Identifying Topics by Position

Chin-Yew Lin and **Eduard Hovy**
Information Sciences Institute
of the University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292, USA
{cyl,hovy}@isi.edu

Abstract

This paper addresses the problem of identifying likely topics of texts by their position in the text. It describes the automated training and evaluation of an Optimal Position Policy, a method of locating the likely positions of topic-bearing sentences based on genre-specific regularities of discourse structure. This method can be used in applications such as information retrieval, routing, and text summarization.

1 Introduction: Topic Spotting by Position

In an increasingly information-laden world, the problem of automatically finding the major topics of texts acquires new urgency. A module that can suggest likely locations of topics in texts, robustly and with an acceptable degree of correctness, would be useful for a number of important applications, including information retrieval, gisting, and automated summarization.

Several methods have been tried to perform Topic Identification. Some involve parsing and semantic analysis of the text, and are therefore less robust over arbitrary input. Others, such as the *Cue Phrase* and *Position* methods, are more robust, though generally somewhat less accurate. Of these, the Position Method, identified in the late 1950's, remains among the best; it can outperform newer methods such as those based on word counting (Salton et al., 1994).

The Position Method springs from the recognition that texts in a genre generally observe a predictable discourse structure, and that sentences of greater topic centrality tend to occur in certain specifiable locations. The text's title, for example, is a very informative position in most genres, as is the Abstract paragraph in scientific articles. Edmundson (Edmundson, 1969) defined the Position Method as follows:

> "...the machine-readable cues are certain general characteristics of the corpus provided by the skeletons of documents, i.e. headings and format. The Location method is based on the hypothesis that: (1) sentences occurring under certain headings are positively relevant; and (2) topic sentences tend to occur very early or very late in a document and its paragraphs."

However, since the paradigmatic discourse structure differs significantly over text genres and subject domains, the Position Method cannot be defined as straightforwardly as Baxendale's (Baxendale, 1958) *title plus first and last sentences of each paragraph*; it has to be tailored to genre and domain. Can one develop ways of tailoring this method?

Furthermore, since the resolution power of the Position Method is the sentence, while the desired output—topics—generally appear at the word or phrase level, the most accurate results of this method may still include too much spurious material to be really useful. How useful is the method in general? By what measure(s) can one evaluate it?

Basic questions about how the Position Method can be tailored for optimality over a genre and how it can be evaluated for effectiveness remain unanswered. To our knowledge, no systematic study has yet been performed, though some variant of it has been used in computational studies (see for example (Edmundson, 1969; Luhn, 1958; Baxendale, 1958)), writing-education classes (for example, (Sjostrom and Hare, 1984)), and has been the subject of cognitive psychological verification (Kieras, 1985).

This paper contains an analysis of the Position Method. We first discuss previous work, then in Section 3 describe the background studies and training of an Optimal Position Policy for a genre of texts, and in Section 4 describe its evaluation.

2 Related Work

Edmundson's (Edmundson, 1969) laid the groundwork for the Position Method. He introduced four clues for identifying significant words (topics) in a text. Among them, *Title* and *Location* are related to the Position Method. Edmundson assigned pos-

itive weights to sentences according to their ordinal position in the text, giving most weight to the first sentence in the first paragraph and the last sentence in the last paragraph. He conducted seventeen experiments to verify the significance of these methods. According to his results, the Title and Location methods respectively scored around 40% and 53% accuracy, where accuracy was measured as the coselection rate between sentences selected by Edmundson's program and sentences selected by a human.

Although Edmundson's work is fundamental, his experiments used only 200 documents for training and another 200 documents for testing. Furthermore, he did not trying out other possible combinations, such as the second and third paragraphs or the second-last paragraph. In order to determine where the important words are most likely to be found, Baxendale (Baxendale, 1958) conducted an investigation of a sample of 200 paragraphs. He found that in 85% of paragraphs the topic sentence was in the first sentence and in 7% the final one. Donlan (Dolan, 1980) stated that a study of topic sentences in expository prose showed that only 13% of paragraphs of contemporary professional writers began with topic sentences (Braddock, 1974). Singer and Donlan (Singer and Dolan, 1980) maintain that a paragraph's main idea can appear anywhere in the paragraph, or not be stated at all.

Arriving at a negative conclusion, Paijmans (Paijmans, 1994) conducted experiments on the relation between word position in a paragraph and its significance, and found that "words with a high information content according to the tf.idf-based weighting schemes do not cluster in the first and the last sentences of paragraphs or in paragraphs that consist of a single sentence, at least not to such an extent that such a feature could be used in the preparation of indices for Information Retrieval purposes." In contrast, Kieras (Kieras, 1985) in psychological studies confirmed the importance of the position of a mention within a text.

3 Training the Rules

3.1 Background

The purposes of our study are to clarify these contradictions, to test the abovementioned intuitions and results, and to verify the hypothesis that the importance of a sentence in a text is indeed related to its ordinal position. Furthermore, we wish to discover empirically which textual positions are in fact the richest ones for topics, and to develop a method by which the optimal positions can be determined automatically and their importance evaluated.

To do all this, one requires a much larger document collection than that available to Edmundson and Baxendale. For the experiments described here, we used the Ziff-Davis texts from the corpus produced for DARPA's TIPSTER program (Harman, 1994). Volume 1 of the Ziff corpus, on which we trained the system, consists of 13,000 newspaper texts about new computers and related hardware, computer sales, etc., whose genre can be characterized as product announcements. The average text length is 71 sentences (34.4 paragraphs). Each text is accompanied by both a set of three to eight topic keywords and an abstract of approx. 6 sentences (both created by a human).

In summary, we did the following: To determine the efficacy of the Position Method, we empirically determined the *yield* of each sentence position in the corpus, measuring against the topic keywords. We next ranked the sentence positions by their average yield to produce the *Optimal Position Policy* (OPP) for topic positions for the genre. Finally, now comparing to the abstracts accompanying the texts, we measured the *coverage* of sentences extracted from the texts according to the policy, cumulatively in the position order specified by the policy. The high degree of coverage indicated the effectiveness of the position method.

3.2 Sentence Position Yields and the Optimal Position Policy

We determined the optimal position for topic occurrence as follows. Given a text **T** and a list of topics keywords $\mathbf{t}_i$ of **T**, we label each sentence of **T** with its ordinal paragraph and sentence number (P_m,S_n). We then removed all closed-class words from the texts. We did not perform morphological restructuring (such as canonicalization to singular nouns, verb roots, etc.) or anaphoric resolution (replacement of pronouns by originals, etc.), for want of robust enough methods to do so reliably. This makes the results somewhat weaker than they could be.

What data is most appropriate for determining the optimal position? We had a choice between the topic keywords and the abstracts accompanying each text in the corpus. Both keywords and abstracts contain phrases and words which also appear in the original texts; on the assumption that these phrases or words are more important in the text than other ones, we can assign a higher importance to sentences with more such phrases or words (or parts of them).[1] Since a topic keyword has a fixed boundary, using it to rank sentences is easier than using an abstract.

For this reason we defined *sentence yield* as the average number of different topic keywords mentioned in a sentence. We computed the yield of each sentence position in each text essentially by counting

[1] How many topic keywords would be taken over verbatim from the texts, as opposed to generated paraphrastically by the human extractor, was a question for empirical determination—the answer provides an upper bound for the power of the Position Method.

the number of different topic keywords contained in the appropriate sentence in each text, and averaging over all texts. Sometimes, however, keywords consist of multiple words, such as "spreadsheet software". In order to reward a full-phrase mention in a sentence over just a partial overlap with a multiword keyword/phrase, we used a formula sensitive to the degree of overlap. In addition, to take into account word position, we based this formula on the Fibonacci function; it monotonically increases with longer matched substrings, and is normalized to produce a score of 1 for a complete phrase match. Our hit function **H** measures the similarity between topic keyword $\mathbf{t}_i$ and a window $\mathbf{w}_{ij}$ that moves across each sentence (P_m,S_n) of the text. A window matches when it contains the same words as a topic keyword $\mathbf{t}_i$. The length of the window equals the length of the topic keyword. Moving the window from the beginning of a sentence to the end, we computed all the $\mathbf{H}_s$ scores and added them together to get the total score $\mathbf{H}_s$ for the whole sentence. We acquired the $\mathbf{H}_s$ scores for all sentences in **T** and repeated the whole process for the each text in the corpus. After obtaining all the $\mathbf{H}_s$ scores, we sorted all the sentences according to their paragraph and sentence numbers. For each paragraph and sentence number position, we computed the average $\mathbf{H}_{avg}$ score.

These average yields for each position are plotted in Figure 1, which shows the highest-yield sentence position to be (P_2,S_1), followed by (P_3,S_1), followed by (P_4,S_1), etc.

Finally, we sorted the paragraph and sentence position by decreasing yield $\mathbf{H}_{avg}$ scores. For positions with equal scores, different policies are possible: one can prefer sentence positions in different paragraphs on the grounds that they are more likely to contains distinctive topics. One should also prefer sentence positions with smaller S_m, since paragraphs are generally short. Thus the Optimal Position Policy for the Ziff-Davis corpus is the list

[(T) (P_2,S_1) (P_3,S_1) (P_2,S_2) {(P_4,S_1) (P_5,S_1) (P_3,S_2)} {(P_1,S_1) (P_6,S_1) (P_7,S_1) (P_1,S_3) (P_2,S_3)} ...]

3.3 Additional Measures and Checks

Throughout the above process, we performed additional measures and checks in order to help us prevent spurious or wrong rules. We collected facts about the training corpus, including the average number of paragraphs per text (PPT), the average number of sentences per paragraph (SPP), and the average number of sentences per human-made summary (SPS). PPT and SPP prevent us from forming a rule such as *25th sentence in the 100th paragraph* when PPT is 15 and SPP is 5. SPS suggests how many sentences to extract. For the ZIFF Vol. 1 corpus, PPT is 34.43, SPP is 2.05, and SPS is 5.76. Most texts have under 30 paragraphs; 97.2% of paragraphs have fewer than 5 sentences. 47.7% of paragraphs have only one sentence (thus the first sentence is also the last), and 25.2% only two. With regard to the abstracts, most have 5 sentences and over 99.5% have fewer than 10.

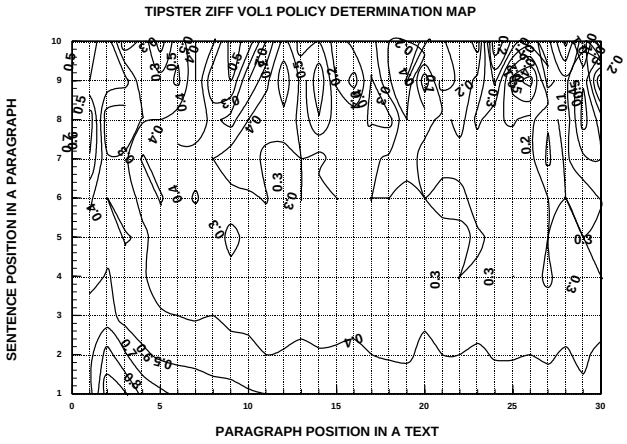

Figure 1: Average yield by paragraph and sentence position; lightest shade shows highest yield.

We also counted how many different topic keywords each specific text unit contains, counted once per keyword. This *different hit* measure *dhit* played an important role, since the OPP should be tuned to sentence positions that bear as many different topic keywords as possible, instead of positions with very high appearances of just a few topic keywords. We can compute *dhit* for a sentence, several sentences, or several paragraphs. *Sentenceyield* is *dhit* score of a sentence. Figure 2 shows *dhit* scores for the first 50 paragraph positions, and Figure 3 *dhit* scores for the last 50 positions (counting backward from the end of each text). Since PPT=34.43, the first and last 50 positions fully cover the majority of texts. The former graph illustrates the immense importance of the title sentence ($dhit = 1.96$), and the importance of the second ($dhit = 0.75$) and third ($dhit = 0.64$) paragraphs relative to the first ($dhit = 0.59$). Paragraphs close to the beginning of texts tend to bear more informative content; this is borne out in Figure 3, which clearly indicates that paragraph positions close to the end of texts do not show particularly high values, while the peak occurs at position P_{-14} with $dhit = 0.42$. This peak occurs precisely where most texts have their second or third paragraphs (recall that the average text length is 13 to 16 paragraphs).

To examine Baxendale's first/last sentence hypothesis, we computed the average *dhit* scores for the first and the last 10 sentence positions in a paragraph as shown in Figure 4 and Figure 5 respectively. The former indicates that the closer a sentence lies

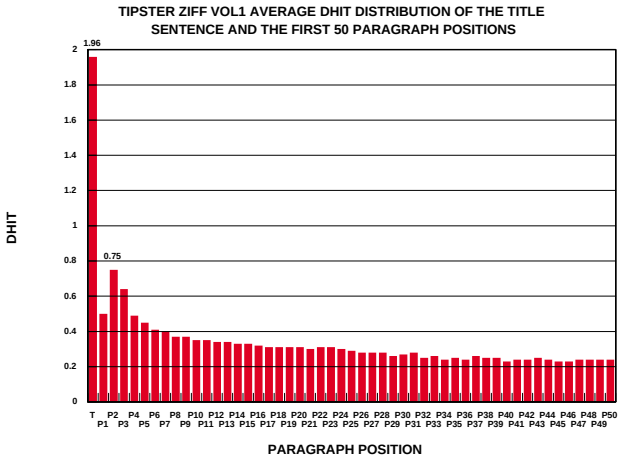

Figure 2: Vol. 1 *dhit* distribution for the title sentence and the first 50 paragraph positions.

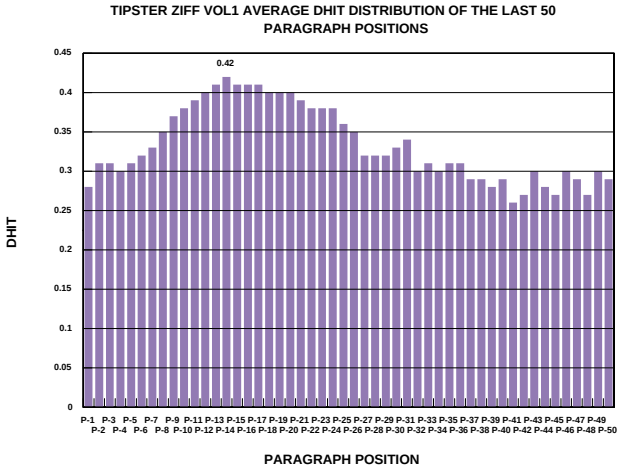

Figure 3: Vol. 1 *dhit* distribution for the last 50 paragraph positions, counting backward.

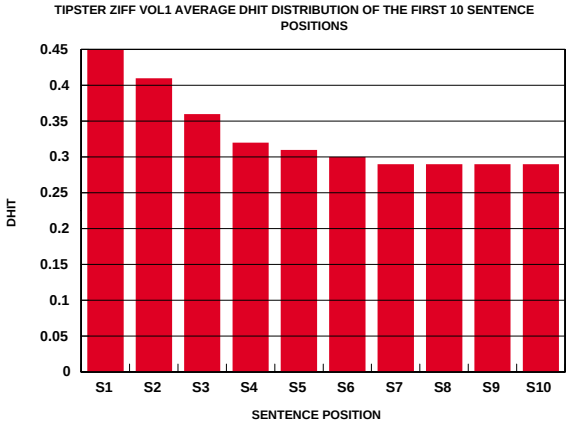

Figure 4: Vol. 1 *dhit* distribution of the first 10 sentence positions in a paragraph.

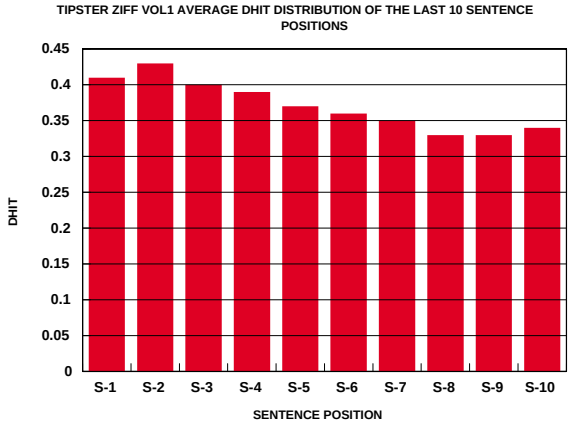

Figure 5: Vol. 1 *dhit* distribution of the last 10 sentence positions in a paragraph.

to the beginning of a paragraph, the higher its *dhit* score is. This confirms the *first sentence* hypothesis. On the other hand, the latter figure does not support the *last sentence* hypothesis; it suggests instead that the *second* sentence from the end of a paragraph contains the most information. This is explained by the fact that 47.7% of paragraphs in the corpus contain only one sentence and 25.2% of the paragraphs contain two sentences, and the SPP is 2.05: the second-last sentence *is* the first!

4 Evaluation

The goal of creating an Optimal Position Policy is to adapt the position hypothesis to various domains or genres in order to achieve maximal topic coverage. Two checkpoints are required:

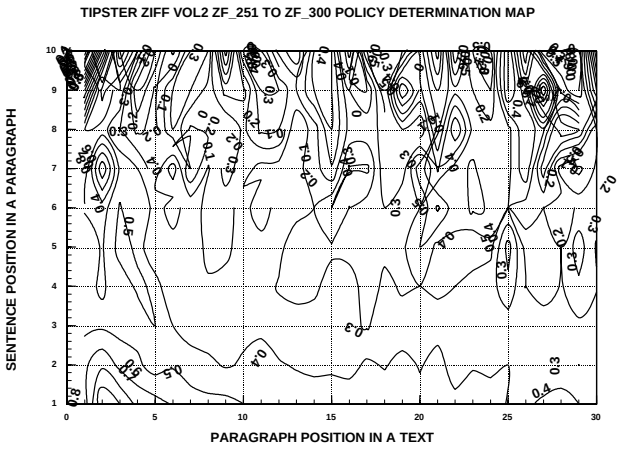

Figure 6: Vol. 2 optimal position Policy Determination Map in contour view.

1. applying the procedure of creating an OPP to another collection in the same domain should result in a similar OPP, and
2. sentences selected according to the OPP should indeed carry more information than other sentences.

Two evaluations were conducted to confirm these points.

In both cases, we compared the sentences extracted according to the OPP to the sentences contained in the human-generated abstracts. Though we could have used topic keywords for both training and evaluation, we decided that the abstracts would provide a more interesting and practical measure for output, since the OPP method extracts from the text full sentences instead of topic phrases. Accordingly, we used as test corpus another, previously unseen, set of 2,907 texts from Vol. 2 of the Ziff-Davis corpus, which contained texts of the same nature and genre as Vol. 1.

4.1 Evaluation I

This evaluation established the validity of the Position Hypothesis, namely that the OPP so determined does in fact provide a way of identifying high-yield sentences, and is not just a list of average high-yield positions of the corpus we happened to pick. following the same steps as before, we therefore derived a new OPP on the test corpus.

The result of the average scores of 300 positions (P_m, S_n) shown in Figure 6, with $1 \leq m \leq 30$ and $1 \leq n \leq 10$, was a contour map highly similar to Figure 1.

Both peak at position (P_2, S_1) and decrease gradually in the X direction and more rapidly in the Y direction. The similarity between the policy determination maps of the training and test sets confirms two things: First, correspondences exist between topics and sentence positions in texts such as the ZIFF-Davis collection. Second, the regularity between topics and sentence positions can be used to identify topic sentences in texts.

4.2 Evaluation II

In the evaluation, we measured the word overlap of sentences contained in the abstracts with sentence(s) extracted from a text according to the OPP. For each measure, we recorded scores cumulatively, choosing first the most promising sentence according to the OPP, then the two most promising, and so on.

We measured word overlap as follows: first, we removed all function (closed-class) words from the abstract and from the text under consideration. Then, for the first 500 sentence positions (the top 1, 2, 3,. . . , taken according to the OPP), we counted the number of times a window of text in the extracted sentences matched (i.e., exactly equalled) a window of text in the abstract. (Again we performed no morphology manipulations or reference resolution, steps which would improve the resulting scores.) We performed the counts for window lengths of 1, 2, 3, 4, and 5 words. If a sentence in an abstract matched more than one sentence extracted by the OP, only the first match was tallied. For each number of sentences extracted, and for each window size, we averaged the counts over all 2,907 texts.

We define some terms and three measures used to assess the quality of the OPP-selected extracts. For an extract E and a abstract A:

w^E_{mi}: a window i of size m in E.

w^A_{mi}: a window i of size m in A.

$|W^E_m|$: total number of windows of size m in E.

$|W^A_m|$: total number of different windows of size m in A, i.e., how many $w^A_{mi} \neq w^A_{mj}$.

hit : $w^E_{mi} = w^A_{mj}$, i.e., words and word sequences in w^E_{mi} and w^A_{mj} are exactly the same.

Precision of windows size m:

$$P_m = \frac{\#\ hits}{|W^E_m|}$$

Recall of windows size m:

$$R_m = \frac{\#\ different\ hits}{|W^A_m|}$$

Coverage of windows size m:

$$C_m = \frac{\#\ sentences\ in\ A\ with\ at\ least\ one\ hit}{\#\ sentences\ in\ A}$$

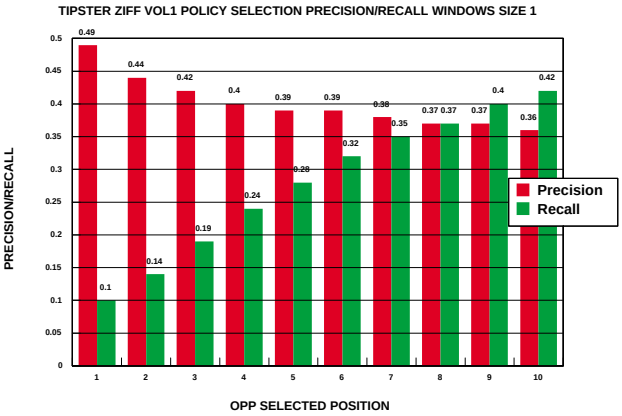

Figure 7: Cumulative precision/recall scores of top ten OPP-selected sentence positions of window size 1.

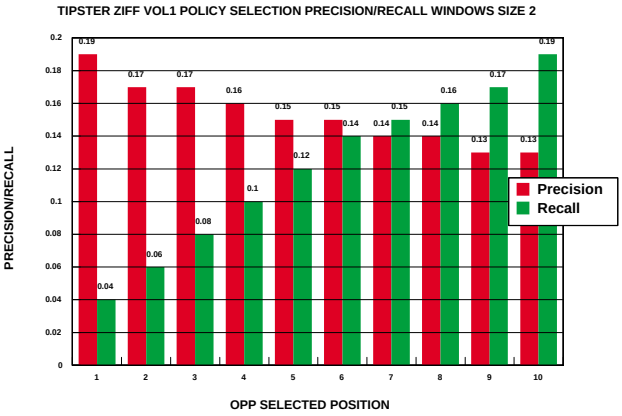

Figure 8: Cumulative precision/recall scores of top ten OPP-selected sentence positions of window size 2.

4.2.1 Precision and Recall

Precision, P_m, measures what percentage of windows of size m in E can also be found in A (that is, P_m indicates what percentage of E is considered important with regard to A). Recall, R_m, measures the diversity of E. A high P_m does not guarantee recovery of all the possible topics in A, but a high R_m does ensure that many different topics in A are covered in E. However, a high R_m alone does not warrant good performance either. For example, an OPP that selects all the sentences in the original text certainly has a very high R_m, but this extract duplicates the original text and is the last thing we want as a summary! Duplicate matches (the same word(s) in different windows) were counted in P but not in R.

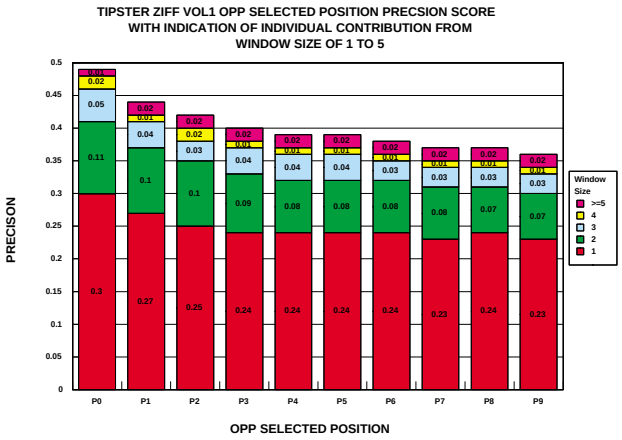

Figure 9: Precision scores show individual contribution from window size 1 to 5.

Figure 7 and Figure 8 show the precision/recall graphs of window sizes 1 and 2 respectively. Figure 7 indicates that the precision score decreases slowly and the recall score increases more rapidly as we choose more sentences according to the OPP. Selecting 7 sentences (is 10% of the average length of a ZIFF text), the precision is 0.38 and the recall 0.35. Considering that the matching process requires exact match and morphological transformation is not used, this result is very encouraging. However, with window size 2, precision and recall scores drop seriously, and more so with even larger windows. This suggests using variable-length windows, sizing according to maximal match. So doing would also avoid counting matches on window size 1 into matches of larger window sizes. The contributions of precision, P^o_m, and recall, R^o_m, from each m-word window alone, can be approximated by:

$$\begin{aligned} P^o_m &\approx P_m - P_{m+1} \\ R^o_m &\approx R_m - R_{m+1} \end{aligned}$$

Figure 9 and Figure 10 show precision and recall scores with individual contributions from window sizes 1 to 5. Precision P_v and recall R_v of variable-length windows can be estimated as follows:

$$\begin{aligned} P_v &\approx \sum_{m=1}^{l} P^o_m \\ R_v &\approx \sum_{m=1}^{l} R^o_m \end{aligned}$$

The performance of variable-length windows compared with windows of size 1 should have a difference less than the amount shown in the segments of window size ≥ 5.

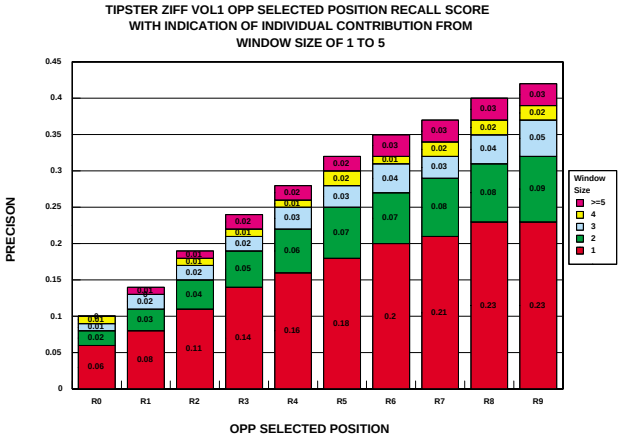

Figure 10: Recall scores show individual contribution from window size 1 to 5.

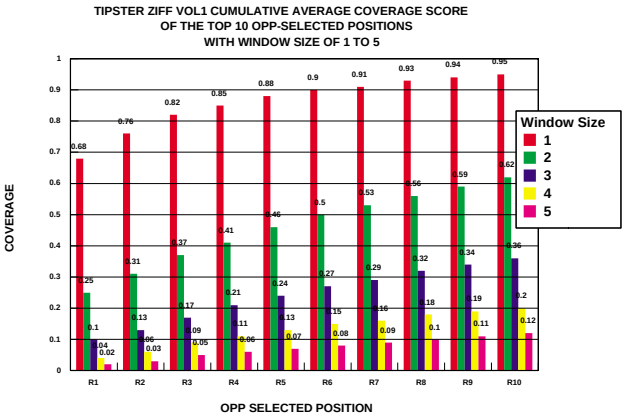

Figure 11: Cumulative coverage scores of top ten sentence positions according to the OPP, with window sizes 1 to 5.

4.2.2 Coverage

Coverage, C_m, tests similarity between E and A in a very loose sense. It counts the number of sentences in A with at least one hit in E (i.e., there exists at least one pair of windows w^A_{mi} and w^E_{mj} such that $w^A_{mi} = w^E_{mj}$). C_m estimates the potential of the OPP procedure. Figure 11 shows the cumulative average coverage scores of the top ten sentence positions of the training set following the OPP. Figure 11 indicates that 68% of sentences in A shared with the title sentence at least one word, 25% two words, 10% three words, 4% four words, and 2% five words. The amount of sharing at least one word goes up to 88% if we choose the top 5 positions according to the OPP and 95% if we choose the top 10 positions!

The contribution of coverage score, C^o_m, solely from m-word match between E and A can be computed as follows:

$$C^o_m = C_m - C_{m-1}$$

The result is shown in Figure 12. Notice that the topmost segment of each column in Figure 12 represents the contribution from matches of at least five words long, since we only have C_m up to $m = 5$. The average number of sentences per summary (SPS) is 5.76. If we choose the top 5 sentence positions according to the OPP, Figure 12 tells us that these 5-sentences extracts E (the average length of an abstract), cover 88% of A in which 42% derives solely from one-word matches, 22% two words, 11% three words, and 6% four words. The average number of sentences per text in the corpus is about 70. If we produce an extract of about 10% of the average length of a text, i.e. 7 sentences, the coverage score is 0.91. This result is extremely promising and confirms the OPP-selected extract bearing important contents.

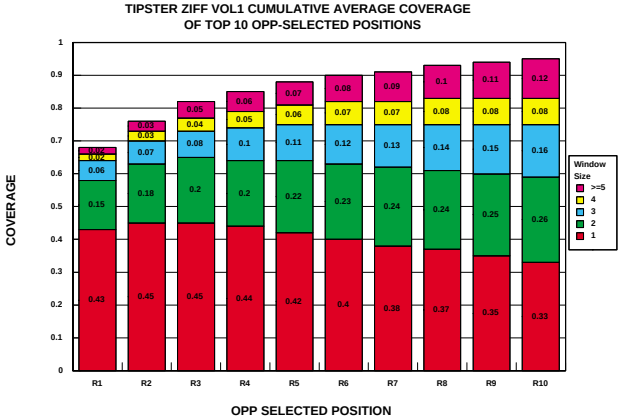

Figure 12: Cumulative coverage scores of top ten sentence positions with contribution marked for each window size.

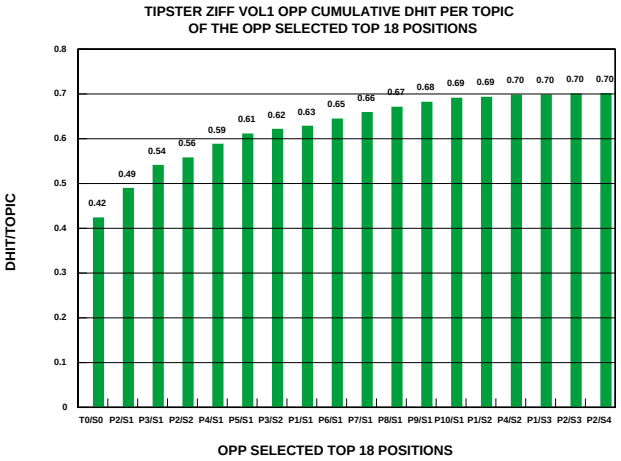

Figure 13: Cumulative *dhit* per topic for the top 18 OPP selected positions.

5 Conclusion

This study provides empirical validation for the Position Hypothesis. It also describes a method of deriving an Optimal Position Policy for a collection of texts within a genre, as long as a small set of topic keywords is defined with each text. The Precision and Recall scores indicate the selective power of the Position method on individual topics, while the Coverage scores indicate a kind of upper bound on topics and related material as contained in sentences from human-produced abstracts.

The results displayed in Figure 13 are especially promising. It is clear that only about 30% of topic keywords are not mentioned in the text directly. This is excellent news: it means that as an upper bound, only about 30% of the humans' abstracts in this domain derive from some inference processes, which means that in a computational implementation only about the same amount has to be derived by processes yet to be determined. Second, the title contains about 50% of the topic keywords; the title plus the two most rewarding sentences provide about 60%, and the next five or so add another 6%. Thus, a fairly small number of sentences provides 2/3 of the keyword topics.

It must be remembered that our evaluations treat the abstract as ideal—they rest on the assumption that the central topic(s) of a text are contained in the abstract made of it. In many cases, this is a good assumption; it provides what one may call the author's perspective of the text. But this assumption does not support goal-oriented topic search, in which one wants to know whether a text pertains to some particular prespecified topics. For a goal-oriented perspective, one has to develop a different method to derive an OPP; this remains the topic of future work.

Ultimately, the Position Method can only take one a certain distance. Because of its limited power of resolution—the sentence—and its limited method of identification—ordinal positions in a text—it has to be augmented by additional, more precise techniques. But the results gained from what is after all a fairly simple technique are rather astounding nonetheless.

References

P. B. Baxendale. 1958. Machine-made index for technical literature — an experiment. *IBM Journal*, pages 354–361, October.

Richard Braddock. 1974. The frequency and placement of topic sentences in expository prose. In *Research in The Teaching of English*, volume 8, pages 287–302.

Dan Dolan. 1980. Locating main ideas in history textbooks. In *Journal of Reading*, pages 135–140.

H. P. Edmundson. 1969. New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285.

Donna Harman. 1994. Data preparation. In R. Merchant, editor, *The Proceedings of the TIPSTER Text Program Phase I*, San Mateo, California. Morgan Kaufmann Publishing Co.

D.E. Kieras, 1985. *Thematic Process in the Comprehension of Technical Prose*, pages 89–108. Lawrence Elrbaum Association, Hillsdale, New Jersey.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal*, pages 159–165, April.

J.J. Paijmans. 1994. Relative weights of words in documents. In L.G.M. Noordman and W.A.M. de Vroomen, editors, *Conference Proceedings of STINFON*. StinfoN.

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.

Harry Singer and Dan Dolan. 1980. *Reading And Learning from Text*. Little Brown, Boston, Mass.

Colleen Langdon Sjostrom and Victoria Chou Hare. 1984. Teaching high school students to identify main ideas in expository text. *Journal of Educational Research*, 78(2):114–118.