Knowledge-based Automatic Topic Identification

Chin-Yew Lin
Department of Electrical Engineering/System
University of Southern California
Los Angeles, CA 90089-2562, USA
chinyew@pollux.usc.edu

Abstract

As the first step in an automated text summarization algorithm, this work presents a new method for automatically identifying the central ideas in a text based on a knowledge-based concept counting paradigm. To represent and generalize concepts, we use the hierarchical concept taxonomy WordNet. By setting appropriate cutoff values for such parameters as concept generality and child-to-parent frequency ratio, we control the amount and level of generality of concepts extracted from the text. [1]

1 Introduction

As the amount of text available online keeps growing, it becomes increasingly difficult for people to keep track of and locate the information of interest to them. To remedy the problem of information overload, a robust and automated text summarizer or information extrator is needed. *Topic identification* is one of two very important steps in the process of summarizing a text; the second step is summary text generation.

A *topic* is a particular subject that we write about or discuss. (Sinclair et al., 1987). To identify the topics of texts, Information Retrieval (IR) researchers use word frequency, cue word, location, and title-keyword techniques (Paice, 1990). Among these techniques, only word frequency counting can be used robustly across different domains; the other techniques rely on stereotypical text structure or the functional structures of specific domains.

Underlying the use of word frequency is the assumption that the more a word is used in a text, the more important it is in that text. This method recognizes only the literal word forms and nothing else. Some morphological processing may help, but pronominalization and other forms of coreferentiality defeat simple word counting. Furthermore, straightforward word counting can be misleading since it misses conceptual generalizations. For example: *"John bought some vegetables, fruit, bread, and milk."* What would be the topic of this sentence? We can draw no conclusion by using word counting method; where the topic actually should be: "John bought some groceries." The problem is that word counting method misses the important concepts behind those words: *vegetables, fruit, etc.* relates to *groceries* at the deeper level of semantics. In recognizing the inherent problem of the word counting method, recently people have started to use artificial intelligence techniques (Jacobs and Rau, 1990; Mauldin, 1991) and statistical techniques (Salton et al., 1994; Grefenstette, 1994) to incorporate the sementic relations among words into their applications. Following this trend, we have developed a new way to identify topics by counting *concepts* instead of words.

2 The Power of Generalization

In order to count concept frequency, we employ a concept generalization taxonomy. Figure 1 shows a possible hierarchy for the concept *digital computer*. According to this hierarchy, if we find *laptop* and *hand-held computer*, in a text, we can infer that the text is about *portable computers*, which is their parent concept. And if in addition, the text also mentions *workstation* and *mainframe*, it is reasonable to say that the topic of the text is related to *digital computer*.

Using a hierarchy, the question is now how to find the most appropriate generalization. Clearly we cannot just use the leaf concepts — since at this level we have gained no power from generalization. On the other hand, neither can we use the very top concept — everything is a *thing*. We need a method of identifying the most appropriate concepts somewhere in middle of the taxonomy. Our current solution uses

[1] This research was funded in part by ARPA under order number 8073, issued as Maryland Procurement Contract # MDA904-91-C-5224 and in part by the National Science Foundation Grant No. MIP 8902426.

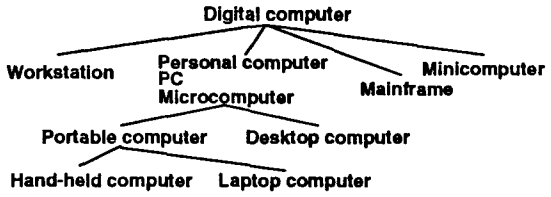

Figure 1: A sample hierarchy for *computer*

concept frequency ratio and starting depth.

2.1 Branch Ratio Threshold

We call the frequency of occurrence of a concept C and it's subconcepts in a text the concept's *weight*[2]. We then define the *ratio* $\mathcal{R}$,at any concept C, as follows:

$$\mathcal{R} = \frac{MAX(weight\ of\ all\ the\ direct\ children\ of\ C)}{SUM(weight\ of\ all\ the\ direct\ children\ of\ C)}$$

$\mathcal{R}$ is a way to identify the degree of summarization informativeness. The higher the ratio, the less concept C generalizes over many children, i.e., the more it reflects only one child. Consider Figure 2. In case (a) the parent concept's ratio is 0.70, and in case (b), it is 0.3 by the definition of $\mathcal{R}$. To generate a summary for case (a), we should simply choose **Apple** as the main idea instead of its parent concept, since it is by far the most mentioned. In contrast, in case (b), we should use the parent concept **Computer Company** as the concept of interest. Its small ratio, 0.30, tells us that if we go down to its children, we will lose too much important information. We define the *branch ratio threshold* ($\mathcal{R}_t$) to serve as a cutoff point for the determination of interestingness, i.e., the degree of generalization. We define that if a concept's *ratio* $\mathcal{R}$ is less than $\mathcal{R}_t$, it is an interesting concept.

2.2 Starting Depth

We can use the ratio to find all the possible interesting concepts in a hierarchical concept taxonomy. If we start from the top of a hierarchy and proceed downward along each child branch whenever the branch ratio is greater than or equal to $\mathcal{R}_t$, we will eventually stop with a list of interesting concepts. We call these interesting concepts the *interesting wavefront*. We can start another exploration of interesting concepts downward from this interesting wavefront resulting in a second, lower, wavefront, and so on. By repeating this process until we reach the leaf concepts of the hierarchy, we can get a set of interesting wavefronts. Among these interesting

[2]According to this, a parent concept always has weight greater or equal to its maximum weighted direct children. A concept itself is considered as its own direct child.

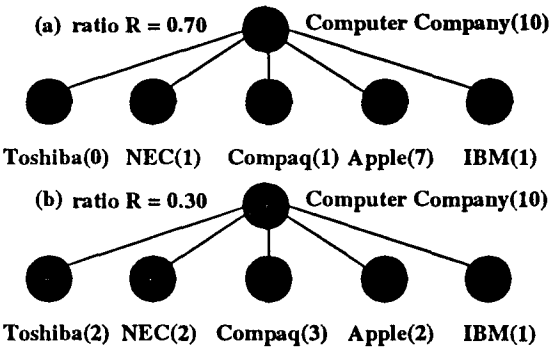

Figure 2: Ratio and degree of generalization

wavefronts, which one is the most appropriate for generation of topics? It is obvious that using the concept counting technique we have suggested so far, a concept higher in the hierarchy tends to be more general. On the other hand, a concept lower in the hierarchy tends to be more specific. In order to choose an adequate wavefront with appropriate generalization, we introduce the parameter *starting depth*, $\mathcal{D}_s$. We require that the branch ratio criterion defined in the previous section can only take effect after the wavefront exceeds the starting depth; the first subsequent interesting wavefront generated will be our collection of topic concepts. The appropriate $\mathcal{D}_s$ is determined by experimenting with different values and choosing the best one.

3 Experiment

We have implemented a prototype system to test the automatic topic identification algorithm. As the concept hierarchy, we used the noun taxonomy from WordNet[3] (Miller et al., 1990). WordNet has been used for other similar tasks, such as (Resnik, 1993) For input texts, we selected articles about information processing of average 750 words each out of *BusinessWeek* (93–94). We ran the algorithm on 50 texts, and for each text extracted eight sentences containing the most interesting concepts.

How now to evaluate the results? For each text, we obtained a professional's abstract from an online service. Each abstract contains 7 to 8 sentences on average. In order to compare the system's selection with the professional's, we identified in the text the sentences that contain the main concepts mentioned in the professional's abstract. We scored how many sentences were selected by both the system and the professional abstractor. We are aware that this evaluation scheme is not very accurate, but it serves as a rough indicator for our initital investigation.

We developed three variations to score the text

[3]WordNet is a concept taxnonmy which consists of synonym sets instead of individual words

sentences on weights of the concepts in the interesting wavefront.

1. the weight of a sentence is equal to the sum of weights of parent concepts of words in the sentence.
2. the weight of a sentence is the sum of weights of words in the sentence.
3. similar to one, but counts only one concept instance per sentence.

To evaluate the system's performance, we defined three counts: (1) *hits*, sentences identified by the algorithm and referenced by the professional's abstract; (2) *mistakes*, sentences identified by the algorithm but not referenced by the professional's abstract; (3) *misses*, sentences in the professional's abstract not identified by the algorithm. We then borrowed two measures from Information Retrieval research:

Recall : $hits/(hits + misses)$
Precision : $hits/(hits + mistakes)$

The closer these two measures are to unity, the better the algorithm's performance. The precision measure plays a central role in the text summarization problem: the higher the precision score, the higher probability that the algorithm would identify the true topics of a text. We also implemented a simple plain word counting algorithm and a random selection algorithm for comparision.

The average result of 50 input texts with branch ratio threshold[4] 0.68 and starting depth 6. The average scores[5] for the three sentence scoring variations are 0.32 recall and 0.35 precision when the system produces extracts of 8 sentences; while the random selection method has 0.18 recall and 0.22 precision in the same experimental setting and the plain word counting method has 0.23 recall and 0.28 precision.

4 Conclusion

The system achieves its current performance without using linguistic tools such as a part-of-speech tagger, syntactic parser, pronoun resoultion algorithm, or discourse analyzer. Hence we feel that the concept counting paradigm is a robust method which can serve as a basis upon which to build an automated text summarization system. The current system draws a performance lower bound for future systems.

We have not yet been able to compare the performance of our system against IR and commerically available extraction packages, but since they do not employ concept counting, we feel that our method can make a significant contribution.

We plan to improve the system's extraction results by incorporating linguistic tools. Our next goal is generating a summary instead of just extracting sentences. Using a part-of-speech tagger and syntatic parser to distinguish different syntatic categories and relations among concepts; we can find appropriate concept types on the interesting wavefront, and compose them into summary. For example, if a noun concept is selected, we can find its accompanying verb; if verb is selected, we find its subject noun. For a set of selected concepts, we then generalize their matching concepts using the taxonomy and generate the list of {selected concepts + matching generalization} pairs as English sentences. There are other possibilities. With a robust working prototype system in hand, we are encouraged to look for new interesting results.

References

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers, Boston.

Paul S. Jacobs and Lisa F. Rau. 1990. SCISOR: Extracting information from on-line news. *Communication of the ACM*, 33(11):88–97, November.

Michael L. Mauldin. 1991. *Conceptual Information Retrieval — A Case Study in Adaptive Partial Parsing.* Kluwer Academic Publishers, Boston.

George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Five papers on wordnet. CSL Report 43, Congnitive Science Labortory, Princeton University, New Haven, July.

Chris D. Paice. 1990. Constructing literature abstracts by computer: Techinques and prospects. *Information Processing and Management*, 26(1):171–186.

Philip Stuart Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships.* Ph.D. thesis, University of Pennsylvania, University of Pennsylvania.

Gerard Salton, James Allan, Chris Buckley, and Amit Singhal. 1994. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, 264:1421–1426, June.

John Sinclair, Patrick Hanks, Gwyneth Fox, Rosamund Moon, and Penny Stock. 1987. *Collins COBUILD English Language Dictionary.* William Collins Sons & Co. Ltd., Glasgow, UK.

[4]This threshold and the starting depth are determined by running the system through different parameter setting. We test ratio = 0.95,0.68,0.45,0.25 and depth = 3,6,9,12. Among them, $\mathcal{R}_t = 0.68$ and $\mathcal{D}_s = 6$ give the best result.

[5]The recall (R) and precision (P) for the three variations are: var1(R=0.32,P=0.37), var2(R=0.30,P=0.34), and var3(R=0.28,P=0.33) when the system picks 8 sentences.