Evaluation of Computational Linguistic Techniques for Identifying Significant Topics for Browsing Applications

Judith L. Klavans, Nina Wacholder, and David K. Evans
Columbia University
Department of Computer Science
and
Center for Research on Information Acess
500 W. 120th Street
New York, NY, 10027, USA
{klavans, nina, devans} @cs.columbia.edu

0. Abstract

Evaluation of natural language processing tools and systems must focus on two complementary aspects: first, evaluation of the accuracy of the output, and second, evaluation of the functionality of the output as embedded in an application. This paper presents evaluations of two aspects of LinkIT, a tool for noun phrase identification linking, sorting and filtering. LinkIT [Evans 1998] uses a head sorting method [Wacholder 1998] to organize and rank simplex noun phrases (SNPs). LinkIT is to identify significant topics in domain-independent documents. The first evaluation, reported in D.K.Evans et al. 2000 compares the output of the Noun Phrase finder in LinkIT to two other systems. Issues of establishing a gold standard and criteria for matching are discussed. The second evaluation directly concerns the construction of the browsing application. We present results from Wacholder et al. 2000 on a qualitative evaluation which compares three shallow processing methods for extracting index terms, i.e., terms that can be used to model the content of documents. We analyze both quality and coverage. We discuss how experimental results such as these guide the building of an effective browsing applications.

1. Goals and Methodology

The goal of this project is to evaluate different approaches to the task of automatically identifying index terms that have been derived without recourse to lexicons or to other kinds of domain-specific information. An index term is defined as a word or phrase that reflects a meaningful representation of a document for use by people. Such index terms are found in abstracts, in library subject codes, and in back of the book indices. A list of index terms for a particular document should represent the content accurately and should thoroughly capture the major topics.

This paper presents the evaluation of two phases of the NP identification and linking system, called LinkIT. The first evaluation focuses on NP identification; the second on the quality of output from LinkIT and from two other approaches to extracting terms from documents. We then discuss the implications of our experimental results for the design of browsing applications such as automatic gisting and an indexer's aid.

2. Phase One - NP Evaluation

In D.K. Evans et al. 2000, we report on our evaluation results for the NP identification task. For this evaluation, we designed an experiment to test LinkIT's performance at NP identification as compared to other NP identifiers. The task consists of identifying the NPs in a test collection of 14 documents. In this experiment, LinkIT's additional capabilities of lexical chain identification and noun group ranking were not evaluated.

Simplex NPs (SNPs) are identified by a system which sequentially parses text that has been tagged with part of speech using a finite state machine. Next, the complete list of SNPs identified in a document is sorted by the head of the phrase, which, at least for English-language common SNPs, is almost always the last word. The intuitive justification for sorting SNPs by head is based on the fundamental linguistic distinction between head and modifier: in general, a head makes a greater contri-

bution to the syntax and semantics of a phrase than does a modifier. If, as a practical matter, it is necessary to rank the contribution to a whole document made by the sequence of words constituting an NP, the head should be ranked more highly than other words in the phrase. This distinction is important in linguistic theory; for example, [Jackendoff 1977] discusses the relationship of heads and modifiers in phrase structure. It is also important in NLP, where, for example, [Strzalkowski 1997] and [Evans and Zhai 1996] have used the distinction between heads and modifiers to add query terms to information retrieval systems.

The data set consisted of NPs from documents wsj_0300 - wsj_0314 of the Penn Wall Street Journal Treebank [Marcus, Santorini, & Marcinkiewicz 1993]. The noun phrases were extracted from the parsed data files of the Treebank. An automatic process was used to extract the smallest unit marked as an NP in the Treebank, and each resulting file was examined to verify the correctness of the NPs extracted. In certain cases, complex noun phrases were manually split into smaller units; for example, NPs that contained a conjunction were split if we judged that there was ambiguity regarding the applicability of the head of the NP to each constituent of the phrase. Each system was tested over the plain text files corresponding to the parsed data files for the test noun phrases. For the initial evaluation, we compared output of the LinkIT system to output from the text chunking tool of [Ramshaw & Marcus, 1995]. The Penn chunker applies the transformation-based learning technique [Brill 1993] to the chunking task.

A human judge rated the acceptability of each NP in the system's output by assigning it to one of six categories representing the relationship between the NP in the gold standard set and the NP in the system output. The evaluation of NP identification is a difficult task since definitions of NPs vary. In this particular evaluation we defined six different classes for characterizing the relationship between an NP in the test set and an NP in the evaluation set but because we are forced to assign relationships between NPs to one of these six categories, we lose information. The categories for evaluation were:

- **Correct** - A perfect match of the two NPs.
- **Missing** - A NP in the gold standard is completely missing from the test set.
- **Under-generated** - A NP in the system output partially matches a NP in the gold standard set, but the words in the NP in the test set are a proper subset of the words in the gold standard NP.
- **Over-generated** - The words in the gold standard NP are a proper subset of the words in the NP in the test set.
- **Mismatch** - There is some overlap between the two NPs but neither is a proper subset of the other. In this case the test set NP contains some word(s) not in the gold standard NP and the gold standard NP contains some word(s) not in the test set NP.
- **False positive** - A NP is not in the gold standard set at all - it is a false positive.

The UPenn Chunker did not appear to perform as well as LinkIT in the test reported in this paper. LinkIT's precision was 79% and the recall 83%, in comparison to 67% recall and 74% precision for the UPenn Chunker. In an independent study, Ramshaw and Marcus report a recall and precision of 93% for base NP chunks trained on a much larger test set (950K words). We can only conclude that the discrepancy is due to the difference in what counts as an NP; we plan to investigate this problem further.

We also performed an evaluation on the Arizona Noun Phraser [Tolle & Chen 2000; Tolle 1997]. It must be stressed that the Arizona Noun Phraser is targeted for an IR task, and as such employs a definition of NPs that is more suited to that domain. However, bearing this and the stringent nature of our evaluation in mind, the Arizona Noun Phraser was achieved recall of 61% and precision of 66%. In the case of the Arizona Noun Phraser, many NPs tested fell into the mismatched NP category, when a more expressive set of relationships might not have penalized it. For example, for the two sequential NPs "a man" and "extraor-

dinary qualities", the Arizona Noun Phraser generated the NP "man with extraordinary qualities". Had it generated the NP "a man with extraordinary qualities" it could be assigned to the over-generation category twice. Since the Arizona Noun Phraser did not include the "a", we were forced to assign the NP "man" to the mismatch category since it contained the "a" from the NP "a man" and the "extraordinary qualities" NP from the following noun phrase.

We concluded that the NP identification in LinkIT was comparable, and at times superior, to other NP identifiers, but that final comparisons would need to be redone based on an agreed upon criterion for the gold standard. Each system has a different interpretation of the notion NP; thus, we informally estimated that such direct comparisons may include an error margin of up to 20%.

3. Phase Two – Comparison of Three Term Identification Systems

In later research [Wacholder et al. 2000], we compared three shallow processing methods for identifying index terms:

- **Keywords (KW)** are terms identified by counting frequency of stemmed words in a document;
- **Technical terms (TT)** are noun phrases (NPs) or subparts of NPs repeated more than twice in a document [Justeson and Katz 1995];
- **Head sorted terms (HS)** are identified by a method in which simplex noun phrases (as defined below) are sorted by head and then ranked in decreasing order of frequency [Wacholder 1998].

Each of these methods, use statistical and/or linguistic properties that apply to any natural language document in any field, and thus are domain-independent. They are also corpus-independent, in that the terms are ranked with respect to each document, without regard to properties of the corpus.

The standard IR technique known as tf*idf [Salton 1989] seeks to identify documents relevant to a particular query by relativizing keyword frequency in a document as compared to frequency in a corpus. This method can be used to locate at least some important concepts in full text. Although it has been effective for information retrieval, for other applications, such as human-oriented indexing, this technique is impractical. Ambiguity of stems (*trad* might refer to *trader* or *tradition*) and of isolated words (*state* might be a political entity or a mode of being) means that lists of keywords have not usually been used to represent the content of a document to human beings. Furthermore, humans have a difficult time processing stems and parts of words out of phrasal context.

For this study, we used the SMART system [Salton 1989] to identify stemmed keywords. We used the technical term finder of Justeson and Katz 1995. Wacholder 1998 proposed the method of Head Sorting (HS) for identifying significant topics that can be used to represent a source document. HS also uses a frequency measure to provide an approximation of topic significance. However, instead of counting frequency of stems or repetition of word sequences, this method counts frequency of a relatively easily identified grammatical element, heads of simplex noun phrases (SNPs). For common NPs (NPs whose head is a common noun), an SNP is a maximal NP that includes premodifiers such as determiners and possessives but not post-nominal constituents such as prepositions or relativizers. For example, *the well-known book* is an SNP but *the well-known book on asteroids* includes two SNPs, *well-known book* and *asteroids*. For proper names, an SNP is a name that refers to a single entity. For example, *Museum of the City of New York*, the name of an organization, is an SNP even though the organizational name incorporates a city name. Others, such as [Church 1988, Strzalkowski 1997], have discussed a similar concept, sometimes called simple or base NPs. We chose to evaluate methods that depend only on document-internal data, independent of corpus, domain or genre. We therefore did not use, for example, tf*idf, the purely statistical technique that is the used by most information retrieval systems, or a hybrid statistical and symbolic

technique for identifying collocations [Smadja 1993].

To compare performance, two groups evaluated the output of the three approaches: professionals and students. Professionals included librarians and publishing professionals familiar with both manual and automatic text indexing. Students included undergraduate and graduate students with a variety of academic interests. Subjects were presented with an article and a list of terms identified by one of the three methods. They were asked to answer the question: "Would this term be useful in an electronic index for this article?" A 1 to 5 rating scale was used, where 1 indicates a high quality term should to be included in the index and 5 indicates a junk term to not be included. For example, the phrase *court-approved affirmative action plans* received an average rating of 1 from the professionals, meaning that it was ranked as useful for the article; the KW *affirmative* received an average rating of 3.75, meaning that it was less useful; and the KW *action* received an average ranking of 4.5, meaning that it was not useful.

Results were measured in terms of two criteria: **quality** and **coverage**. Quality refers to ranking terms high on the 1 to 5 scale from highest to lowest. Coverage refers to the thoroughness with which the terms cover the significant topics in the document. Results showed that TTs are superior with respect to quality; however, there are only a small number of TTs per document, so they do not provide adequate coverage in that they are not fully representative of the document as a whole. In contrast, KWs provide good coverage but relatively poor quality in that KWs are vague, and not well filtered. SNPs, which have been sorted using HS and filtered, provide a better balance of quality and coverage.
To sum, our second evaluation showed that:

- The **KW** approach identifies some useful index terms, but they are mixed in with a large number of low-ranked terms.
- The **TT** approach identifies high quality terms, but with low coverage, i.e., relatively few indexing terms.
- The **HS** approach achieves a balance between quality and coverage.

Examples of output of these three different techniques can be found in the Appendix. Our results for the three types of terms, by document, are shown in Figure 1. All results were included, even partial ratings.

Doc	Avg KW rating	Avg TT rating	Avg HS rating
900405-0109	3.08	1.45	2.71
900516-0043	3.73	2.19	2.71
900517-0062	2.98	1.7	3.25
Avg of Avgs	**3.27**	**1.79**	**2.89**

Figure 1: Average ratings of 3 types of index terms

TTs received the highest ratings for all three documents—an average of 1.79 on the scale of 1 to 5, with 1 being the best rating. HS came in second, with an average of 2.89, and KW came in last with an average of 3.27. Note that averaging conceals the fact that the number of TTs is much lower than the other two types of terms, as shown in Figure 1. Figure 2 shows the total number of terms rated at or below specified rankings, allows us to measure quality and coverage. (1 is the highest rating; 5 is the lowest.) This figure shows that the HS method identifies more high quality terms than the TT method does.

	Number of terms ranked at or better than			
Method	**2**	**3**	**4**	**5**
KW	27	75	124	166
HS	41	96	132	160
TT	15	21	21	21

Figure 2: Running total of terms identified at or below a specified rank

TT clearly identifies the highest quality terms: 100% of TTs receive a rating of 2 or better. However, only 8 TTs received a rating of 2 or better (38% of the total), while 41 HSs received a rating of 2 or better (26% of the total). This indicates that the TT method misses many high quality terms. KW, the least discriminating method in terms of quality, also provides better coverage than does TT.

This result is consistent with our observation that TT identifies the highest quality terms, but there are very few of them: an average of 7 per 500 words compared to over 50 for HS and KW. Therefore there is a need for additional high quality terms. The list of HSs received a higher average rating than did the list of KWs, as shown in Figure 2. This is consistent with our expectation that phrases containing more content-bearing modifiers would be perceived as more useful index terms than would single word phrases consisting only of heads.

4. Next Steps – Incorporating Results into Applications

Our results show that single words in isolation are judged differently than the same word when presented in the context of a larger phrase. This finding alone has implications in the design of indexing tools since the way that index terms are presented to the human indexer in a browsing tool will affect its usefulness as much as the term itself. We plan on building several tools, including an indexer's aid for scholarly publishing and a gisting tool for browsing which incorporate our results.

We have performed a qualitative evaluation of three techniques for identifying significant terms in a document, driven by an indexing task. Our results show that the head sorting technique outperforms two other indexing methods, technical terms and keywords, as measured by balance of quality and coverage. We have used human judges to evaluate the effectiveness of each method. This research is a contribution to the overall evaluation of computational linguistic tools in terms of their usefulness for human-oriented computational applications, such as the creation of a profile or thumbnail of a document. In future research, we will utilize these results to combine techniques for building effective browsing tools.

5. Acknowledgements

This work was supported by NSF Grant IRI-97-12069, as part of the Information and Data Management Project (http://www.columbia.edu/cu/cria/SigTops/) and also by NSF Grant CDA-97-53054.

6. References

Boguraev, Branimir and Kennedy, Christopher (1998) "Applications of term identification terminology: domain description and content characterisation", *Natural Language Engineering* 1(1):1-28.

Brill, Eric (1993). Automatic grammar induc tion and parsing free text: A transformation-based approach. In *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 237-242.

Church, Kenneth Ward (1988) "A stochastic parts program and noun phrase parser for unrestricted text", in *Proceedings of the Second Conference on Applied Natural Language Processing*, pp. 136-143.

Evans, David A. and Chengxiang Zhai (1996) "Noun-phrase analysis in unrestricted text for information retrieval", *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp.17-24. 24-27.

Evans, David Kirk (1998) LinkIT Documentation, Columbia University Department of Computer Science Report.

Evans, David Kirk, Judith Klavans, and Nina Wacholder (2000) "Document processing with LinkIT", *Proceedings of RIAO (Recherches d'Informations Assistee Par Ordinateur = Computer-Assisted Information Retrieval) Conference*, Paris, France.

Jackendoff, Ray (1977) X-bar Syntax: A Study of Phrase Structure, MIT Press, Cambridge, MA.

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

Marcus M. P., B. Santorini, and M. A. Marcinkiewicz, (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* (19).

Ramshaw, Lance A. and Mitchell P. Marcus (1995). Text Chunking Using Transformation-Based Learning. *Proceedings of the Third Association for Computational Linguistics Workshop on Very Large Corpora*.

Salton, Gerald (1989) Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley, Reading, MA.

Smadja, Frank (1993) "Retrieving collocations from text", *Computational Linguistics* 19(1): 143-177.

Strzalkowski, Tomek (1997) "Building effective queries in natural language information retrieval", *Proceedings of the Applied Natural Language Processing Conference of the Association for Computational Linguistics*, Washington, DC, pp.299-306.

Tolle, Kristin M. and Hsinchun Chen (2000). Comparing Noun Phrasing Techniques for Use with Medical Digital Library Tools. *Journal of the American Society for Information Science Association* 51(4): 352-370.

Tolle, Kristin M. (1997). Improving Concept Extraction from Text Using Noun Phrasing Tools: An Experiment in Medical Information Retrieval. Master Thesis. University of Arizona, Department of Management Information Systems.

Wacholder, Nina (1998) "Simplex NPS sorted by head: a method for identifying significant topics within a document", *Proceedings of the Workshop on the Computational Treatment of Nominals*, COLING-ACL '98, Montreal, Canada. pp. 70-79.

Wacholder, Nina, David Kirk Evans, Judith L. Klavans (2000) "Evaluation of Automatically Identfied Index Terms for Browsing Electronic Documents", *Proceedings of the Applied Natural Language Processing and North American Chapter of the Association for Computational Linguistics (ANLP-NAACL) 2000*. Seattle, Washington.

0. Appendix A: Terms identified in WSJ900516-0043

0.1 Head Sorted

workers
maintenance workers
flour mill workers
other flour mill workers
many flour mill workers
elevator workers
mills
flour mill(s)
cancer(s)
members
union members
former members
researchers
NCI researchers
chemical(s)
department
maintenance department(s)
study
Federal Study
preliminary study
recent study
years
eight-fold higher risk
almost three-fold higher risk
fumigants
chemical fumigants
journal
Wall Street Journal
semi-monthly Journal
deaths
fewer deaths
24 lymphoma deaths
lymphoma deaths
Health
cells
white blood cells
bins
grain bins
union
American Federation of Grain Millers union
system
lymph system(s)

lymphoma
non-Hodgkins lymphoma
pesticides
handlers
grain handlers
new finding
not unexpected finding
suspicion(s)
grain
elevators
grain storage elevators
grain elevators

0.2 Keywords

mill/mills
worked/working
flour
grain
cancer/cancers/Cancer
chemical/chemicals
lymphoma
risk
researchers
members
fumigate/fumigants
years
study
increased
high
found
department/departments
death/deaths
white
union
pesticide/pesticides
maintenance
lymph
Journal
handled/handlers
elevator/elevators
dying
blood
Wall
unexpected/unexpectedly
system/systems
suspicions
Street
storage/stored
responsible
reporter/reported
phosphine
occupational
noted
NCI
nation/National
Hodgkins
health/Health
fold
finding
Federation/Federal/federation
exposure/exposures/exposed
cells
bins
applied
American

0.3 Technical Terms

flour mill
chemical fumigants
grain bin
grain handler
white blood cell
lymphoma death
lymph system
maintenance department
higher risk