A Grammatico-Statistical Approach to Discourse Partitioning

Tadashi Nomoto and Yoshihiko Nitta
Advanced Research Laboratory, Hitachi Ltd.*
email:{nomoto, nitta}@harl.hitachi.co.jp

Abstract

The paper presents a new approach to text segmentation — which concerns dividing a text into coherent discourse units. The approach builds on the theory of discourse segment (Nomoto and Nitta, 1993), incorporating ideas from the research on information retrieval (Salton, 1988). A discourse segment has to do with a structure of Japanese discourse; it could be thought of as a linguistic unit demarcated by *wa*, a Japanese topic particle, which may extend over several sentences. The segmentation works with discourse segments and makes use of coherence measure based on *tf·idf*, a standard information retrieval measurement (Salton, 1988; Hearst, 1993). Experiments have been done with a Japanese newspaper corpus. It has been found that the present approach is quite successful in recovering articles from the unstructured corpus.

Introduction

In this paper, we describe a method for discovering coherent texts from the unstructured corpus. The method is both linguistically and statistically motivated. It derives a linguistic motivation from the view that discourse consists of what we call *discourse segments*, minimal coherent units of discourse (Nomoto and Nitta, 1993), while statistically it is guided by ideas from information retrieval (Salton, 1988). Previous quantitative approaches to text segmentation (Hearst, 1993; Kozima, 1993; Youmans, 1991) have paid little attention to a statistically important structure that a discourse might have and defined it away as a lump of words or sentences. Part of our concern here is with explicating possible effects of a discourse segment on the quantitative structuring of discourse.

In what follows, we will describe some important features about discourse segment and see how it can be incorporated into a statistical analysis of discourse. Also some comparison is made with other approaches such as (Youmans, 1991), followed by discussion on the results of the present method.

Theory of Discourse Segment

The theory of discourse segment (Nomoto and Nitta, 1993) carries with it a set of empirical hypotheses about structure of Japanese discourse. Among them is the claim that Japanese discourse is constructed from a series of linguistics units called *discourse segment*. The discourse segment is thought of as a topic-comment structure, where a topic corresponds to the subject matter and a comment a discussion about it. In particular, Japanese has a special way of marking the topic: by suffixing it with a postpositional particle *wa*. Thus in Japanese, a topic-comment structure takes the form:

$$\underbrace{\bullet\bullet\bullet}_{topic}\ wa\ \underbrace{\bullet\bullet\bullet\bullet\bullet}_{comment}$$

where "•" represents a word. The comment part could become quite long, extending over quite a few sentences (Mikami, 1960). Now Japanese provides for a variety of ways to mark off a topic-comment structure; the *wa*-marking is one such and a typographical device such as a line- or a page-break is another. For the present discussion, we take a discourse segment to be a block of sentences bounded by a text break and/or a *wa*-marked element.

$$|\ \underbrace{T_1\, S_1\, S_2 \cdots S_n}_{Discourse\ Segment}\ |\ \underbrace{T_2\, S_{n+1}\, S_{n+2} \cdots S_{n+m}}_{Discourse\ Segment}\ |,$$

where "T" denotes a boundary marker, "S" a sentence, and "|" a segment gap. For the semantics of a discourse segment, Nomoto and Nitta (1993) observes an interesting tendency that zero (elliptical) anaphora occurring within the segment do not refer across the segment boundary; that is, their references tend to be resolved internally[1].

Now we take a very simple view about the global structure of discourse: syntactically, discourse is just

[1]Here and throughout we use 01 for a subject(NOMinative) zero; 02 for a object(ACCusative) zero; TOP for a topic case; DAT for a dative(indirect object) case; PASS for a passive morpheme.

| Taro$_{<i>}$ -wa 01$_{<i>}$ rojin$_{<j>}$ -ni seki
TOP *old man* *DAT* *seat*

-wo yuzutte -ageta node, 01$_{<i>}$ 02$_{<j>}$
ACC *give* *help* *because*

orei -wo iwar eta. |
thank *say* *PASS*

"Because Taro gave the old man a favor of giving a seat, he thanked Taro."

Note that all the instances of 01 and 02 have internal antecedents: *Taro* and *rojin*.

*2520 Hatoyama Saitama 350-03 Japan
tel. +81-492-96-6111 fax. +81-492-96-6006

a chronological juxtaposition of contiguous, disjoint blocks of sentences, each of which corresponds to a discourse segment; semantically, discourse is a set of anaphoric islands set side by side. Thus a discourse should look like Figure 1, where G denotes a discourse segment. Furthermore, we do not intend the dis-

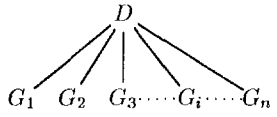

Figure 1: Discourse Structure

course structure to be anything close to the ones that rhetorical theories of discourse (Hovy, 1990; Mann and Thompson, 1987; Hobbs, 1979) claim it to be, or *intentional structure* (Grosz and Sidner, 1986) ; indeed we do not assume any functional relation, i.e. causation, elaboration, extension, etc., among the segments that constitute a discourse structure. The present theory is not so much about the rhetoric or the function of discourse as about the way anaphora are interpreted.

It is quite possible that a set of discourse segments are not aggregated into a single discourse but may have diverse discourse groupings (Nomoto and Nitta, 1993). This happens when discourses are interleaved or embedded into some other. An interleaving or an embedding of discourse is often invoked by changes in narrative mode such as direct/indirect speech, quoting, or interruption; which will cue the reader/hearer to suspend a current discourse flow, start another, or resume the interrupted discourse.

A Quantitative Structuring of Discourse

Vector Space Model

Formally, a discourse segment is represented as a *term vector* of the form:

$$G_i = (g_{i1}, g_{i2}, g_{i3}, \ldots, g_{it})$$

where a g_i represents a nominal occurrence in G_i. In the information retrieval terms, what happens here is that a discourse segment G_i is *indexed* with a set of terms g_{i1} through g_{it}; namely, we characterize G_i with a set of indices $g_{i1}, \ldots, g_{it}$. A term vector can either be *binary*, where each term in the vector takes 0 or 1, i.e., absence or presence, or *weighted*, where a term is assigned to a certain importance value. In general, the weighted indexing is preferable to the binary indexing, as the latter policy is known to have problems with precision (Salton, 1988)[2]. The weighting policy that we will adopt is known as *tf· idf*. It is an indicator of term importance w_{ij} defined by:

$$w_{ij} = tf_{ij} \cdot \log \frac{N}{df_j}$$

where *tf* (term frequency) is the number of occurrences of a term T_j in a document D_i; *df* (document frequency) is the number of documents in a collection of N documents in which T_j occurs; and the importance, w_{ij} is given as the product of *if* and the inverse *df* factor ,or *idf*, $\log N/df_j$. With the *tf·idf* policy, high-frequency terms that are scattered evenly over the entire documents collection are considered to be less important than those that are frequent but whose occurrences are concentrated in particular documents[3]. Thus the *tf·idf* indexing favors rare words, which distinguish the documents more effectively than common words.

With the indexing method in place, it is now possible to define the coherence between two term vectors. For term vectors $X = (x_1, x_2, \ldots, x_t)$ and $Y = (y_1, y_2, \ldots, y_t)$, let the coherence be defined by:

$$C(X,Y) = \frac{2\sum_{i=1}^{t} w(x_i)w(y_i)}{\sum_{i=1}^{t} w(x_i)^2 + \sum_{i=1}^{t} w(y_i)^2}$$

where $w(x_i)$ represents a *tf·idf* weight assigned to the term x_i. The measure is known as *Dice coefficient*[4].

Experiments

Earlier quantitative approaches to text partitioning (Youmans, 1991; Kozima, 1993; Hearst, 1993) work with an arbitrary block of words or sentences to determine a structure of discourse. In contrast, we work with a block of discourse segments. It is straightforward to apply the *tf·idf* to the analysis of discourse; one may just treat a block of discourse segments as a document unit by itself and then define the term frequency (*tf*), the document frequency (*df*), and the size of documents collection (N), accordingly. Coherence would then be determined by the number of terms segment blocks share and *tf·idf* weights the terms carry. Thus one pair of blocks will become more cohesive than another if the pair share more of the terms that are locally frequent.

The partitioning proceeds in two steps. We start with the following:

1. Collect all the nominal occurrences found in a corpus[5]
2. Divide the collection into disjoint discourse segments.
3. Compare all pairs of adjacent blocks of discourse segments.

[2]Precision measures the proportion of correct items retrieved against the total number of retrieved items.

[3]Precision depends on the size of documents collection; as the collection gets smaller in size, index terms become less extensive and more discriminatory. The *idf* factor could be dispensed with in such cases.

[4]Other measures standardly available in the information retrieval include *inner product, cosine coefficient,* and *Jaccard coefficient.*

[5]This is done by JUMAN, a Japanese morphological analyzer (Matsumoto et al., 1993).

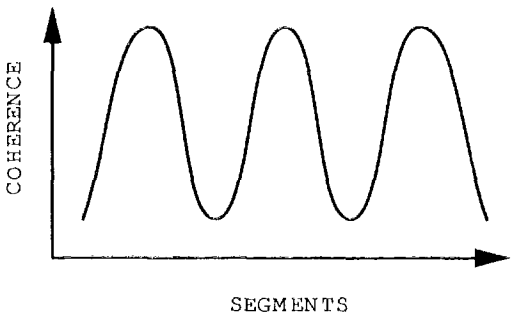

Figure 2: A coherence curve

4. Assign a coherence/similarity value to each pair.

Next, we examine the coherence curve for hills and valleys and partition it manually. Valleys (viz, low coherence values) are likely to signal a potential break in a discourse flow, whereas hills (viz, high coherence values) would signal local coherency. Figure 2 shows how a coherence curve might appear.

Coherence is measured at every segment with a paired comparison window moving along the sequence of segments. Or more precisely, given a segment d_j and a block size n, what we do is to compare a block spanning d_{j-n+1} through d_j and one spanning d_{j+1} through d_{j+n-1}. The measurement is plotted at the jth position on the x-axis. If either of the comparison

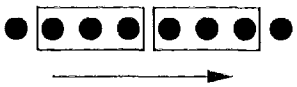

Figure 3: A Moving Paired Window

windows is underfilled, no measurement will be made. The graph that the procedure gives out is smoothed (with an appropriate width) to bring out its global trends. The length of a single segment, i.e., noun counts, varies from text to text, genre to genre, ranging from a few words (a junior high science book) to somewhere around 60 words (a newspaper editorial).

We performed experiments on a four-week collection of editorials from *Nihon Keizai Shimbun*, a Japanese economics newspaper, which contains the total of 1111 sentences with some 10,000 nouns, and 556 discourse segments. The corpus was divided into segmental sets of nouns semi-automatically[6] Coherence was measured for each adjacent pair of segments, using the *Dice* coefficient. It was found that the block size of 10 segments yields good results. Figure 4 shows a coherence graph covering about a week's amount of the corpus. The graph is smoothed with the width of 5. We see that article boundaries (vertical lines) coincide fairly well with major minima on the graph: with only one miss at 65, which falls on a paragraph boundary.

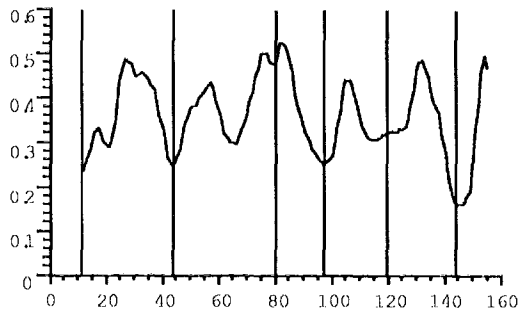

Figure 4: A Dice Analysis

Experiments with various block sizes suggest that the choice of block size relates in some way to the structure of discourse; an increasing block size would extract a more global or general structure of discourse.

Youmans (1991) has suggested a information measurement based on word frequency. It is intended to measure the ebb and flow of 'new information' in discourse. The idea is simply to count the number of new words introduced over a moving interval and produce what he calls a *vocabulary management profile (VMP)*, or measurements at intervals. Now given a discourse $D = \{w_1, \ldots, w_n\}$, the k-th interval of the size Δ is defined by $I_k = \{w_k, \ldots, w_r\}$ where:

$$r = \begin{cases} k + \Delta - 1 & \text{if } k \leq n - \Delta \\ n & \text{otherwise} \end{cases}$$

Measurements are made at intervals I_1 through I_{n-1}.

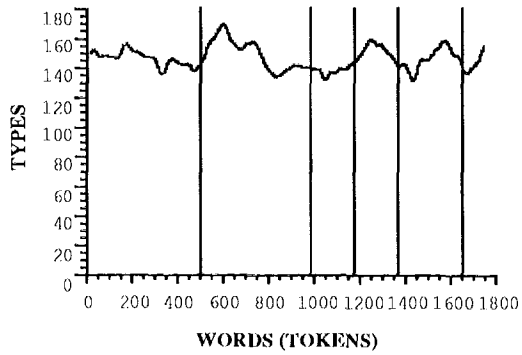

Figure 5: A VMP Analysis

What we like to see is how the scheme compares with

[6]The manual part consists in hand-filtering the corpus to eliminate non-topic marking instances of the particle *wa*, i.e., those that are suffixed to case particles such as *to* (CONJUNCTIVE), *de* (LOCATIVE/INTRUMENTAL), *he* (DIRECTIONAL), *kara* (SOURCE), *ni* (DATIVE), etc., or to a particular form of verbal inflection (*renyou-kei*, i.e. infinitive); thus *wa* is treated as non-topical unless it occurs as a postposition to the bare noun.

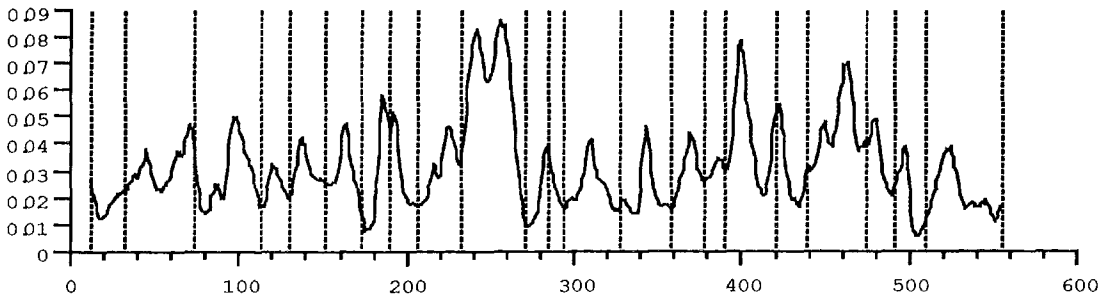

Figure 6: The Dice on the *Nikkei* corpus

ours. Figure 5 shows the results of a VMP analysis for the same nominal collection as above. The interval is set to 300 words, or the average length of a paired window in the previous analysis. The y-axis corresponds to the number of new words (TYPE) and the x-axis to an interval position (TOKEN). As it turns out, the VMP fails to detect any significant pattern in the corpus. One of the problems with the analysis has to do with its generality (Kozima, 1993); a text with many repetitions and/or limited vocabulary would yield a flattened VMP, which, of course, does not tell us much about its inner structurings. Indeed, this could be the case with Figure 5. We suspect that the VMP scheme fares better with a short-term coherency than with a long-term or global coherency.

Evaluation

Figure 6 demonstrates the results of the Dice analysis on the nikkei collection of editorial articles. What we see here is a close correspondence between the Dice curve and the global discourse structure. Evaluation here simply consists of finding major minima on the graph and locating them on the list of those discourse segments which comprise the corpus. The procedure is performed by hand.

Correspondences evaluation has been problematical, since it requires human judgments on how discourse is structured, whose reliability is yet to be demonstrated. It was decided here to use copious boundary indicators such as an article or paragraph break for evaluating matches between the Dice analysis and the discourse. For us, discourse structure reduces to just an orthographic structure[7].

In the figure, article boundaries are marked by dashes. 7 out of 27 local minima are found to be incorrect, which puts the error rate at around 25%. We obtained similar results for the Jaccard and cosine coefficient. A possible improvement would include adjusting the document frequency (*df*) factor for index terms; the average df factor we had for the Nikkei corpus is around 1.6, which is so low as to be negligible[8].

[7] Yet, there is some evidence that an orthographic structure is *linguistically* significant (Fujisawa et al., 1993; Nunberg, 1990).

[8] However, the average df factor would increase in proportion to the growth of corpus size. It is likely, therefore, that with a

Another interesting possibilty is to use an alternative weighting policy, the *weighted inverse document frequency* (Tokunaga and Iwayama, 1994). A *widf* value of a term is its frequency within the document divided by its frequency throughout the entire document collection. The *widf* policy is reported to have a marked advantage over the *idf* for the text categorization task.

Recall and Precision

As with the document analysis, the effectiveness of text segmentation appears to be dictated by recall/precision parameters where:

$$recall = \frac{number\ of\ correct\ boundaries\ retrieved}{total\ number\ of\ correct\ boundaries}$$

$$precision = \frac{number\ of\ correct\ boundaries\ retrieved}{total\ number\ of\ boundaries\ retrieved}$$

A boundary here is meant to be a minimum on the coherence graph. Precision is strongly affected by the size of block or interval[9]; a large-block segmentation yields less boundaries than a small-block segmentation. (Table 1). Experiments were made on the *Nikkei* corpus to examine the effects of the block size parameter on text segmentation. The corpus was divided into equally sized blocks, ranging from 5 to 35 words in length. The window size was kept to 10 blocks. Shown in Figure 7 are the results given in terms of recall and precision. Also, a partitioning task with discourse segments, whose length varies widely, is measured for recall and precision, and the result is represented as G. Averaging recall and precision values for each size gives

Block Size	5	10	15	20	25	30	35
Boundaries	52	25	24	20	21	19	12

Table 1: Block size (in word) and the number of boundaries retrieved.

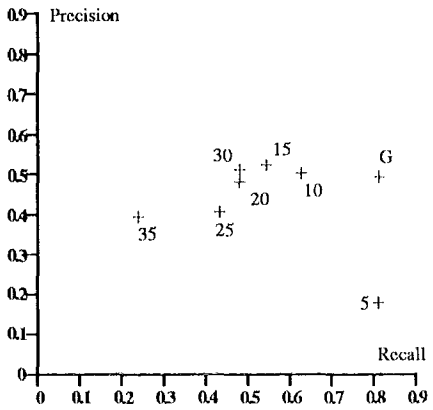

Figure 7: Recall and Precision

an ordering:

$$35 < 25 < 20 < 5 < 30 < 15 < 10 < G.$$

G ranks highest, whose average value, 0.66, is higher than any other. '10' comes second (0.61)[10]. (It is an interesting coincidence that the average length of discourse segments is 13.7 words.) The results demonstrate in quantitative terms the significance of discourse segments.

It is worth pointing out that the method here is rather selective about a level of granularity it detects, namely, that of a news article. It is possible, however, to have a much smaller granularity; as shown in Table 1, decreasing the block size would give a segmentation of a smaller granularity. Still, we chose not to work on fine-grained segmentations because they lack a reliable evaluation metric[11].

Conclusion

In this paper, we have described a method for partitioning an unstructured corpus into coherent textual units. We have adopted the view that discourse consists of contiguous, non-overlapping discourse segments. We have referred to a vector space model for a statistical representation of discourse segment. Coherence between segments is determined by the Dice coefficient with the *tf·idf* term weighting.

We have demonstrated in quantitative terms that the method here is quite successful in discovering articles from the corpus. An interesting question yet to be answered is how the corpus size affects the document frequency and coherence measurements. Another problem has to do with relating the present discussion to rhetorical analyses of discourse.

larger corpus, we might get better results.

[9]Blocks here are intended to mean minimal textual units into which a discourse is divided and for which coherence is measured.

[10]In general, recall is inversely proportionate to precision; a high precision implies a low recall and vice versa.

[11]Passonneau and Litman (1993) reports a psychological study on the human reliability of discourse segmentation.

References

Shinji Fujisawa, Shigeru Masuyama, and Shozo Naito. 1993. An Inspection on Effect of Discourse Contraints pertaining to Ellipsis Supplement in Japanese Sentences. In *Kouen-Ronbun-Shuu 3 (conference papers 3)*. Information Processing Society of Japan. In Japanese.

Barbara Grosz and Candance Sidner. 1986. Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204.

Marti A. Hearst. 1993. TextTiling: A Quantitative Approach to Discourse Segmentation. Sequoia 2000 93/24, University of California, Berkeley.

Jerry R. Hobbs. 1979. Coherence and Corefernce. *Cognitive Science*, 3(1):67–90.

Eduard H. Hovy. 1990. Parsimonious and Profligate Approaches to the Question of Discourse Structure Relations. In *5th ACL Workshop on Natural Language Generation*, Dawson, Pennsylvania.

Hideki Kozima. 1993. Text Segmentation Based on Similarity Between Words. In *Proceedings of the 31st Annual Meeting of the ACL*.

W. C. Mann and S. A. Thompson. 1987. Rhetorical Structure Theory. In L. Polyani, editor, *The Structure of Discourse*. Ablex Publishing Corp., Norwood, NJ.

Yuji Matsumoto, Sadao Kurohashi, Takehito Utsuro, Yutaka Taeki, and Makoto Nagao, 1993. *Japanese Morphological Analysis System JUMAN Manual*. Kyoto University. In Japanese.

Akira Mikami. 1960. *Zou wa Hana ga Nagai (The elephant has a long trunk.)*. Kuroshio Shuppan, Tokyo.

Tadashi Nomoto and Yoshihiko Nitta. 1993. Resolving Zero Anaphora in Japanese. In *ACL Proceedings of Sixth European Conference*, pages 315–321, Utrecht, The Netherlands.

Geoffrey Nunberg. 1990. *The Linguistics of Punctuation*, volume 18 of *CSLI Lecture notes*. CSLI.

Rebecca J. Passonneau and Diane J. Litman. 1993. Intention-based Segmentation: Human Reliability and Correlation with Linguistic Cues. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. The Association for Computational Linguistics. Ohio State University, Columbus, Ohio, USA.

Gerald Salton. 1988. *Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley, Reading, MA.

Takenobu Tokunaga and Makoto Iwayama. 1994. Text Categorization based on Weighted Inverse Document Frequency. unpublished manuscript. submitted to ACM SIGIR 1994.

Gilbert Youmans. 1991. A New Tool for Discourse Analysis:The Vocabulary Management Profile. *Language*, 67:763–789.