CATEGORIZING AND STANDARDIZING PROPER NOUNS FOR EFFICIENT 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
{wjpaik,liddy,memckenn}@mailbox.syr.edu

[2] College of Engineering and Computer Science
Syracuse University
Syracuse, NY 13244
esyu@mailbox.syr.edu

Abstract

In this paper, we describe the most recent implementation and evaluation of the proper noun categorization and standardization module of the DR-LINK document detection system being developed at Syracuse University, under the auspices of ARPA's TIPSTER program. We also discuss the expansion of group common nouns and group proper nouns to enhance retrieval recall. Successful proper noun boundary identification within the part of speech tagger is essential for successful categorization. The proper noun classification module is designed to assign a category code to each proper noun entity, using 30 categories generated from corpus analysis. Standardization of variant proper nouns occurs at three levels of processing. Expansion of group proper nouns and group common nouns is performed on queries. Standardization and categorization is performed on queries and documents. DR-LINK's overall precision for proper noun categorization was 93%, based on 589 proper nouns occurring in the evaluation set.

1. Introduction

In information retrieval, proper nouns, group proper nouns, and group common nouns present unique problems. Proper nouns are recognized as an important source of information for detecting relevant documents in information retrieval and extracting contents from a text (Rau, 1991). Yet most of the unknown words in texts which degrade the performance of natural language processing information retrieval systems are proper nouns. Group proper nouns (e.g., Middle East) and group common nouns (e.g., third world) will not match on their constituents unless the group entity is mentioned in the document. 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 ARPA's TIPSTER program.

Our approach to solving the group common noun and the group proper noun problem has been to expand the appropriate terms in a query, such as 'third world,' to all possible names and variants of third world entities. For all proper nouns, our system 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 query requires entities of a particular type. Standardization provides a means of efficiently categorizing and retrieving documents containing variant forms of a proper noun.

2. Proper Noun Boundary Identification

In our most recent implementation, which has improved from our initial attempt (Paik et al, in press), documents are first processed using a probabilistic part of speech tagger (Meeter et al, 1991). Then a proper noun boundary identifier utilizes proper noun part of speech tags from the previous stage to bracket adjacent proper nouns. Additionally, heuristics developed through corpus analysis are applied to bracket proper noun phrases with embedded conjunctions and prepositions as one unit. For example, a list of proper nouns will be bracketed with non-adjacent proper nouns, if 'of' is an embedded preposition. Some examples of preceding proper nouns include Council, Ministry, Secretary, University, etc.

The success of ratio of our proper noun boundary identification module is approximately 96% in comparison to our initial system's 95% (Paik et al, in press). This improvement was achieved by the re-ordering of the data flow. A general-purpose phrase bracketter, which was applied before the proper noun boundary identification heuristics for non-adjacent proper nouns, is now applied to texts after all the proper noun categorization and standardization steps. Thus, we have eliminated one major source of error, which is the conflict between the general-purpose noun phrase bracketter and the proper noun boundary identification heuristics. For example, embedded prepositions in a proper noun phrase are sometimes recognized as the beginnings of prepositional phrases by the general-purpose phrase bracketter. The remaining 3% of error is due mainly to incorrect proper noun tags assigned to the uncommon first word of a sentence by the part of speech tagger.

3. Proper Noun Classification Scheme

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.

The system categorizes all identified proper nouns using several methods. The first approach is to compare the proper noun with a list of all identified prefixes, infixes and suffixes for possible categorization based on these lexical clues. If the system cannot identify a category in this stage, the proper noun is passed to an alias database to determine if the proper noun has an alternate name form. If this is the case, the proper noun is standardized and categorized at this point. If there is no match in the alias database, the proper noun moves to the knowledge-base look up. These databases have been constructed using online lexical resources including the Gazetteer, the World Factbase, and the Executive Desk Reference. If the knowledge-base look up is not successful, the proper noun is run through a context hueristics application developed from corpus analysis, which suggests certain categories of proper nouns. For example, if a proper noun is followed by a comma and another proper noun, which has been identified as a state, we will label the proper noun as a city name, e.g., Time, Illinois. Finally, if the proper noun has still not been categorized, it is compared against a list of first names generated from the corpus for a final personal name categorization check. If the proper noun has not been categorized at this stage, it will be labeled with the 'miscellaneous' category code.

For the categorization system to work efficiently, variant terms must be standardized. This procedure is performed at three levels, with the prefixes, infixes and suffixes standardized first. Next, the proper nouns in alias forms are standardized into the official form where available. These standardization techniques improve the retrieval performance. Finally, if a proper noun was mentioned at least twice in a document, for instance, Analog Devices, Inc. and later as Analog Devices, a partial string match of a proper noun is co-indexed for reference resolution. This technique allows for a full representation of a proper noun entity. Figure 2 illustrates the flow of the proper noun categorization system within the first stages of DR-LINK processing.

When standardization and categorization have been completed, a new field is added to both the query and the document containing the proper noun and the corresponding category codes. These fields are then used for efficient matching and representation.

4. Use of Proper Nouns in Matching

Both the lexical entry for the proper noun or the category code may be used for matching documents to queries. For example, if a query is about a boarder incursion, we can limit the potentially relevant documents to those documents which contain at least two different country names, flagged by the two country category codes in the proper noun field. Using the standardized form of a proper noun reduces the number of possible variants which the system would otherwise need to search for.

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 for a query about sanctions against Japan by European Community countries are likely to mention actions against Japan by member countries by name

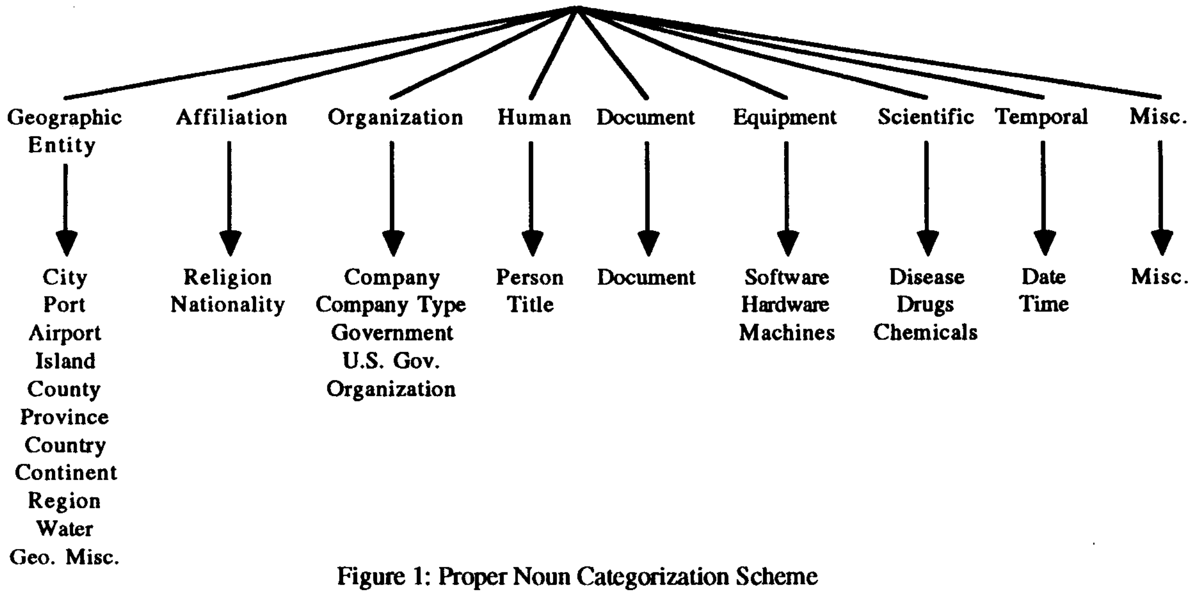

Figure 1: Proper Noun Categorization Scheme

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 the 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.

In our new implementation, categorized and standardized proper nouns are combined with Text Structure (Liddy et al, in press-b) information for matching queries against documents. Text Structure is a recognition of a discernible, predictable schema of texts of a particular type. The Text Structurer module in the DR-LINK system delineates the discourse-level organization of document content so that processing at later stages can focus on those components identified by the Text Structurer as being the most likely location in the document where the information requested in a query is to be found.

All proper nouns in a document collection are indexed in an inverted file with the document accession number, the Text Structure component in which the proper noun was located, and the category code. 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 from certain Text Structure components in each query. These requirements are then run against the proper noun inverted file to rank documents according to the extent to which they match these requirements. Also, the categorization information of proper nouns is currently used in a later module of the system, which extracts concepts and relations from text to produce a more refined representation. For example, proper nouns may reveal the location of a company or the nationality of an individual.

We do not have information retrieval evaluation results based on the new implementation using the proper noun information in conjunction with the Text Structure information. However, in previous testing of our initial system which did not utilize Text Structure information (Paik et al, in press), reranking of documents received from the SFC module, based on the degree of proper noun requirements matching a set of queries against a document collection, resulted in placing all the relevant documents within the top 28%

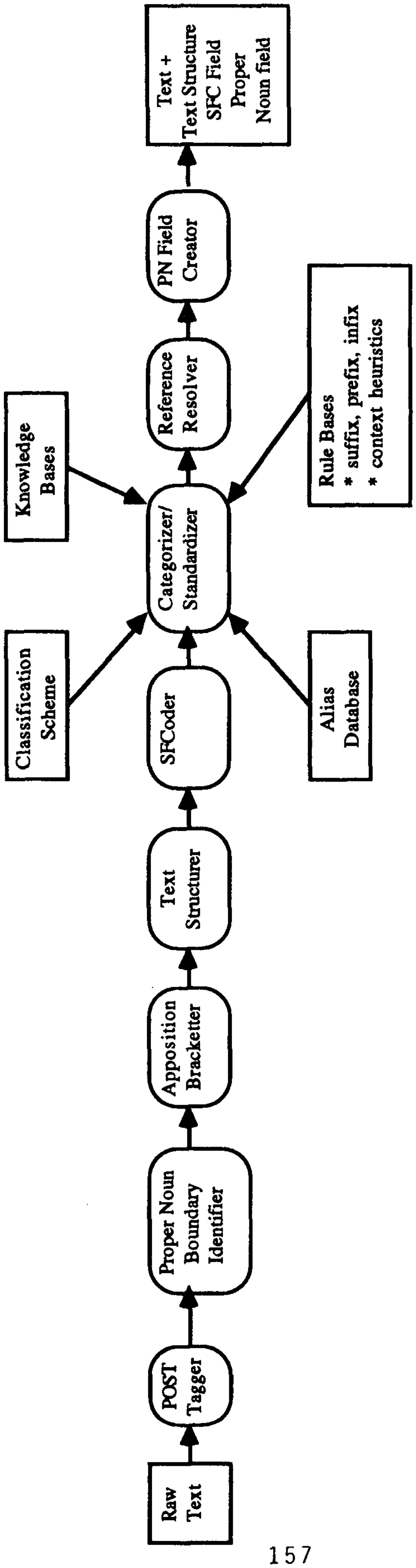

Figure 2: DR-LINK Proper Noun Categorizer

of the document collection. 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 never intended to function as a stand-alone retrieval system.

Finally, 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. The standardization and categorization features permit queries and documents to be matched with greater precision, while the expansion functions of group proper nouns and group common nouns improve recall.

5. Performance Evaluation

While we are currently processing more than one gigabyte of text using the new version of the proper noun categorizer for the TIPSTER 24 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. This document set was also used in evaluating our initial version of the categorizer (Paik et al, in press). Table 1 demonstrates the performance of the categorizer on 589 proper nouns occurring in the test set. In addition to 589 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 11 proper nouns were incorrectly recognized due to unusual proper noun phrases such as, 'Virginia Group to Alleviate Smoking in Public', which the proper noun boundary identification heuristics has failed to bracket.

64 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	Precision *
City	44	0	1.00
Port	10	2	0.83
Province	24	0	1.00
Country	67	0	1.00
Continent	1	0	1.00
Region	1	7	0.13
Geo. Misc.	0	3	0.00
Religion	2	0	1.00
Nationality	33	1	0.97
Company	88	12	0.88
Government	5	1	0.83
U.S. Gov.	23	5	0.92
Organization	13	0	1.00
Person	96	9	0.90
Title	44	2	0.96
Document	3	1	0.75
Machine	0	1	0.00
Date	27	0	1.00
Misc.	64	0	1.00
TOTAL	545	44	0.93
TOTAL-Misc.	481	44	0.92

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

Table 1: DR-LINK Proper Noun Categorizer Performance

In our initial implementation (Paik et al, in press), errors in categorizing person and city names accounted for 68% of the total errors. To improve performance, we added a list of common first names, which was semi-automatically extracted from Associated Press and Wall Street Journal corpora, as a special lexicon to consult when there is no match using all categorization procedures. This addition improved our precision of categorizing person names from the initial system's 46% to 90%.

The errors in categorizing city names, in our initial categorizer, were mainly due to two problems. They are:

1) The locational source of the news, when mentioned at the beginning of the document, is usually capitalized in Wall Street Journal. This special convention of

newspaper texts caused miscategorizing the locational proper nouns (usually city names) as a miscellaneous; and
2) City names which were not in our proper noun knowledge base

The first problem was handled in our new proper noun categorizer by moving the locational information of the news story to a new field, '<DATELINE>', and normalizing capitalization (from all upper case texts to mixed case) at the document preprocessing stage before the part of speech tagging. For example, if a story is about a company in Dallas then the text will be as below:

```
<DOC>
DALLAS: American Medical Insurance Inc. said that ...
...
</DOC>
```

After the new preprocessing module is applied, the text will be as below:

```
<DATELINE> Dallas </DATELINE>
<DOC>
 American Medical Insurance Inc. said that ...
...
</DOC>
```

For the second problem, we incorporated a context rule for identifying city names to our categorizer. The rule is that city names are followed by a country name or a province name from the United States and Canada unless the name is very well known. For example, 'Van Nuys', can now be categorized as a city name as it is preceded by a valid United States province name.

... Van Nuys, Calif. ...

By adding the above new procedures to our categorization system as well as some well known city names which are not province capitals or heavily populated places based on IDA's Gazetteer to our proper noun knowledge base, the precision of categorizing city names has improved from initial system's 25% to 100%.

The overall precision of our new proper noun categorizer has improved to 93% from 77% based on our initial attempt (Paik et al, in press) including proper nouns which are correctly categorized as miscellaneous. This significant advancement was achieved by adding a few sensible context heuristics and modification of the knowledge base. These additions or modifications were based on the analysis of randomly selected documents.

We feel the limitations of not manually updating our proper noun knowledge base for uncommon proper nouns when confronted with proper nouns such as 'Place de la Reunion' and 'Pacific Northwest'. Thus, we are currently developing a strategy based on context clues using locational prepositions as well as appositional phrases to improve categorization of uncommon proper nouns.

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.

	Total Correct	Total Incorrect	Total Missing	Recall *
With Miscellaneous Category	545	44	11	0.91
Without Miscellaneous Category	481	44	11	0.90

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

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

Table 2: DR-LINK Categorizer Overall Recall

8. Conclusion

To compare our proper noun categorization results to the evaluation of a system with similar goals in the literature, we chose Coates-Stephens' (1992) result on acquiring genus information of proper nouns to compare 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. However, it should be noted that our method for assigning categories to proper nouns is different from Coates-Stephens' method, as we rely more on built-in knowledge bases while his approach relies more on context.

Based on 100 unseen documents which had 535 unknown proper nouns, FUNES (Coates-Stephens, 1992) successfully acquired genus information for 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's proper noun categorizer's overall precision, which is computed with the same formula, was 93%, 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, successfully analyzed 94% of the 1,000 names. Our precision figure of categorizing person names was 90%.

Finally, the evaluation result from Rau's (1991) company name extractor is compared to the precision figure of our company name categorization. Both systems 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 88%. However, it should be noted that our result is based on all company names, including those which did not have any clear company name 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-a). An overview of DR-LINK and its approach to document filtering. Proceedings of the Human Language Technology Workshop. Princeton, NJ: March 1993.

Liddy, E.D., McVearry, K., Paik, W., Yu, E.S., & McKenna, M. (In press-b). Development, Implementation & Testing of a Discourse Model for Newspaper Texts. 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.

Paik, W., Liddy, E.D., Yu, E.S., & McKenna, M. (In press). Interpreting Proper Nouns for Information Retrieval. Proceedings of the Human Language Technology Workshop. Princeton, NJ: March 1993.

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