Proceedings of HLT-NAACL 2003
Main Papers , pp. 189-196
Edmonton, May-June 2003

Toward a Task-based Gold Standard for Evaluation of NP Chunks and Technical Terms

Nina Wacholder
Rutgers University
nina@scils.rutgers.edu

Peng Song
Rutgers University
psong@paul.rutgers.edu

Abstract

We propose a gold standard for evaluating two types of information extraction output -- noun phrase (NP) chunks (Abney 1991; Ramshaw and Marcus 1995) and technical terms (Justeson and Katz 1995; Daille 2000; Jacquemin 2002). The gold standard is built around the notion that since different semantic and syntactic variants of terms are arguably correct, a fully satisfactory assessment of the quality of the output must include task-based evaluation. We conducted an experiment that assessed subjects' choice of index terms in an information access task. Subjects showed significant preference for index terms that are longer, as measured by number of words, and more complex, as measured by number of prepositions. These terms, which were identified by a human indexer, serve as the gold standard. The experimental protocol is a reliable and rigorous method for evaluating the quality of a set of terms. An important advantage of this task-based evaluation is that a set of index terms which is different than the gold standard can 'win' by providing better information access than the gold standard itself does. And although the individual human subject experiments are time consuming, the experimental interface, test materials and data analysis programs are completely re-usable.

1 Introduction

The standard metrics for evaluation of the output of NLP systems are precision and recall. Given an arguably correct list of the units that a system would identify if it performed perfectly, there should in principle be no discrepancy between the units identified by a system and the units that are either useful in a particular application or are preferred by human beings for use in a particular task. But when the satisfactory output can take many different forms, as in summarization and generation, evaluation by precision and recall is not sufficient. In these cases, the challenge for system designers and users is to effectively distinguish between systems that provide generally satisfactory output and systems that do not.

NP chunks (Abney 1991; Ramshaw and Marcus 1995; Evans and Zhai 1996; Frantzi and Ananiadou 1996) and technical terms (Dagan and Church 1994; Justeson and Katz 1995; Daille 1996; Jacquemin 2001; Bourigault et al. 2002) fall into this difficult-to-assess category. NPs are recursive structures. For the maximal NP *large number of recent newspaper articles on biomedical science and clinical practice*, a full-fledged parser would legitimately identify (at least) seven NPs in addition to the maximal one: *large number*; *recent newspaper articles*; *large number of recent newspaper articles*; *biomedical science*; *clinical practice*; *biomedical science and clinical practice*; and *recent newspaper articles on biomedical science and clinical practice*. To evaluate the performance of a parser, NP chunks can usefully be evaluated by a gold standard; many systems (e.g., Ramshaw and Marcus 1995 and Cardie and Pierce 1988) use the Penn Treebank for this type of evaluation. But for most applications, output that lists a maximal NP and each of its component NPs is bulky and redundant. Even a system that achieves 100% precision and recall in identifying all of the NPs in a document needs criteria for determining which units to use in different contexts or applications.

Technical terms are a subset of NP chunks. Jacquemin (2001:3) defines *terms* as multi-word "vehicles of scientific and technical information".[1] The operational difficulty, of course, is to decide whether a specific term is a vehicle of scientific and technical information (e.g., *birth date* or *light truck*). Evaluation of mechanisms that filter out some terms while retaining others is subject to this difficulty. This is exactly the kind of case where context plays a significant role in deciding whether a term conforms to a definition and where experts disagree.

In this paper, we turn to an information access task in order to assess terms identified by different techniques. There are two basic types of information access mechanisms, searching and browsing. In searching, the user generates the search terms; in

[1] Jacquemin does not use the modifier *technical*.

browsing, the user recognizes potentially useful terms from a list of terms presented by the system. When an information seeker can readily think up a suitable term or linguistic expression to represent the information need, direct searching of text by user-generated terms is faster and more effective than browsing. However, when users do not know (or can't remember) the exact expression used in relevant documents, they necessarily struggle to find relevant information in full-text search systems. Experimental studies have repeatedly shown that information seekers use many different terms to describe the same concept and few of these terms are used frequently (Furnas et al. 1987; Saracevic et al. 1988; Bates et al. 1998). When information seekers are unable to figure out the term used to describe a concept in a relevant document, electronic indexes are required for successful information access.

NP chunks and technical terms have been proposed for use in this task (Boguraev and Kennedy 1997; Wacholder 1998). NP chunks and technical terms have also been used in phrase browsing and phrase hierarchies (Jones and Staveley 1999; Nevill-Manning et al. 1999; Witten et al. 1999; Lawrie and Croft 2000) and summarization (e.g., McKeown et al. 1999; Oakes and Paice 2001). In fact, the distinction between task-based evaluation of a system and precision/recall evaluation of the quality of system output is similar to the extrinsic/intrinsic evaluation of summarization (Gallier and Jones 1993).

In order to focus on the subjects' choice of index terms rather than on other aspects of the information access process, we asked subject to find answers to questions in a college level text book. Subjects used the Experimental Searching and Browsing Interface (ESBI) to browse a list of terms that were identified by different techniques and then merged. Subjects select an index term by clicking on it in order to hyperlink to the text itself. By design, ESBI forces the subjects to access the text indirectly, by searching and browsing the list of index terms, rather than by direct searching of the text.

Three sets of terms were used in the experiment: one set (HS) was identified using the head-sorting method of Wacholder (1998); the second set (TT) was identified by an implementation of the technical term algorithm of Justeson and Katz (1995); a third set (HUM) was created by a human indexer. The methods for identifying these terms will be discussed in greater detail below.

Somewhat to our surprise, subjects displayed a very strong preference for the index terms that were identified by the human indexer. Table 1 shows that when measured by percentage terms selected, subjects chose over 13% of the available human terms, but only 1.73% and 1.43% of the automatically selected terms; by this measure the subjects' preference for the human terms was more than 7 times greater than the preference for either of the automatic techniques. (In Table 1 and in the rest of this paper, all index term counts are by type rather than by token, unless otherwise indicated.)

	HUM	HS	TT
Total number of terms	**673**	**7980**	**1788**
Number of terms selected	**89**	**114**	**31**
Percentage of terms selected	**13.22%**	**1.43%**	**1.73%**

Table 1: Percentage of terms selected by human subjects relative to number of terms in the entire index.

This initial experiment strongly indicates that 1) people have a demonstrable preference for different types of index terms; 2) these human terms are a very good gold standard. If subjects use a greater proportion of the terms identified by a particular technique, the terms can be judged better than the terms identified by another technique, even if the terms are different. Any automatic technique capable of identifying terms that are preferred over these human terms would be a very strong system indeed. Furthermore, the properties of the terms preferred by the experimental subjects can be used to guide design of systems for identifying and selecting NP chunks and technical terms.

In the next section, we describe the design of the experiment and in Section 3, we report on what the experimental data shows about human preferences for different kinds of index terms.

2 Experimental design

Our experiment assesses the index terms vis a vis their usefulness in a strictly controlled information access task. Subjects responded to a set of questions whose answers were contained in a 350 page college-level text (Rice, Ronald E., McCreadie, Maureen and Chang, Shan-ju L. (2001) *Accessing and Browsing Information and Communication*. Cambridge, MA: MIT Press.) Subjects used the Experimental Searching and Browsing Interface (ESBI) which forces them to access text via the index terms; direct text searching was prohibited. 25 subjects participated in the experiment; they were undergraduate and graduate students at Rutgers University. The experiments were conducted by graduate students at the Rutgers University School of Communication, Information and Library Studies (SCILS).

2.1 ESBI (Experimental Searching and Browsing Interface)

Subjects used the Experimental Searching and Browsing Interface (ESBI) to find the answers to the questions. After an initial training session, ESBI presents the user with a Search/Browse screen (not shown); the question appears at the top of the screen. The subject may enter a string to search for in the index, or click on the "Browse" button for access to the whole index. At this point, "search" and "browse" apply only to the list of index terms, not to the text. The user may either browse the entire list of index terms or may enter a search term and specify criteria to select the subset of terms that will be returned. Most people begin with the latter option because the complete list of index terms is too long to be easily browsed. The user may select (click on) an index term to view a list of the contexts in which the term appears. If the context appears useful, the user may choose to view the term in its full context; if not, the user may either do additional browsing or start the process over again.

Figure 1 shows a screen shot of ESBI after the searcher has entered the string *democracy* in the search box. This view shows the demo question and the workspace for entering answers. The string was (previously) entered in the search box and all index terms that include the word *democracy* are displayed. Although it is not illustrated here, ESBI also permits substring searching and the option to specify case sensitivity.

Regardless of the technique by which the term was identified, terms are organized by grammatical head of the phrase. Preliminary analysis of our results has shown that most subjects like this analysis, which resembles standard organization of back-of-the-book indexes.

Readers may notice that the word *participation* appears at the left-most margin, where it represents the set of terms whose head is *participation*. The indented occurrence represents the individual term. Selecting the left-most occurrence brings up contexts for all phrases for which *participation* is a head. Selecting on the indented occurrence brings up contexts for the noun *participation* only when it is not part of a larger phrase. This is explained to subjects during the pre-experimental training and an experimenter is present to remind subjects of this distinction if a question arises during the experiment.

Readers may also notice that in Figure 1, one of the terms, *participation require*, is ungrammatical. This particular error was caused by a faulty part-of-speech tag. But since automatically identified index terms typically include some nonsensical terms, we have left these terms in – these terms are one of the problems that information seekers have to cope with in a realistic task-based evaluation.

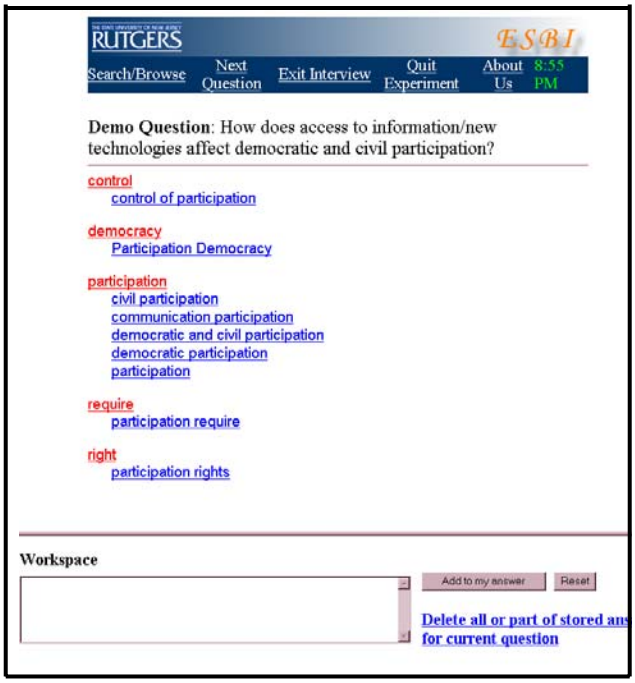

Figure 1: ESBI Screen shot

2.2 Questions

After conducting initial testing to find out what types of questions subjects founder hard or easy, we spent considerable effort to design a set of 26 questions of varying degrees of difficulty. To obtain an initial assessment of difficulty, one of the experimenters used ESBI to answer all of the questions and rate each question with regard to how difficult it was to answer using the ESBI system. For example, the question *What are the characteristics of Marchionini's model of browsing?* was rated very easy because searching on the string *marchionini* reveals an index term *Marchionini's* which is linked to the text sentence: *Marchionini's model of browsing considers five interactions among the information-seeking factors of "task, domain, setting, user characteristics and experience, and system content and interface"* (p.107). The question *What factors determine when users decide to stop browsing?* was rated very difficult because searching on *stop* (or synonyms such as *halt, cease, end, terminate, finish,* etc.) reveals no helpful index terms, while searching on *factors* or *browsing* yields an avalanche of over 500 terms, none with any obvious relevance.

After subjects finished answering each question, they were asked to rate the question in terms of its difficulty. A positive correlation between judgments

of the experimenters and the experimental subjects (Sharp et al., under submission) confirmed that we had successfully devised questions with a range of difficulty. In general, questions that included terms actually used in the index were judged easier; questions where the user had to devise the index terms were judged harder.

To avoid effects of user learning, questions were presented to subjects in random order; in the one hour experiment, subjects answered an average of about 9 questions.

2.3 Terms

Although the primary goal of this research is to point the way to improved techniques for automatic creation of index terms, we used human created terms to create a baseline. For the human index terms, we used the pre-existing back-of-the-book index, which we believe to be of high quality.[2]

The two techniques for automatic identification were the technical terms algorithm of Justeson and Katz (1995) and the head sorting method (Dagan and Church (1994); Wacholder (1998). In the implementation of the Justeson and Katz' algorithm, technical terms are multi-word NPs repeated above some threshold in a corpus; in the head sorting method, technical terms are identified by grouping noun phrases with a common head (e.g., *health-care workers* and *asbestos workers*), and selecting as terms those NPs whose heads appear in two or more phrases. Definitionally, technical terms are a proper subset of terms identified by Head Sorting. Differences in the implementations, especially the preprocessing module, result in there being some terms identified by Termer that were not identified by Head Sorting.

Table 2 shows the number of terms identified by each method. (*Because some terms are identified by more than one technique, the percentage adds up to more than 100%.) The fewest terms (673) were identified by the human method; in part this reflects the judgment of the indexer and in part it is a result of restrictions on index length in a printed text. The largest number of terms (7980) was identified by the head sorting method. This is because it applies looser criteria for determining a term than does the Justeson and Katz algorithm which imposes a very strict standard--no single word can be considered a term, and an NP must be repeated in full to be considered a term.

[2] Jim Snow prepared the index under the supervision of SCILS Professor James D. Anderson.

	HUM	HS	TT	Total
Total number of terms	673	7980	1788	9992
Percentage of total number of terms	6.73%	79.86%	17.89%	*

Table 2: Number of terms in index by method of identification

Wacholder et al. (2000) showed that when experimental subjects were asked to assess the usefulness of terms for an information access task without actually using the terms for information access showed that the terms identified by the technical term algorithm, which are considerably fewer than the terms identified by head sorting, were overall of higher quality than the terms identified by the head sorting method. However, the fact that subjects assigned a high rank to many of the terms identified by Head Sorting suggested that the technical term algorithm was failing to pick up many potentially useful index terms.

In preparation for the experiment, all index terms were merged into a single list and duplicates were removed, resulting in a list of nearly 10,000 index terms.

2.4 Tracking results

In the experiment, we logged the terms that subjects searched for (i.e., entered in a search box) and selected. In this paper, we report only on the terms that the subjects selected (i.e., clicked on). This is because if a subject entered a single word, or a subpart of a word in the search box, ESBI returned to them a list of index terms; the subject then selected a term to view the context in which it appears in the text. This term might have been the same term originally searched for or it might have been a superstring. The terms that subjects selected for searching are interesting in their own right, but are not analyzed here.

3 Results

At the outset of this experiment, we did not know whether it would be possible to discover differences in human preferences for terms in the information access task reported on in this paper. We therefore started our research with the null hypothesis that all index terms are created equal. If users selected index terms in roughly the same proportion as the terms

occur in the text, the null hypothesis would be proven.

The results strongly discredit the null hypothesis. Table 3 shows that when measured by percentage of terms selected, subjects selected on over 13.2% of the available human terms, but only 1.73% and 1.43% respectively of the automatically selected terms. Table 3 also shows that although the human index terms formed only 6% of the total number of index terms, 40% of the terms which were selected by subjects in order to view the context were identified by human indexing. Although 80% of the index terms were identified by head sorting, only 51% of the terms subjects chose to select had been identified by this method. (*Because of overlap of terms selected by different techniques, total is greater than 100%)

	HM	**HS**	**TT**	Total
All terms	673	7980	1788	9992
Percentage of all terms	6.73%	79.9%	17.9%	*
Total number of terms selected	89	114	31	223
Percentage of terms selected	39.9%	51.1%	13.9	*
Percentage of available terms selected	13.2%	1.43%	1.73%	

Table 3: Subject selection of index terms, by method.

To determine whether the numbers represent statistically significant evidence that the null hypothesis is wrong, we represent the null hypothesis ($H_{T)}$) as (1) and the falsification of the null hypothesis (H_A) as (2).

$$H_T: P_1/\mu_1 = P_2/\mu_2 \quad (1)$$

$$H_A: P_1/\mu_1 \neq P_2/\mu_2 \quad (2)$$

P_i is the expected percentage of the selected terms that are type *i* in all the selected terms; μ_i is the expected percentage if there is no user preference, i.e. the proportion of this term type *i* in all the terms. We rewrite the above as (3).

$$H_T: X = 0 \quad H_A: X \neq 0 \quad X = P_1/\mu_1 - P_2/\mu_2 \quad (3)$$

Assuming that X is normally distributed, we can use a one-sample t test on X to decide whether to accept the hypothesis (1). The two-tailed t test (df =222) produces a p-value of less than .01% for the comparison of the expected and selected proportions of a) human terms and head sorted terms and b) human terms and technical terms. In contrast, the p-value for the comparison of head-sorted and technical terms was 33.7%, so we draw no conclusions about relative preferences for head sorted and technical terms.

We also considered the possibility that our formulation of questions biased the terms that the subjects selected, perhaps because the words of the questions overlapped more with the terms selected by one of the methods. [3] We took the following steps:

1) For each search word, calculate the number of terms overlapping with it from each source.
2) Based on these numbers, determine the proportion of terms provided by each method.
3) Sum the proportions of all the search words.

As measured by the terms the subjects saw during browsing, 22% were human terms, 62% were head sorted terms and 16% were technical terms. Using the same reasoning about the null hypothesis as above, the p-value for the comparison of the ratios of human and head sorted terms was less than 0.01%, as was the comparison of the ratios of the human and technical terms. This supports the validity of the results of the initial test. In contrast, the p-value for the comparison of the two automatic techniques was 77.3%.

Why did the subjects demonstrate such a strong preference for the human terms? Table 4 illustrates some important differences between the human terms and the automatically identified terms. The terms selected on are longer, as measured in number of words, and more complex, as measured by number of prepositions per index terms and by number of content-bearing words. As shown in Table 5, the difference of these complexity measures between human terms and automatically identified terms are statistically significant.

Since longer terms are more specific than shorter terms (for example, *participation in a democracy* is longer and more specific than *democracy*), the results suggest that subjects prefer the more specific terms. If this result is upheld in future research, it has practical implications for the design of automatic term identification systems.

	Number of terms selected	Average length of term in words	Prepositions per index term	Content-bearing words per index term
HM	89	6.22	1.4	4.54
HS	114	2.59	0.026	2.23
TT	31	2.26	0	2.26

Table 4: Measures of index term complexity

	Average length of term in number of words	Number of prepositions per index term	Number of content-bearing words per index term
HM vs HS	<0.01%	<0.01%	<0.01%
HM vs TT	<0.01%	<0.01%	<0.01%
HS vs TT	0.57%	8.33%	77.8%

Table 5: Result of two-independent-sample two-tailed t-test on index term complexity. The numbers in the cells are p-value of the test.

4.3 Relationship between Term Source and Search Effectiveness

In this paper, our primary focus is on the question of what makes index terms 'better', as measured by user preferences in a question-answering task. Also of interest, of course, is what makes index terms 'better' in terms of how accurate the resulting users' answers are. The problem is that any facile judgment of free-text answer accuracy is bound to be arbitrary and potentially unreliable; we discuss this in detail in [26]. Nevertheless, we address the issue in a preliminary way in the current paper. We used an ad hoc set of canonical answers to score subjects' answers on a scale of 1 to 3, where 1 stands for 'very accurate', 2 stands for 'partly accurate' and 3 represents 'not at all accurate'. Using general loglinear regression (Poisson model) under the hypothesis that these two variables are independent of each other, our analysis showed that there is a systematic relationship (significance probability is 0.0504) between source of selected terms and answer accuracy. Specifically, in cases where subjects used more index terms identified by the human indexer, the answers were more accurate. On the basis of our initial accuracy judgments, we can therefore draw the preliminary conclusion that terms that were better in that they were preferred by the experimental subjects were also better in that they were associated with better answers. We plan to conduct a more in-depth analysis of answer accuracy and will report on it in future work.

But the primary question addressed in this paper is how to reliably assess NP chunks and technical terms. These results constitute experimental evidence that the index terms identified by the human indexer constitute a gold standard, at least for the text used in the experiment. Any set of index terms, regardless of the technique by which they were created or the criteria by they were selected, can be compared vis a vis their usefulness in the information access task.

4 Discussion

The contribution of this paper is the description of a task-based gold-standard method for evaluating the usefulness and therefore the quality of NP chunks and technical terms. In this section, we address a number of questions about this method.

1) **What properties of terms can this technique be used to study?**
   - **One word or many.** There are two parts to the process of identifying NP terms: NP chunks that are candidate terms must be identified and candidate terms must be filtered in order to select a subset appropriate for use in the intended application. Justeson and Katz (1995) is an example of an algorithm where the process used for identifying NP chunks is also the filtering process. A byproduct of this technique is that single-word terms are excluded. In part, this is because it is much harder to determine in context which single words actually qualify as terms. But dictionaries of technical terminology have many one-word terms.
   - **Simplex or complex NPs** (e.g., Church 1988; Hindle and Rooth 1991; Wacholder 1998) identify simplex or base NPs – NPs which do not have any component NPs -- at least in part because this bypasses the need to solve the quite difficult attachment problem, i.e., to determine which simpler NPs should be combined to output a more complex NP. But if people find complex NPs more useful than simpler ones, it is important to focus on improvement of techniques to reliably identify more complex terms.
   - **Semantic and syntactic terms variants.** Daille et al. (1996), Jacquemin (2001) and others address the question of how to identify semantic (synonymous) and syntactic variants. But independent of the question of how to recognize variants is the question of which variants are to be preferred for different kinds of uses.
   - **Impact of errors.** Real-world NLP systems have a measurable error rate. By conducting experiments in which terms with errors are include in the set of test terms, the impact of

these errors can be measured. The usefulness of a set of terms presumably is at least in part a function of the impact of the errors, whether the errors are a by-product of the algorithm or the implementation of the algorithm.

2) **Could the set of human index terms be used as a gold standard without conducting the human subject experiments?** This of course could be done, but then the terms are being evaluated by a fixed standard – by definition, no set of terms can do better than the gold standard. This experimental method leaves open the possibility that there is a set of terms that is better than the gold standard. In this case, of course, the gold standard would no longer be a gold standard -- perhaps we would have to call it a platinum standard.

3) **How reproducible is the experiment?** The experiment can be re-run with any set of terms deemed to be representative of the content of the Rice text. The preparation of the materials for additional texts is admittedly time-consuming. But over time a sizable corpus of experimental materials in different domains could be built up. These materials could be used for training as well as for testing.

4) **How extensible is the gold standard?** The experimental protocol will be validated only if equally useful index terms can be created for other texts. We anticipate that they can.

5) **How can this research help in the design of real world NLP systems?** This technique can help in assessing the relative usefulness of existing techniques for identifying terms. It is possible, for example, there already exist techniques for identifying terms that are superior to the two tested here. If we can find such systems, their algorithms should be preferred. If not, there remains a need for development of algorithms to identify single word terms and complex phrases.

6) **Do the benefits of this evaluation technique outweigh the costs?** Given the fundamental difficulty of evaluating NP chunks and technical terms, task-based evaluation is a promising supplement to evaluation by precision and recall. These relatively time-consuming human subject experiments surely will not be undertaken by most system developers; ideally, they should be performed by neutral parties who do not have a stake in the outcome.

7) **Should automated indexes try to imitate human indexers?** Automated indexes should contain terms that are most easily processed by users. If the properties of such terms can be reliably discovered, developers of systems that identify terms intended to be processed by people surely should pay attention.

5 Conclusion

In this paper we have reported on a rigorous experimental technique for black-box evaluation of the usefulness of NP chunks and technical terms in an information access task. Our experiment shows that it is possible to reliably identify human preferences for sets of terms.

The set of human terms created for use in a back-of-the-book index serves as a gold standard. An advantage of the task-based evaluation is that a set of terms could outperform the gold standard; any system that could do this would be a good system indeed.

The two automatic methods that we evaluated performed much less well than the terms created by the human indexer; we plan to evaluate additional techniques for term identification in the hope of identifying automatic methods that identify index terms that people prefer over the human terms. We also plan to prepare test materials in different domains, and assess in greater depth the properties of the terms that our experimental subjects preferred; our goal is to develop practical guidelines for the identification and selection of technical terms that are optimal for human users. We will also study the impact of semantic differences between terms on user preferences and investigate whether terms which are preferred for information access are equally suitable for other NLP tasks.

6 Acknowledgements

We are grateful to the other members of the Rutgers NLP-I research group, Lu Liu, Mark Sharp, and Xiaojun Yuan, for their valuable contribution to this project. We also thank Paul Kantor, Judith L. Klavans, Evelyne Tzoukermann , Min Yen Kan, and three anonymous reviewers for their helpful suggestions. Funding for this research has been provided by the Rutgers University Information Science and Technology Council.

References

Abney, Steven (1991) Parsing by chunks. *Principle-Based Parsing*, edited by Steven Abney, Robert Berwick and Carol Tenny. Kluwer: Dordrecht.

Bates, Marcia J. (1998) Indexing and access for digital libraries and the Internet: human, database and domain factors. *Journal of the American Society for Information Science*, 49(13), 1185-1205.

Boguraev, Branimir and Kennedy, Christopher (1997) Salience-based content characterization of text. *ACL EACL Workshop on Intelligent Scalable Text Summarization*, 2-9.

Bourigault, Didier, Jacquemin, Christian and L'Homme, Marie Claude (2001) *Recent Advances in Computational Terminology*. John Benjamins: Philadelphia, PA.

Church, Kenneth Ward (1988) A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. *Proceedings of Second Applied Natural Language Processing Conference*, pp.136-143.

Dagan, Ido and Church, Kenneth (1994) TERMIGHT: Identifying and translating technical terminology. *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing*, pp.34-40.

Daille Beatrice (1996) Study and implementation of combined techniques for automatic extraction of terminology. *The Balancing Act*, pp.49-66. Edited by Judith L. Klavans and Philip Resnik. MIT Press, Cambridge, MA.

Daille, Beatrice, Habert, Benoit., Jacquemin, Christian, & Royaute, Jean (2000) Empirical observation of term variations and principles for their description. *Terminology*, 3(2):197-258.

Furnas, George, Landauer, Thomas, Gomez, Louis & Dumais, Susan T. (1987) The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11), 964-971.

Galliers, Julia Rose & Jones, Karen Sparck (1995) Evaluating natural language processing systems. Lecture Notes in Artificial Intelligence. Springer, New York, 1995.

Jacquemin, Christian (2001). Spotting and Discovering Terms through Natural Language Processing. Cambridge, MA: MIT Press.

Jones, Steve and Staveley, Mark S. (1999) Phrasier: a system for interactive document retrieval using keyphrases. *Proceedings of the 22nd annual international ACM SIGIR conference*, pp.160-167.

Justeson, John S. & Slava M. Katz (1995) "Technical terminology: some linguistic properties and an algorithm for identification in text", *Natural Language Engineering* 1(1):9-27.

Hindle, Donald and Rooth, Matt (1993) Structural ambiguity and lexical relations. *Computational Linguistics* 19(1):103-120.

Lawrie, Dawn and Croft, W. Bruce (2000) Discovering and comparing topic hierarchies. *Proceedings of RIAO 2000 Conference*, 314-330.

McKeown, Kathy, Klavans, Judith, Hatzivassiloglou, Vasileios, Barzilay, Regina and Eskin, Eleazar (1999) Towards multidocument summarization by reformulation: Progress and prospects. *Proceedings of AAAI-99*, pp.453-460.

Nevill-Manning, Craig, Witten, Ian and Paynter, Gordon W. (1999). Lexically-generated subject hierarchies for browsing large collections. *Int'l Journal on Digital Libraries*, *2*(2-3):111-123.

Oakes, Michael P. and Paice, Chris D. (2001) *Term extraction for automatic abstracting*. In Bourigault et al., eds.

Ramshaw, Lance A., and Marcus, Mitchell P. (1995) Text chunking using transformation-based learning. *Proceedings of the Third ACL Workshop on Very Large Corpora*, pp. 82-94.

Rice, Ronald E., Maureen McCreadie & Shan-ju L. Chang (2001). Accessing and Browsing Information and Communication. Cambridge, MA: MIT Press.

Saracevic, Tefko, Paul Kantor, Alice Y. Chamis & Donna Trivison (1988) A study of information seeking and retrieving: I. Background and methodology. Journal of the American Society for Information Science, 39(3), 161-176.

Sharp, M., Liu, L., Yuan, X., Song, P., & Wacholder, N. (2003). Question difficulty effects on question answering involving mandatory use of a term index. Under submission.

Wacholder, N., Sharp, M., Liu, L., Yuan, X., & Song, P. (2003). Experimental study of index terms and information access. Under submission.

Wacholder, Nina (1998) "Simplex noun phrases clustered by head: a method for identifying significant topics in a document", *Proceedings of Workshop on the Computational Treatment of Nominals*, pp.70-79. COLING-ACL, October 16, 1998.

Wacholder, Nina, Judith L. Klavans and David Kirk Evans (2000) "Evaluation of automatically identified index terms for browsing electronic documents", *Proceedings of the NAACL/ANLP2000, Seattle, Washington*.

Witten, Ian H., Paynter, Gordon W., Eibe, Frank, Gutwin, and Nevill-Manning Craig G. KEA: practical automatic keyphrase extraction. *Proceedings of the fourth ACM Conference on Digital Libraries*, pp.254-255.