper Names as Keywords for Clusters of Texts

Proper names are natural keywords to characterize the contents of text clusters. Without proper name recognition, "International Business Machines" is just a string containing three common nouns that may or may not be informative keywords. Recognizing it as a proper name enlarges the set of possible keywords for the document. Second, proper name recognition allows the disambiguation of names from non-names, such as "white" in "white shirt" vs. "Bob White," which enhances the accuracy of keyword selection.

The alias forms generated by NameTag are also of great value to IDF calculations, since we can select one of the name forms of a particular entity occurring within a document ("President Clinton," "Bill Clinton," "Clinton") to be the canonical form for all the name forms. This reduces the chances that alternate forms of the same name will be used as distinct keywords for the same document.

As discussed in (Aone, Bennett, and Gorlinsky, 1996), we quantitatively evaluated the accuracy of clustering, and the use of proper name recognition enhanced the F-measure (a combined measure of recall and precision) from 50% to 61% in clustering in ClariNet news feed.

4.1.2 Tailoring of NameTag to Clustering

The name recognition technology embodied in NameTag had to be customized for MMF, particularly for the closed-caption texts from CNN Headline News. It had to handle all *upper-case* closed caption texts, which pose some challenges due to the absence of case information. In general, lexical information has to be available for name recognition in upper-case text, which tends to have a retarding effect on system performance. In addition, since the closed captions are transcriptions of speech by anchor persons or reporters, characteristics of spoken language had to be accommodated (e.g., "OK" is Oklahoma in a text while it is an answer in a caption).

4.1.3 High Precision vs. High Recall Name Recognition

The proper name recognition used in MMF has to be highly accurate. In name recognition, as in other areas of language technology, there is a trade-off between recall and precision. In applications such as text clustering, high precision is preferred over high recall so that the system does not introduce errors in keyword selection. That is, not recognizing "BILL CLINTON" is more acceptable than mis-recognizing "BILL FOLDER" as a person name.

To handle this, NameTag provides three settings, depending on what kind of application it is being used for: "High Precision," "High Accuracy," and "Normal." The first setting ensures that all names identified are correct, even at the expense of some possibly missed ones. The second setting focuses on identifying all possible names within a document, even at the cost of some false positives. The third setting is a balanced one, aiming at achieving the highest possible combined scores. For MMF, we used the "High Precision" setting to optimize the Cobweb clustering performance.

5 Proper Name Recognition Integrated with Manual Text Indexing

SRA recently developed an an operational indexing system, the Human Indexing Assistant (HIA), which assists the human indexing of an incoming flow of documents. The task involves human indexers who process a large incoming stream of documents and fill out a template with names of products and equipment, as well as companies and individuals involved in the manufacture of those items. Integral to it is the use of NameTag.

HIA's GUI presents the user with three screens, the first containing the original document to be indexed, the second the template to be filled out, and the third used for iconic representations of the indexed material. This third screen serves as a working area where the user, having filled out a template for a name to be indexed, can iconify it, place it in the third screen and make links between it and other iconified objects such as company names. As part of the indexing process, NameTag is first launched from the indexing interface, and it highlights in the first screen the proper names in the document to be indexed. The indexer can then select what they think is appropriate to index and paste the names into templates in the second screen.

5.1 Issues Concerning Proper Name Recognition for Manual Text Indexing

For this application, we used the "high recall" setting of NameTag, as discussed in Section 4.1.3. It was important that as many potential names as possible be identified. It is a part of the indexers' job that no names be missed during indexing. Inserting NameTag into the process required that it gain the indexers' confidence that it could indeed hit all possible names. The indexers were not particularly concerned with misidentification of names, as these could be quickly passed over by the human

(e.g., "BILL FOLDER" as a personal name in all upper-case text). Once the users had confidence in NameTag, it was possible for them to stop reading documents *in toto*, thus producing great increases in the throughput of the operation.

As a side issue in this sort of application, it should be pointed out that information on quality of the indexing when done *without* automated or semi-automated help is rarely available or reliable. The users impose requirements that they themselves may not meet consistently ("whatever you do, your system can't miss any names"), and the developers must work within what may be a more or less fictional framework. However, dealing with issues of this kind is critical to success. Successful insertion of HIA into the workplace involved getting the indexers to buy into its value for them. Ultimately, the deciding factor was the clearly faster rate of indexing with the system than with the line-editing-oriented, totally manual system being replaced.

6 Conclusion

We have discussed three diverse applications in which proper name recognition as embodied in NameTag has played an important role. Clearly, the identification of names can improve the performance of other algorithms and systems. In return, the nature of the application in which name recognition is being used affects whether the name recognition should aim at high coverage or high accuracy.

References

Aone, Chinatsu, Scott William Bennett, and James Gorlinsky. 1996. Multi-media Fusion through Application of Machine Learning and NLP. In *AAAI Spring Symposium Working Notes on Machine Learning in Information Access.*

Aone, Chinatsu, Hatte Blejer, Mary Ellen Okurowski, and Carol Van Ess-Dykema. 1994. A Hybrid Approach to Multilingual Text Processing: Information Extraction and Machine Translation. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA).*

Aone, Chinatsu, Nicholas Charocopos, and James Gorlinsky. 1997. An Intelligent Multilingual Information Browsing and Retrieval System Using Information Extraction. In *Proceedings of the Fifth Conference on Applied Natural Language Processing.*

Church, Kenneth and William Gale. 1995. Inverse document frequency (idf): A measure of deviations from poisson. In *Proceedings of the 3rd Workshop on Very Large Corpora.*

Fisher, Douglas H. 1987. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2.

Fluhr, Christian. 1995. Multilingual information retrieval. In Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors, *Survey of the State of the Art in Human Language Technology.* Oregon Graduate Institute.

Kay, Martin. 1995. Machine translation: The disappointing past and present. In Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors, *Survey of the State of the Art in Human Language Technology.* Oregon Graduate Institute.

Oard, Douglas W. and Bonnie J. Dorr, editors. 1996. *A Survey of Multilingual Text Retrieval.* Technical Report UMIACS-TR-96-19. Institute for Advanced Computer Studies, University of Maryland.