INTERPRETATION OF PROPER NOUNS FOR INFORMATION RETRIEVAL

Woojin Paik [1], *Elizabeth D. Liddy* [1], *Edmund Yu* [2], *Mary McKenna* [1]

[1] School of Information Studies
Syracuse University
Syracuse, NY 13244

[2] College of Engineering and Computer Science
Syracuse University
Syracuse, NY 13244

1. INTRODUCTION

Most of the unknown words in texts which degrade the performance of natural language processing systems are proper nouns. On the other hand, proper nouns are recognized as a crucial source of information for identifying a topic in a text, extracting contents from a text, or detecting relevant documents in information retrieval (Rau, 1991).

In information retrieval, proper nouns in queries frequently serve as the most important key terms for identifying relevant documents in a database. Furthermore, common nouns (e.g. 'developing countries') or group proper nouns (e.g. 'U.S. government') in queries sometimes need to be expanded to their constituent set of proper nouns in order to serve as useful retrieval terms. We have implemented two solutions to this problem: one approach is to expand a term in a query such as 'U.S. government' to all possible names and variants of United States government entities. Another approach assigns categories from a proper noun classification scheme to every proper noun in both documents and queries to permit proper noun matching at the category level. Category matching is more efficient than keyword matching if the request is for an entity of a particular type. For example, queries about government regulations of use of agrochemicals on produce from abroad, require presence of the following proper noun categories: government agency, chemical and foreign country.

Our proper noun classification scheme, which was developed through corpus analysis of newspaper texts, is organized as a hierarchy which consists of 9 branching nodes and 30 terminal nodes. Currently, we use only the terminal nodes to assign categories to proper nouns in texts. Based on an analysis of 588 proper nouns from a set of randomly selected documents from Wall Street Journal, we found that our 29 meaningful categories correctly accounted for 89% of all proper nouns in texts. We reserve the last category as a miscellaneous category. Figure 1 shows a hierarchical view of our proper noun categorization scheme.

2. BOUNDARY IDENTIFICATION

The proper noun processor herein described is a module in the DR-LINK System (Liddy et al, in press) for document detection being developed under the auspices of DARPA's TIPSTER Program. In our implementation, documents are first processed using a probabilistic part of speech tagger (Meeter et al, 1991) and general-purpose noun phrase bracketter which identifies proper nouns and proper noun phrases in texts. We have developed a special purpose proper noun phrase boundary identification module which extends the proper noun bracketting to include proper noun phrases with embedded conjunctions and prepositions. The module utilizes heuristics developed through corpus analysis. The success ratio is approximately 95%. Incorrectly identified proper noun phrases are due mainly to two reasons: 1) the part of speech tagger identifies common words as proper nouns; and, 2) conflicts between the general-purpose noun phrase bracketter and the special-purpose proper noun boundary identifier. While the first source of error is difficult to fix, we are currently experimenting with applying the special purpose proper noun boundary identifier before the general-purpose noun phrase bracketter. Our preliminary results show that this would result in a 97% correct ratio for identifying boundaries of proper nouns.

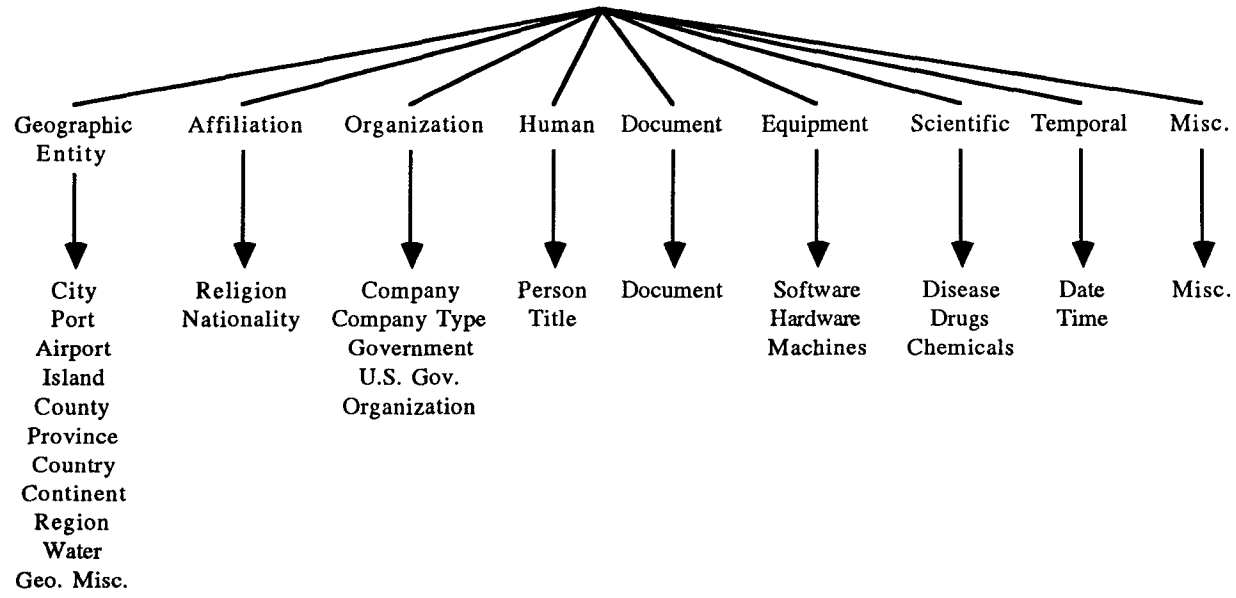

Figure 1: Proper Noun Categorization Scheme

3. CATEGORIZATION

Next, the system categorizes all the identified proper nouns using several methods:

1) comparison to lists of known prefixes, infixes and suffixes for each category of proper noun;
2) consulting an alias database consisting of alternate names for some proper nouns;
3) look-up in a proper noun knowledge-base of proper nouns and their categories extracted from online lexical resources (e.g., World Factbase, Gazetteer), and finally;
4) applying context heuristics developed from corpus analysis of the contexts which suggest certain categories of proper nouns.

While being categorized, the proper nouns are standardized in three ways:

1) prefixes, infixes, and suffixes of proper nouns are standardized;
2) proper nouns in alias forms are translated into their official form, and;
3) the partial string of a proper noun which was mentioned in full earlier in the document is co-indexed for reference resolution.

A new field containing the list of each standardized proper noun and its category code is added to the document for later use in several stages of matching and representation. The first two techniques improve retrieval performance, while the co-indexing of references produces a full representation of a proper noun entity and all its accompanying information. Figure 2 shows a schematic view of DR-LINK's proper noun categorizer.

4. USE OF PROPER NOUN IN MATCHING

When matching documents to queries, either the lexical entry for the proper noun can be matched or the match can be at the category level, as each proper noun occurring in a document is recorded in the proper noun field of the document along with its appropriate category code. For example, if a query is about a business merger, we can limit the potentially relevant documents to those documents which contain at least two different company names, flagged by two company category codes in the proper noun field. For many queries, using the standardized form of a proper noun reduces the number of possible variants which the system would otherwise need to search for. For example, 'MCI Communications Corp.', 'MCI Communications', and 'MCI', are all standardized as 'MCI Communications CORP' by our proper noun categorizer. This process is similar in purpose to the common practice in standard retrieval matching of reducing variants by stemming. However, stemming is not a viable means for standardizing proper names.

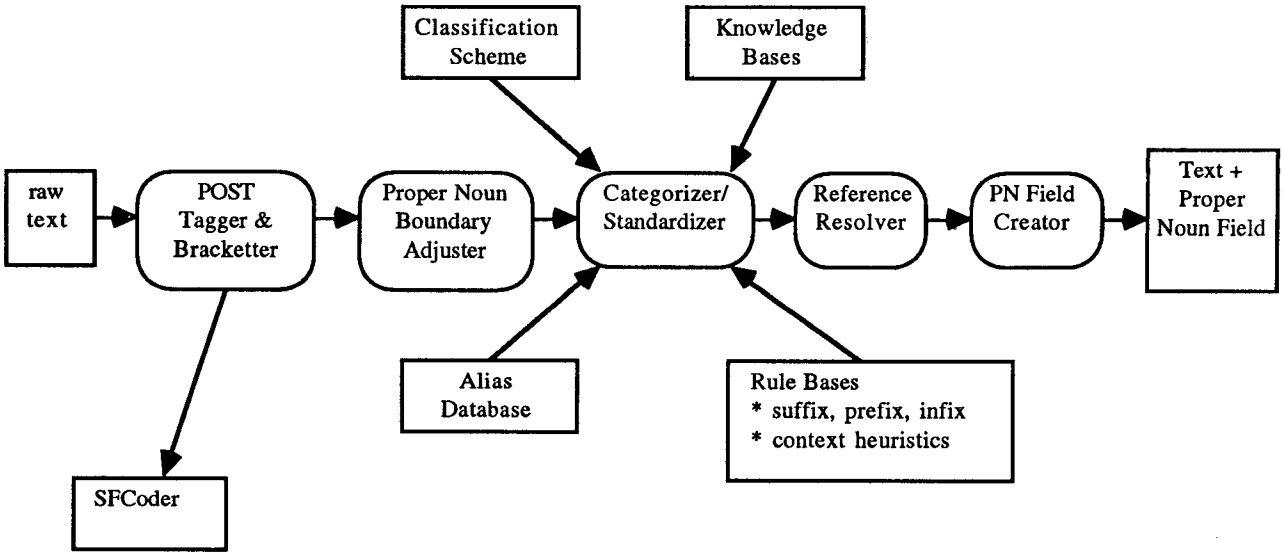

Figure 2: DR-LINK Proper Noun Categorizer

While the category matching strategy is useful in many cases, an expansion of a group proper noun such as 'European Community', which occurs in a query, to member country names is also beneficial. Relevant documents to a query about sanctions against Japan by European Community countries are likely to mention actions against Japan by member countries by name rather than the term in the query, European Community. We are currently using a proper noun expansion database with 168 expandable entries for query processing. In addition, certain common nouns or noun phrases in queries such as 'socialist countries' need to be expanded to the names of the countries which satisfy the definition of the term to improve performance in detecting relevant documents. The system consults a list of common nouns and noun phrases which can be expanded into proper nouns and actively searches for these terms during the query processing stage. Currently, the common noun expansion database has 37 entries.

The creation and use of proper noun information is first utilized in DR-LINK system as an addition to the subject-content based filtering module which uses a scheme of 122 subject field codes (SFCs) from a machine readable dictionary rather than keywords to represent documents. Although SFC representation and matching provides a very good first level of document filtering, not all proper nouns reveal subject information, so the proper noun concepts in texts are not actually represented in the SFC vectors.

For processing the queries for their proper noun requirements, we have developed a Boolean criteria script which determines which proper nouns or combinations of proper nouns are needed by each query. This requirement is then run against the proper noun field of each document to rank documents according to the extent to which they match this requirement. In the recent testing of our system, these values were used to rerank the ranked list of documents received from the SFC module. The results of this reranking placed all the relevant documents within the top 28% of the database. It should also be noted that the precision figures on the output of the SFC module plus the proper noun matching module produced very reasonable precision results (.22 for the 11-point precision average), even though the combination of these two modules was not intended to function as a stand-alone retrieval system.

Also, the categorization information of proper nouns is currently used in the system's later module which extract concepts and relations from text to produce a more refined representation. For example, proper nouns reveal the location of a company or the nationality of an individual. The proper noun extraction and categorization module, although developed as part of the DR-LINK System, could be used to provide improved document representation for any information retrieval system, because it permits queries and documents to be matched with greater precision and the expansion functions improve recall.

5. PERFORMANCE EVALUATION

While we have processed more than one gigabyte of text using the current version of the proper noun categorizer for the TIPSTER 18 month testing, the evaluation of

the proper noun categorizer herein reported is based on 25 randomly selected Wall Street Journal documents which were compared to the proper noun categorization done by a human. Table 1 shows the categorizer's performance over 588 proper nouns occurring in the test set. In addition to 588 proper nouns, 14 common words were incorrectly identified as proper nouns due to errors by the part of speech tagger and typos in the original text; and the boundaries of 17 proper nouns were incorrectly recognized by the general-purpose phrase bracketter error.

	Total Correct	Total Incorrect	Precision *
City	11	33	0.25
Port	10	2	0.83
Province	23	1	0.96
Country	66	1	0.99
Continent	1	0	1.00
Region	1	7	0.13
Religion	2	0	1.00
Nationality	32	2	0.94
Company	87	13	0.87
Government	5	1	0.83
U.S. Gov.	20	8	0.71
Organization	9	1	0.90
Person	48	57	0.46
Title	42	4	0.91
Document	1	2	0.33
Machine	0	1	0.00
Date	27	0	1.00
Misc.	65	0	1.00
TOTAL	450	133	0.77
TOTAL-Misc.	385	133	0.74

$$\text{* Precision} = \frac{\text{Total \# Correct}}{\text{Total \# Correct + Total \# Incorrect}}$$

Table 1: DR-LINK Proper Noun Categorizer Performance

65 proper nouns were correctly categorized as miscellaneous as they did not belong to any of our 29 meaningful categories. This may be considered a coverage problem in our proper noun categorization scheme, not an error in our categorizer. Some examples of the proper nouns belonging to the miscellaneous category are: 'Promised Land', 'Mickey Mouse', and 'IUD'. The last row of Table 1 shows the overall precision of our categorizer based on the proper nouns which belong to the 29 meaningful categories.

	Total Correct	Total Incorrect	Total Missing	Recall *
With Miscellaneous Category	450	133	17	0.75
Without Miscellaneous Category	385	133	17	0.72

$$\text{* Recall} = \frac{\text{Total \# Correct}}{\text{Total \# Actual}}$$

Total # Actual =
Total # Correct + Total # Incorrect + Total # Missing

Table 2: DR-LINK Categorizer Overall Recall

Most of the wrongly categorized proper nouns are assigned to the miscellaneous category, not mis-categorized to another meaningful category. The only notable case where a proper noun was mis-categorized as another meaningful category, occurred between the city and the province categories. Our categorizer assigned the province category (IDA's Gazetteer calls states provinces) to 'New York' when the proper noun was actually referring to the name of the city.

Errors in the categorization of person and city names account for 68% of the total errors. To correct the categorization errors in person names, we are currently experimenting with a list of common first names as a special lexicon to consult when there is no match in prefix and suffix lists nor any context clues to other meaningful categories. The main reason for mis-categorizing city names as miscellaneous proper nouns was due to a special convention of newspaper text. The locational source of the news, when mentioned at the beginning of the document, is usually capitalized. For example, if the story is about a company in Dallas then the text will start as below:

DALLAS: American Medical Insurance Inc. said that ...

This problem will be solved in the new version of our proper noun categorizer by incorporating a capitalization normalizer, which converts words in all upper case to lower case except the first character of a word, before the part of speech tagging. We are also in the process of incorporating context information for identifying city names in our categorizer based on the observation that city names are usually followed by a country name or a province name from the United States and Canada.

Low precision in categorizing region names such as 'Pacific Northwest' is due to incomplete coverage of possible region names in the proper noun database. We are currently developing a strategy based on context clues using locational prepositions.

Table 2 shows the overall recall figure of our categorizer which is affected by the proper noun phrase boundary identification errors caused by the general-purpose phrase bracketter.

6. CONCLUSION

In comparing our proper noun categorization result to others in the literature, Coates-Stephens' (1992) result on acquiring genus information of proper nouns was contrasted to our overall precision. While his approach is to acquire information about unknown proper nouns' detailed genus and differentia description, we consider our approach of assigning a category from a classification scheme of 30 classes to an unknown proper noun generally similar in purpose to his acquisition of genus information.

Based on 100 unseen documents which had 535 unknown proper nouns, FUNES (Coates-Stephens, 1992) successfully acquired genus information of 340 proper nouns. Of the 195 proper nouns not acquired, 92 were due to the system's parse failure. Thus, the success ratio based on only the proper nouns which were analyzed by the system, was 77%. DR-LINK proper noun categorizer's overall precision, which is computed with the same formula, was 75%, including proper nouns which were correctly categorized as miscellaneous.

Katoh's (1991) evaluation of his machine translation system, which was based on translating the 1,000 most frequent names in the AP news corpus, 94% of the 1,000 names were analyzed successfully. Our precision figure of categorizing person names was 46%. However, Katoh's system kept a list of 3,000 entries as a system lexicon before the testing. Thus, a considerable number of the 1,000 most frequent names would have been already known, while DR-LINK system's proper noun categorizer had only 47 entries of person names in the proper noun knowledge base before the testing. Therefore, we believe that the performance of our person name categorization will improve significantly by the addition of a list of common first names in our knowledge base.

Finally, the evaluation result from Rau's (1991) company name extractor is compared to the precision figure of our company name categorization. Both system relied heavily on company name suffixes. Rau's result showed 97.5% success ratio of the program's extraction of company names that had company name suffixes. Our system's precision figure was 87%. However, it should be noted that our results are based on all company names, even those which did not have any clear suffixes or prefixes.

REFERENCES

Coates-Stephens, S. (1992). The Analysis and Acquisition of Proper Names for Robust Text Understanding. Unpublished doctoral dissertation, City University, London.

Katoh, N., Uratani, N., & Aizawa, T. (1991). Processing Proper Nouns in Machine Translation for English News. Proceedings of the Conference on 'Current Issues in Computational Linguistics', Penang, Malaysia.

Liddy, E.D., Paik, W., Yu, E.S., & McVearry, K.: (In press). An overview of DR-LINK and its approach to document filtering. Proceedings of the Human Language Technology Workshop. Princeton, NJ: March 1993.

Meteer, M., Schwartz, R. & Weischedel, R. (1991). POST: Using probabilities in language processing. Proceedings of the Twelfth International Conference on Artificial Intelligence. Sydney, Australia.

Rau L. (1991). Extracting Company Names from Text. Proceedings of the Seventh Conference on Artificial Intelligence Applications. Miami Beach, Florida.