CRL's APPROACH TO MET

Jim Cowie
jcowie@nmsu.edu
Computing Research Laboratory
New Mexico State University, Box 30001/3CRL, Las Cruces, NM, 88003

Summary

From February to April CRL carried out investigations into the modification of our English name recognition software developed for MUC-6 [1] to Chinese and Spanish. In addition a Japanese system, developed under Tipster Phase I [2], was modified to comply with the MET task. Finally learning methods developed for MUC-6 were adapted to handle Chinese. All systems performed with good levels of accuracy and it is clear that further tuning and refinement, for which there was no time or resources, would lead to even higher levels of performance.

A Pattern Based Approach

Three pattern based systems were produced for Spanish, Japanese, and Chinese. The basic operation of each is the same, although in the case of Chinese very few patterns were produced due to time and resource constraints. The data for each system was derived from the training texts plus additional data derived from corpora and other CRL data developed for machine translation. No attempt was made to segment the texts used in the Chinese system. The Japanese system used information produced by the Juman segmentor, from Kyoto University.

It is important to have basic tools for handling two byte character sets. A lookup program and a sort have to be able to handle character order based on two bytes. These had already been developed as part of CRL's Tipster phase I work. Given these and a tool to be used for human tagging and a program to extract tagged text of different types and place it in files it is possible to develop a naive system which recognizes previously tagged names very rapidly.

Thus -

```
HUMAN + TEXTS + TEXT TAGGING TOOL -
> TAGGED TEXTS

TAGGED TEXTS + EXTRACTION PROGRAM -
> KNOWN LISTS

KNOWN LISTS + SORT - > UP LISTS
```

and finally

```
LOOKUP LISTS + TEXT + LOOKUP PRO-
GRAM -> TAGGED TEXTS
```

The obvious problems here are that new names will appear and short forms will be recognized in contexts where they are incorrect. For Chinese, however, this approach already gave reasonably high performance both on test and evaluation data.

The next stage is to develop lexicons containing semantic tags for component parts of names, and also for indicators that occur in the surrounding text. For example beginning organization words in CRL's Spanish system include the following - ASOCIACION, BANCO, BANKHAUS, BOLSA. These too can be partly derived from the training data and then extended manually. Lists of proper names, locations, titles, professions and known words in the language are also used. For personal names in Spanish the English name lists which were partitioned into unambiguous names (in English e.g. CHARLES) and ambiguous names (e.g. BAKER) were repartitioned to lists of ambiguous and unambiguous names in Spanish using a morphological analyzer and a Spanish dictionary (CHARLES is now ambiguous and BAKER not). The Japanese system was already constructed in this manner and similar developments on a lesser scale carried out for Chinese. Sets of patterns which recognize the basic semantic tags are now developed and coded up as 'lex' pattern recognizers. So our system now looks like -

```
LOOKUP LISTS + TEXT + LOOKUP PRO-
GRAM -> PARTIALLY TAGGED TEXTS
(PTT)

SEMANTIC LISTS + P T T + LOOKUP
PROGRAM - > MARKED TEXT

MARKED TEXTS + PATTERN RECOGNIZERS
-> TAGGED TEXTS
```

At this point it is possible to further develop patterns and data by scoring system performance against some, unseen, set of pre-tagged texts. Some short forms of names are, however, particularly difficult to identify in isolation and certain segments of the text, in particular capitalized headlines in Spanish, are also difficult if the system uses capitalization. This is most easily handled by a post tag stage which takes names discovered in the text and generates a list of names which are then converted to appropriate short forms. A final pass through the whole texts using the lookup tool and the list of full and abbreviated names catches many of the short forms.

There is obviously some point at which the recognition and classification gets difficult. The rules for MET made a distinction between an organization name and the same name used to indicate a locations; this can be a very vague distinction and one probably better handled by other stage in the system. Also a wide range of entities were allowed as organizations, typical patterns for pop group names can be a lot more eccentric than company names, a grainier division would separate some difficult types from more conventionally usages. Finally there are some forms where a great deal of context and/ .or world knowledge are needed to arrive at a correct classification. These fortunately are only a few percent of the total cases in the type of texts used for MET.

Dates and numbers were also recognized using patterns of semantic types. These performed extremely accurately with high precision and recall.

Problems

The main problems were lack of time and people to work on the task. An estimated two man months of effort was put into all three systems. An additional problem lies in the original design of our systems which do not allow multiple semantic tags on words or phrase. Thus a word might be a part of a person name and a location but it is tagged as a location due to the ordering of the tagging process. This in turn requires more complex and less accurate patterns to be used, which allow locations as part of peoples names and so on. Clearly a more complicated lexical structure would allow simpler patterns. Given the short development time for the MET evaluation it was not possible to develop the necessary infrastructure to support this more complex lexicon.

A Learning Based Approach

Our learning system has already been described in the MUC-6 Proceedings [2]. It is based on automatically creating decision trees using Quinlan's ID3 algorithm. These decision trees use a limited context of text in the case of Chinese four pairs of characters before and four pairs after a point in the text to decide if a name starts (or ends) at this position. High precision, and low recall results were possible using the extremely small amount of training data available for MET. CRL extended this data by hand tagging all the texts provided for MET. Additional data was then generated by tagging known names in XinHua newswire and extracting these and their contexts to provide further training data. This increase in positive examples increased both precision and recall significantly. The final problem with adapting the method to Chinese was finding unmatched brackets and fixing them. One decision tree finds the start of an organization, but the other fails to identify the end. In English, once the missing bracket is identified, a simple heuristic based on capitalization, conjunction and some other common words (e.g. of) was used to fix the missing bracket. For Chinese we developed a frequency list of characters occurring before and after each name type. This was used to compute the most likely insertion point for the missing bracket. This approach needs further analysis to determine its success.

The adaptation of the English learning system to Chinese took around one and a half man months.

References

1. Cowie, J, "Description of the CRL/NMSU Systems Used for MUC-6" in *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Morgan-Kaufman, 1996.

2. Cowie, J., Guthrie, L.,et al. "The DIDEROT System" in *Proceeding of the Tipster Text Program (Phase I)*, 223-235, Morgan-Kaufman,1993.