Proceedings of the IASTED International Conference
Knowledge Sharing and Collaborative Engineering
November 22-24, 2004, St. Thomas, US Virgin Islands

Using Noun Phrase Centrality to Identify Topics for Extraction based Summaries

Zhuli Xie, Peter C. Nelson
Department of Computer Science
University of Illinois at Chicago
Chicago IL 60607, U.S.A.
{zxie, nelson}@cs.uic.edu

Weimin Xiao, Thomas M. Tirpak
Physical Realization Research Center
Motorola Labs
Schaumburg IL 60196, U.S.A.
{awx003, T.Tirpak}@motorola.com

ABSTRACT

In this paper, we use a Social Network Analysis method and decision tree analysis to study the distribution and relationship of Noun Phrases in documents and their corresponding abstracts. Initial results have shown significant improvement in extraction based text summarization by applying systematic predictions of the Noun Phrases that appear in both the documents and in their corresponding abstracts.

KEY WORDS
Text Mining, Information Retrieval, and Text Summarization

1. Introduction

Vast amount of new text documents appear every day in almost every scientific field. Researchers often depend on Information Retrieval (IR) techniques to obtain relevant documents for their needs. However, collection of tens, hundreds, or even thousands of documents returned from an IR tool may pose an insurmountable obstacle to finding the right information in timely manner. Automatic text summarization plays an important role in reducing the workload of digesting the content in the returned documents.

In our previous study of text summarization, we established a framework that uses machine learning to discover the underlying summarization mechanism [1]. We developed a prototype system based on this framework to produce **extractive summaries**, i.e., selecting a certain number of sentences from a given text document to form the summary. Text features are typically the focus in the automatic text summarization literature [2][3][4][5][6][7]. Superficial features, such as term frequency, sentence location, length of the sentence, etc., have been widely studied, and can easily be obtained and processed statistically. Using these conventional features, our prototype system generated fairly good summaries. However, those features did not capture contextual information from the text. Thus, one should not expect that the system will always produce a summary covering the essential topics of a document.

For certain types of summaries, a topic list may indicate the content of a document [5][8]. Topics are usually characterized in the form of noun phrases. Thus, for a given text document, if we can find some generic rules to identify the topics (noun phrases) that appear in the document that are the most likely to appear in its summary, the quality of the extractive summarization can be improved. This paper presents this approach and discusses a new text feature — Noun Phrase Centrality which our experiments have shown to be quite useful.

2. Noun Phrase Centrality

The study of organizational communication by [9] introduced centrality, a technique that was originally used in Social Network Analysis (SNA), i.e., the study of certain social relationships among a group of actors. An actor can be an individual or an organization in social networks. Each actor has an important property — centrality indicating the degree to which the other actors in the network revolve around him/her. Actors with high centrality are more prominent or important in the network. We have noticed that in a text document, a topic or con-

cept is often introduced in the form of noun phrase (NP). Centering this topic or concept, the subsequent sentences or paragraphs further develop it, indicating the reasons that the author brings up the topic. For example, in the introduction part of one of our sample documents (Figure 1), a concept－synchronous tree-adjoining grammars (TAG)－is introduced in the first sentence. In the second sentence, the meaning of this concept is explained. Following that, in the second paragraph, the author explained the intension of introducing synchronous TAGs, and in the third paragraph the author further introduced his work with this technique.

Such phenomena were also studied by [10], proposing a Centering Theory that models the "coherence of utterances within a discourse segment".

Introduction

The formalism of synchronous tree-adjoining grammars, a variant of standard tree-adjoining grammars (TAG), was intended to allow the use of TAGs for language transduction in addition to language specification. Synchronous TAGs specify relations between language pairs; each language is specified with a standard TAG, and the pairing between strings in the separate languages is specified by synchronizing the two TAGs through linking pairs of elementary trees.

This paper concerns the formal definitions underlying synchronous tree- adjoining grammars. In previous work, the definition of the transduction relation defined by a synchronous TAG…First, the weak-generative expressivity of TAGs is increased... Second, the lack of a simple recursive characterization of the derivation … makes the design of parsing algorithms difficult if not impossible.

In this paper, we describe how synchronous TAG derivation can be redefined so as to eliminate these problems. The redefinition relies … Furthermore,… However,…, some linguistic analyses may no longer be statable. We comment on some possible negative ramifications of this fact.

Figure 1. An example of topics/concepts introduced by a noun phrase

Not all noun phrases refer to some topics or concepts. A noun phrase may specifically refer to a person, an organization, an object, an attribute, a state, or a number, etc. In our experiments, a text document can contain hundreds of noun phrases. Determining which noun phrases are actually important topics in a given text will be meaningful in many research areas in natural language processing, such as document retrieval, classification, and automatic text summarization. Inspired by the work of [9], we have established a network for all noun phrases in a given text with each noun phrase represented by a node. It is assumed that important topics presented in the text are the prominent nodes within the network, and then, the idea of Centrality from SNA can be used to measure the prominence of the noun phrases in the text.

3. Formation of a noun phrase network

We process a text document in four steps. First, the text is tokenized and stored into an internal representation with structural information. Second, the tokenized text is tagged through a Brill tagging algorithm [11] POS tagger[11]. Third, the noun phrases in each sentence are parsed according to 35 parsing rules as shown in Figure 2. If a new noun phrase is found in the sentence, a new node is formed and added to the network. If the noun phrase already exists in the network, the node containing it will be identified, and a link will be added between two nodes if the nodes are parsed out sequentially in the same sentence. Finally, after all sentences have been processed, the centrality of each node in the network is updated. The processes of forming a noun phrase network are shown in Figure 3.

4. Abstract Noun Phrase Distribution Analysis

We refer to the noun phrases that appear in the abstract of a document as "abstract noun phrases". The "body text" refers to the text of a document excluding the abstract. As our goal is to predict which noun phrases from the body text will most likely appear in the abstract, a natural step is to find relations between all abstract noun phrases and

[1] The POS tagger we used can be obtained from http://web.media.mit.edu/~hugo/montytagger/

noun phrases from the body text. If this part of NPs is too small, it may not significantly affect the abstraction, since the abstract may contain many topics or concepts. Fortunately, as we will show in Section 4.2., the NPs from body text play a significant role in the abstract NPs.

4.1. CMP-LG Corpus

In our experiment, a corpus of 183 documents was used. The documents are from the Computation and Language collection and have been marked in XML with tags providing basic information about the document such as title, author, abstract, body, sections, etc. This corpus is a part of the TIPSTER Text Summarization Evaluation Conference (SUMMAC) effort acting as a general resource to the information retrieval, extraction and summarization communities. We excluded the five documents from this corpus which did not have abstracts.

4.2. Experiment Results

We analyzed how many abstract NPs can be found in the body text for the 178 documents. On average, 68% of noun phrases in the abstract can be found in other sections of the article. In Figure 4, we can see that almost 90% of the documents had over 50% of the abstract NPs coming from the body text. In the next section, we will discuss how we try to find relationships between the abstract noun phrases and body text noun phrases and use these relationships to predict the abstract NPs.

NX --> CD	NX --> NNS
NX --> CD NNS	NX --> PRP
NX --> NN	NX --> WP$ NNS
NX --> NN NN	NX --> WDT
NX --> NN NNS	NX --> EX
NX --> NN NNS NN	NX --> WP
NX --> NNP	NX --> DT JJ NN
NX --> NNP CD	NX --> DT CD NNS
NX --> NNP NNP	NX --> DT VBG NN
NX --> NNP NNPS	NX --> DT NNS
NX --> NNP NN	NX --> DT NN
NX --> NNP NNP NNP	NX --> DT NN NN
NX --> JJ NN	NX --> DT NNP
NX --> JJ NNS	NX --> DT NNP NN
NX --> JJ NN NNS	NX --> DT NNP NNP
NX --> PRP$ NNS	NX --> DT NNP NNP NNP
NX --> PRP$ NN	NX -->DT NNP NNP NN NN
NX --> PRP$ NN NN	

Figure 2. NP Parsing Rules

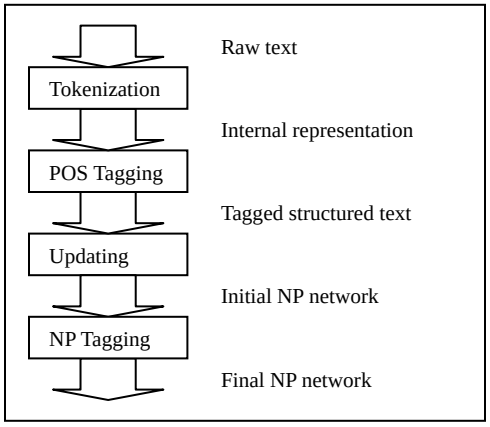
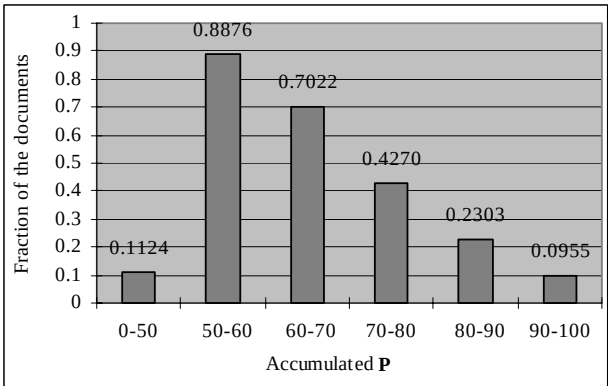


Figure 3. NP network formulation chart

5. Predicting Abstract Noun Phrases

5.1. Using the Noun Phrase Centrality Heuristic

As we discussed in Section 2, if a noun phrase (denoted as NP1 here) refers to a topic addressed in an article, it is very likely that the NP1 will appear in the article repeatedly. In the procedure of forming the NP network for the article, the node containing the NP1 will establish many links to other nodes. Thus, its centrality will be relatively higher than other peripheral nodes in which the NPs refer to less important topics or just specific objects. Given this heuristic, we performed an experiment, in which the nodes with highest centralities are retrieved, and the NPs contained in them are compared with the actual abstract NPs, on the CMP-LG corpus. To evaluate this method, we intended to use Precision, which measures the



(**P**:The percentage of the abstract NPs from the body text in all the abstract NPs)

Figure 4. Abstract NPs from body text analysis result

fraction of correctly predicted abstract NPs of all the predicted NPs, and Recall, which measures the fraction of correctly predicted abstract NPs in all Common NPs[2].

After establishing the NP network for the article and ranking the nodes according to their centralities, we must decide how many nodes should be retrieved. This number should not be too big; otherwise the Precision value will be very low, although the Recall will be higher. If this number is very small, the Recall will decrease correspondingly. We adopted a compound metric — F-measure, to balance the node selection:

F-measure=2*Precision*Recall/(Precision+Recall)

Based on our study of 178 documents in the CMP-LG corpus, we tried to find whether the best number of nodes to be retrieved has any relation to the number of NPs in the body text, or the number of abstract NPs.

From Figure 5(a), we can see that the number of common NPs is roughly proportional to the number of NPs in the abstract. We then obtained a linear regression model for the data shown in Figure 5(a) and used this model to calculate the number of nodes we should retrieve from the NP network, given the number of abstract NPs in a document is known a priori:

Number of Common NPs =
0.555 * Number of Abstract NPs + 2.435

One could argue that the number of abstract NPs is unknown a priori and thus the proposed method is of limited use. However, the user can provide an estimate based on the desired number of words in the summary. Here we can adopt the same way of asking the user to provide a limit for the NPs in the summary. We used the actual number of NPs the author used in his/her abstract in our experiment.

It should be noted that there are no readily observable relations between the number of common NPs and the number of NPs in the body text in Figure 5(b). Our experiment result shown in Figure 6 suggests that simply using the regression formula is not satisfactory. The average F-measure value is only 0.22. Among the 178 documents, seven of them received a score of zero (0) for both Precision and Recall[3], which means none of the retrieved nodes contained a common NP.

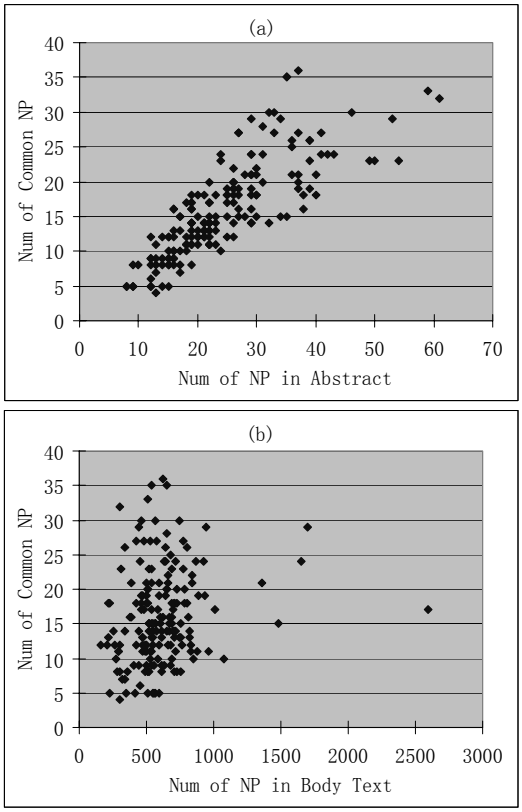


Figure 5. Scatter plots of Common NPs

5.2. Randomly Picking Noun Phrases

One way to study the relationship between common NPs and the NPs in the body text is to regard it as a stochastic process, i.e., the author just picks the NPs from the body text at random. We simulated such a process. The results in Figure 6 show that although the Centrality Heuristic performs not so well in predicting the abstract NPs, it does much better than random selection.

5.3. C4.5 Decision Tree with Rule Induction

Our experiments with the Centrality heuristic have shown that it has a certain effect on determining the relationship between the abstract NPs and NPs in the body text. Nevertheless, our results indicate that the relationship is not solely determined by the centrality, since the values of Precision, Recall, and F-measure are more than 0.2. Other factors, such as position, may participate in

[2] Common NPs refers to the NPs which appear in the body text and the abstract.

[3] In this case, the F-measure is defined to be 0.

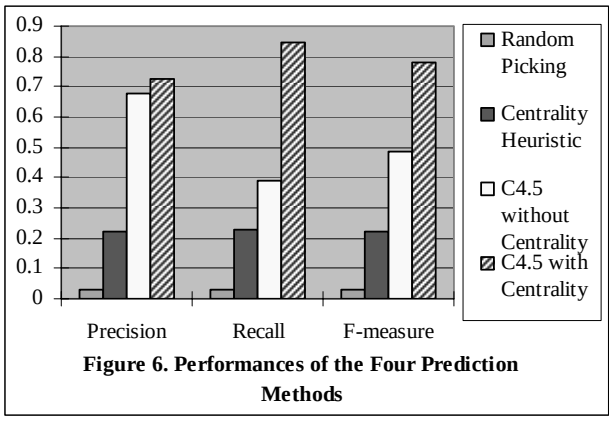


Figure 6. Performances of the Four Prediction Methods

determining the relationship. In order to further study this relationship, we employed C4.5 decision trees with rule induction, trying to find other important factors. For the NPs in the body text, we selected eight attributes:

Position: The order of a NP appearing in the text, normalized by the total number of NPs.

Pronoun type: If the NP contains a pronoun, it is marked as one of four classes: NOMInative, ACCUsative, POSSessive, and AMBIguous (you, her, and it). All other NPs belong to class NONE [12].

Article: Three classes are defined for this attribute: INDEfinite (contains a or an), DEFInite (contains the), and NONE (all others).

Head noun POS tag: A head noun is the last word in the NP. Its POS tag is used here.

Proper name: Whether the NP is a proper name, by looking at the POS tags of all words in the NP.

Centrality: Obtained from the NP network.

Number: Whether the NP is just one number.

In abstract: Whether the NP appears in the author-provided abstract. This attribute is the target for the C4.5 to classify.

The 178 documents have generated more than 100,000 training records. Among them only a very small portion (2.6%) belongs to the positive class. When using decision tree C4.5 on such imbalanced attribute, it is very common that the class with absolute advantages will be favored [13][14]. To reduce the preference bias, one way is to boost the weak class by replicating instances in the minority class [14][15]. In our experiments, the 178 documents were arbitrarily divided into three roughly equal groups, generating 36,157, 37,600, and 34,691 records, respectively. After class balancing, the records are increased to 40,109, 42,210, and 38,499. The three data sets were then run through C4.5 with 10-fold cross-validation. We performed two experiments: one with the feature Centrality, and one without it, in order to see how Centrality affects the predictions. The results are shown in Table 1, where the numbers for Set 1, 2, and 3 are average values for the 10 tests in the cross-validations. The mean values of the metrics are also shown in the Figure 6 in comparison with the Centrality Heuristic and Random Selection. We observed that the Precision, Recall, and F-measure achieved by C4.5 greatly outperform the heuristic and random methods. Meanwhile, using C4.5 with Centrality greatly outperforms using it without Centrality: for the mean values, the Precision is increased by 7%; the Recall of the former is more than twice as the latter, and the F-measure of the former is improved to 160% of the latter! We also studied the rules generated by C4.5. For rules which covered 100 or more instances, we found that 98% of them contain attributes **Position** and **Centrality** while the attribute **Pronoun type** does not appear even once. A typical rule for predicting that a noun phrase should be an abstract noun phrase is:

IF
  Article = NONE and **Number** = False and
  Centrality > 0.017403 and **Position** <= 0.354412
THEN
  NPinAbstract = True

6. Conclusion and Future work

We have presented a new approach to predict summary topics by introducing an important text feature and treating the problem as a classification task. In comparison with the Centrality Heuristic and Random Selection, the approach achieves very promising results for the CMP-LG corpus. The rules induced by C4.5 identified that Position and Centrality are crucial factors in deter-

	C4.5 without Centrality			C4.5 with Centrality		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Set 1	0.7094	0.5196	0.5990	0.7142	0.8051	0.7562
Set 2	0.6874	0.2703	0.3865	0.7501	0.9100	0.8218
Set 3	0.6370	0.3799	0.4739	0.7152	0.8264	0.7665
Mean	0.6779	0.3899	0.4865	0.7265	0.8472	0.7815

Table 1. Results for three data sets

mining the relationship of abstract NPs and NPs in the body text. Using this approach, we can implement a machine-generated topic list to be covered in a summary, which is very close to what a human summarizer would use. The Centrality measure provides a new way to quantify the prominence of a noun phrase in the text. Prior to our study, the Centrality had been used to compare the similarity of two text units in [10]. This feature can be applied to other Natural Language Processing applications which require a means to measure contextual information. Future work will be towards using the NPs predicted in our approach to generate a summary for a given text, which is a problem of language generation if the extractive type summary can not satisfy the user's requirements. We would also like to improve the NP Centrality measure by considering pronoun resolution issues. Furthermore, it is also important to explore how the Centrality is related to the term frequency.

7. Acknowledgement

Thanks to Motorola Labs for their support through the Illinois Manufacturing Research Center.

References:

[1] Z. Xie, X. Li, B. Di Eugenio, W. Xiao, T. Tirpak, and P. Nelson, Using Gene Expression Programming to Construct Sentence Ranking Function for Text Summarization, *Proc. of The 20th International Conference on Computational Linguistics*, Geneva, Switzerland, 2004.

[2] H. Edmundson, New methods in automatic abstracting, *Journal of ACM*, *16*(2), 1969, 264-285.

[3] J. Kupiec, J. Pedersen, and F. Chen, A trainable document summarizer, *Proc. 18th ACM- SIGIR Conference*, Seattle, Washington, 1995, 68-73.

[4] I. Mani, *Automatic Summarization* (Amsterdam/ Philadelphia: John Benjamins Publishing Company, 2001).

[5] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, M. Kan, B. Schiffman, and S. Teufel, Columbia Multi-Document Summarization: Approach and Evaluation, *Proc. of the Document Understanding Conference (DUC01)*. Edmonton, Canada 2001.

[6] E. Hovy, and C. Lin, Automated Text Summarization in SUMMARIST, *Advances in Automatic Text Summarization* (Cambridge, MA: The MIT Press, 1999, 81-94).

[7] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell, Summarizing Text Documents: Sentence Selection and Evaluation Metrics, *Proc. SIGIR '99*, Berkeley, California. 1999, 121-128.

[8] B. Liu, C. Chin, and H. Ng, Mining Topic-Specific Concepts and Definitions on the Web, *Proc. of the Twelfth International World Wide Web Conference (WWW-2003)*, Budapest, Hungary. 2003, 20-24.

[9] S. Corman, T. Kuhn, R. McPhee, and K. Dooley. Studying complex discursive systems: Centering resonance analysis of organizational communication, *Human Communication Research*, *28*(2), 2002, 157-206.

[10] B. J. Grosz, S. Weinstein, and A. K. Joshi, Centering: A framework for modeling the local coherence of a discourse, *Computational Linguistics*, *21*, 1995, 203-225.

[11] E. Brill, Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging, *Computational Linguistics*, *21*(4), 1995, 543-566.

[12] C. Cardie and K. Wagstaff, Noun Phrase Coreference as Clustering, *Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999, 82-89.

[13] N. Japkowicz, The class imbalance problem: significance and strategies, *Proc. of the 2000 International Conference on Artificial Intelligence (ICAI2000)*, 2000.

[14] M. Kubat, and S. Matwin, Addressing the curse of imbalanced data sets: one-sided sampling, *Proc. of the Fourteenth International Conference on Machine Learning*, Morgan Kauffman, 1997, 179–186.

[15] N. Chawla, K. Bowyer, L. Hall, and W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, *Proc. of the International Conference on Knowledge Based Computer Systems*, India, 2000.