NameTag^{TM} Japanese and Spanish Systems as Used for MET

Chinatsu Aone
Systems Research and Applications Corp. (SRA)
4300 Fair Lakes Court
Fairfax, VA 22033
aonec@sra.com, http://www.sra.com *

1 Introduction

We have participated in the Multilingual Entity Task (MET) for Japanese and Spanish using SRA's multilingual text-indexing software called NameTag^{TM}. Its English version was used for the Named Entity Task (NE) in MUC-6 [2]. The NameTag Japanese and Spanish systems were customized to accommodate the MET-specific requirements and were able to achieve high performance in both recall and precision.

	Jp MET	Sp MET	Eg MUC-6
Tokenizer	y	y	y
Patterns	y	y	y
Lexical data	y	y	y
Alias generator	y	y	y
Morph analyzer	n	optional	n
Segmenter	y	n	n

Table 1: NameTag configurations

2 MET System Description

For MET, we used NameTag in its Japanese and Spanish configurations. NameTag is an automated text indexing system that recognizes and classifies names and other key phrases such as time and numeric expressions. It is an enhanced offspring, implemented in C++, of the preprocessing module of SRA's multilingual natural language processing system [1]. NameTag combines dynamic pattern recognition with static lexical look-up to achieve high recall and precision at high speed.

The NameTag engine is designed for multilingual capabilities. The same engine is used for different languages using language-specific "plug-ins" such as tokenizers, patterns, lexical data, alias generators, morphological analyzers, and segmenters. Table 1 compares plug-ins used for different languages/tasks.

NameTag has several unique features beside being able to handle multiple languages. First, it can generate and link aliases of names automatically using language-specific alias generators. For example, the Spanish system can recognize "Peña" as an alias for "José Francisco Peña Gómez" by generating a paternal name alone. The Japanese system currently performs a limited organization alias recognition such as "DAIWA" as an alias for "DAIWABANK."

Second, SRA's Japanese segmenter, which uses a large lexicon, morphological analysis, and heuristic-based rules to segment a sentence into words, is fully integrated into the NameTag engine. Thus, the system can utilize the results of name recognition in the subsequent segmentation process and increase segmentation accuracy.

Third, NameTag can take advantage of SGML markers to improve performance. For MET, headlines (marked by <SLUG> or <HL>) were processed after the main body of text (marked by <TXT>). This was advantageous for Japanese because any entity recognized in the main body was utilized in segmentation of headlines. It also increased precision for Spanish because only lexicon and alias lookups were applied to all-upper-case headlines, avoiding spurious or erroneous names generated by patterns.

Finally, NameTag provides a GUI-based multilingual development environment which facilitates rapid development of patterns.

3 MET Results and Analysis

The MET final blind tests were conducted using 100 Kyodo articles for Japanese and 100 AFP articles for Spanish. These articles were retrieved using the keyword "press conference." Thus, they encompassed various subject domains, including business, politics, sports, and arts, unlike the Wall Street Journal arti-

*I'd like to thank Mila Ramos-Santacruz and Misa Miyachi, who assisted me with manual tagging of development texts and lexical development for Spanish and Japanese respectively, and Kevin Hausman, who is the principal developer of the C++ NameTag engine.

cles used for MUC-6. In addition, the MET guidelines had additional requirements not found in MUC-6, such as tagging of relative dates and tagging organizations as locations when they are used as facilities.

Despite the differences in types of articles and the additional requirements, the NameTag Japanese and Spanish systems achieved high performance in both recall and precision. The Japanese system is slower (15MB/h on a SPARC 20) than its Spanish and English counterparts (93 MB/h and 80 MB/h respectively) because of the segmentation overhead. Both the Japanese and Spanish systems still have room for higher recall because of shorter development time but also partly because of difficult language-specific issues to be solved, which we will discuss below.

3.1 Japanese-specific Issues

The MET evaluation has revealed several Japanese-specific challenges which must be solved in order for the system to achieve even higher performance. First, we have encountered what we call *chicken-and-egg* problems. Good name recognition requires good segmentation, as name recognition patterns rely on properties of words segmented by the segmenter such as part-of-speech and other linguistic attributes. However, good segmentation, in turn, relies on good name recognition, as names are usually not in the lexicons and thus tend to cause segmentation errors. As discussed in Section 2, NameTag can utilize the results of name recognition in subsequent segmentation to partially solve this problem. Additionally, it is essential that the segmenter be more robust and accurate in order to improve performance on name recognition and other Japanese text processing tasks even further.

Another *chicken-and-egg* problem was encountered in constructions where a person name and an organization name appear next to each other in a sentence and there is no delimiter between the two (e.g. "ABCDEFG" where ABC is a person name and DEFG an organization name with no space or other punctuation in-between.) Here, recognizing the person name requires recognizing the adjacent organization name first while recognizing the organization name requires recognizing the person name first. In these cases, the system often mistags the whole string as a person and misses the organization name.

The second big challenge is dealing with Japanese aliases, which are more complex than English aliases. The NameTag Japanese system currently generates aliases like "SILICONGRAPHICS" for "SILICONGRAPHICSCORP." by stripping off certain corporate designators at the end of names. But it does not currently generate an alias which is a character subsequence of its full name like "NIKKOU" for "NIHONKOUKUU." Since aliases are, by definition, already recognized as names in a given article, they often appear in contexts where patterns do not apply. In these cases, not generating aliases results in missing names (i.e., loss in recall).

3.2 Spanish-specific Issues

In addition to the general differences between MET and the MUC-6 NE task described earlier, there were a few Spanish-specific issues which had to be tackled for MET.

In the MET Spanish articles, the capitalization convention was rather unpredictable (e.g., "Oficina de lucha contra la droga," "puerto cubano de Mariel"). Thus, capitalization clue was not as relevant in Spanish MET texts as English WSJ texts in recognizing proper names. Consequently, the Spanish system needed to perform *deeper* analysis of the texts to achieve comparable results.

The presence of "de" in Spanish person names has made person name recognition more difficult, as "de" is also a preposition, and sometimes caused a Spanish version of *chicken-and-egg* problem. For example, the system thought "Valle Rivas" in "Olijela del Valle Rivas" was a location as "valle" also means "valley." On the other hand, it tagged "Roverto Marquevich de San Isidro" as a full person name though "de" is a preposition and "San Isidro" is a location name.

4 Summary

The MET evaluation has proved that NameTag can be ported to other languages with a level of performance similar to English, despite various language-specific challenges. We plan to port it to other unsegmented languages such as Chinese and Thai in addition to other European languages.

References

[1] Chinatsu Aone, Sharon Flank, Paul Krause, and Doug McKee. SRA: Description of the SOLOMON System as Used for MUC-5. In *Proceedings of Fourth Message Understanding Conference (MUC-5)*, 1993.

[2] George Krupka. SRA: Description of the SRA System as Used for MUC-6. In *Proceedings of Sixth Message Understanding Conference (MUC-6)*, 1995.