Automatic Detection of Discourse Structure by Checking Surface Information in Sentences

Sadao Kurohashi and Makoto Nagao
Dept. of Electrical Engineering, Kyoto University
Yoshida-honmachi, Sakyo, Kyoto, 606, Japan

Abstract

In this paper, we propose an automatic method for detecting discourse structure using a variety of clues existing in the surface information of sentences. We have considered three types of clue information: clue expressions, occurrence of identical/synonymous words/phrases, and similarity between two sentences. Experimental results have shown that, in the case of scientific and technical texts, considerable part of the discourse structure can be estimated by incorporating the three types of clue information, without performing sentence understanding processes which requires giving knowledge to computers.

1 Introduction

To understand a text or dialogue, one must track the **discourse structure** (DS), specifying how sentences are combined and what kind of relations (**coherence relations**) they have. Work on DS has mainly focused on such questions as what kind of knowledge should be employed, and how inference may be performed based on such knowledge (e.g., Grosz and Sidner 1986; Hobbs 1985; Zadrozny and Jensen 1991). However, by examining the current status of work both on automatic extraction and on manual coding of knowledge, detailed knowledge with broad coverage availability to computers is unlikely to be constructed for the present. On the other hand, recent rapid increase in the amount of on-line texts has forced us to analyze not only isolated sentences but also discourses using present available knowledge.

We propose here an automatic method for estimating DS in scientific and technical texts by a variety of keys existing in the surface information of sentences. One important key for DS is clue words (e.g., Cohen 1984; Grosz and Sidner 1986; Reichman 1985). Furthermore, we have considered two more important clues. One is the occurrence of identical/synonymous words/phrases for detecting **topic chaining** or **topic-dominant chaining** relation (Polanyi and Scha 1984); the other is a certain similarity between two sentences for detecting their coordinate relation. The judgment based on such clue information is not absolute but just probable. Therefore, we have incorporated the above mentioned three factors into one evaluation measure to estimate the most plausible DS.

2 Discourse Structure Model and Coherence Relations

Studies of DS have been reported by a large number of researchers (e.g., Cohen 1984; Dalgren 1988; Grosz and Sidner 1986; Hobbs 1985; Mann 1984; Polanyi and Scha 1984; Reichman 1985; Zadrozny and Jensen 1991). What has been commonly suggested is that the DS resulting from the recursive embedding and sequencing of **discourse units** has the form of a tree (discourse history parse tree). However, there has been a variety of definition for discourse units, constituents of the tree, and coherence relations. In this research we have adopted the simplest model in the interest of focusing on how to detect DS automatically. In our model, each sentence is considered a discourse unit, and each node of the discourse history parse tree is a sentence and each link a coherence relation.[1]

Coherence relations existing in a text, as Reichman (1985) pointed out, greatly depend on the genre of the text; narrative, argument, news article, conversation, and scientific report. Among a number of the coherence relations suggested so far, we selected the following set of the relations which accounted for intuitions concerning our target texts, namely scientific and technical texts (`Si` denotes the former sentence and `Sj` the latter).

List : `Si` and `Sj` involve the same or similar events or states, or the same or similar important constituents, like s4-3 and s4-6 in Appendix.

Contrast : `Si` and `Sj` involve contrasting events or states, or contrasting important constituents.

Topic chaining : `Si` and `Sj` have distinct predications about the same topic, like s1-13 and s1-19.

Topic-dominant chaining : A dominant constituent apart from a given topic in `Si` becomes a topic in `Sj`, like s4-4 and s4-5.

Elaboration : `Sj` gives details about a constituent introduced in `Si`, like s1-16 and s1-17.

Reason : `Sj` is the reason for `Si`, like s1-13 and s1-14.

Cause : `Sj` occurs as a result of `Si`, like s1-17 and s1-18.

[1] At present, we regard a sentence marked off by a period as a discourse unit. Coherence relations are existing also between clauses in a sentence. We think our approach examining surface clue information can be adapted to extract their relations, and we intend to extend our system to handle them.

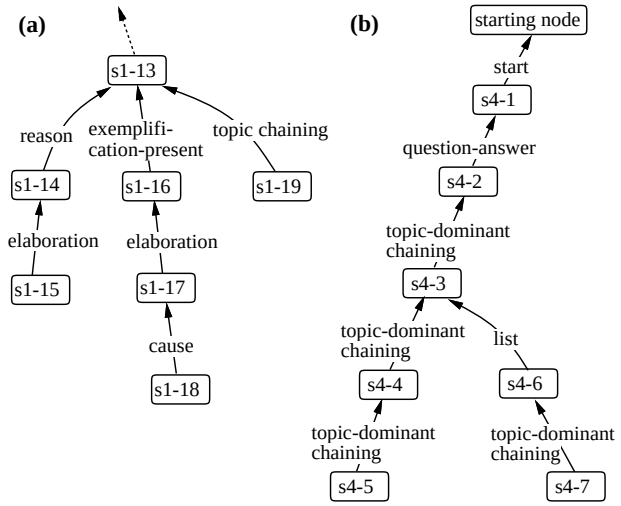


Figure 1: Examples of discourse structures.

Change : The event or state in `Si` changes in `Sj` (usually as time passes).

Exemplification-present : An example of the event, state or constituent in `Si` is introduced in `Sj`, like s1-13 and s1-16.

Exemplification-explain : An example of the event, state or constituent in `Si` is explained in `Sj`.

Question-answer : `Sj` is the answer to the question in `Si`, like s4-1 and s4-2.

The DSs for the sample text in Appendix is shown in Figure 1.

As in many previous approaches, we also make the following assumption in the DS model: a new sentence coming in can be connected to the node on the right most edge in the DS tree (hereafter, we call a new sentence an **NS**, and a possible connected sentence on the right edge in the DS tree a **CS**: Figure 2). This means that, after detailed explanations for one topic terminate, and a new topic is introduced, details of the old topic are hidden in inner nodes and are no longer referred to.

3 Automatic Detection of Discourse Structure

3.1 Outline

Considering our DS model, what the DS analysis should do is clear; for each NS, it tries to find the correct CS and the correct relation between them. In order to estimate them, we have directed our attention to three types of clue information: 1) clue expressions indicating some relations, 2) occurrence of identical/synonymous words/phrases in topic chaining or topic-dominant chaining relation, 3) similarity between two sentences in list or contrast relation. By the method described later we can transform such information into reliable scores for some relations. As an

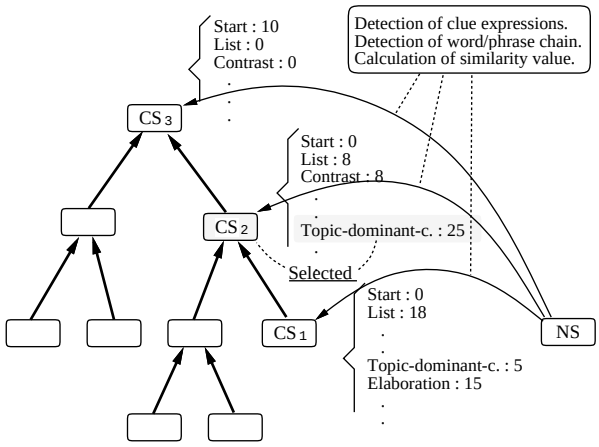


Figure 2: Ranking relations to CSs by three types of clue information.

NS comes in, for each CS we calculate reliable scores for all relations by examining the above three types of clues. As a final result, we choose the CS and the relation having the maximum reliable score (Figure 2).

As an initial state a DS has one node, **starting node**. We always give a certain score for the special relation, **start**, between an NS and the starting node. When any other relation to any CS does not have larger score for an NS, it is connected to the starting node by start relation. This means that the NS is the starting sentence of a new large segment, like paragraph or section, in the DS.

3.2 Detection of Clue Expressions

We prepared heuristic rules for finding clue expressions by pattern matching and relating them to proper relations with reliable scores. A rule consists of the following parts:

- `condition` for rule application :
  - `rule applicable range` (how far in the sequence of CSs the rule can be applied to)
  - `relation of CS to its previous DS`
  - `dependency structure pattern for CS`
  - `dependency structure pattern for NS`
- `corresponding relation` and `reliable score`.

`Patterns for CS and NS` are matched not for word sequences but for dependency structures of both sentences.[2] We use a powerful pattern matching facility for dependency structures, where a wild card matching any partial dependency structure, regular expressions, AND-, OR-, NOT-operators, etc. are available (Murata and Nagao 1993). We apply each rule for the pair of a CS and an NS. If the condition of the rule is satisfied, the specified `reliable score` is given

[2]Input to our system is a sequence of parsed sentences, dependency structures, by our developed parser (Kurohashi and Nagao 1992a). In Japanese the dependency structure of a sentence consists of head/modifier relations between **bunsetsus**, each of which is composed of a content word and suffix words.

Table 1: Examples of heuristic rules for clue expressions.

Rule-1
range : 1
relation of CS : ∗
CS : ∗
NS : NAZE-NARA (because) → ∗
relation : reason
score : 20

Rule-2
range : ∗
relation of CS : ∗
CS : ∗ → X ∗ → ∗
NS : ∗ → X-NO (of) → REI ∗ (example) → ∗
relation : exemplification-present
score : 30

Rule-3
range : 1
relation of CS : exemplification-present
CS : ∗
NS : ∗
relation : elaboration
score : 25

"A → B" denotes a head/modifier relation, where "A" depends on "B".
"∗" denotes a wild card.

to the `corresponding relation` between the CS and the NS.

For example, Rule-1 in Table 1 gives a score to the reason relation between two adjoining sentences (note the `rule applicable range` is '1') if the NS starts with the expression "NAZE-NARA (because)". Rule-2 in Table 1 is applied not only for the neighboring CS but also for farther CSs, by specifying the occurrence of identical words ("X") in the `condition`. We also can specify the `relation of CS to its previous DS` as a condition, like Rule-3 in Table 1. This rule considers the fact that when some examples are introduced by exemplification-present relation, detailed explanations for them often follow.

3.3 Detection of Word/Phrase Chain

In general a sentence can be divided into two parts; a topic part and a non-topic part. When two sentences are in a topic chaining relation, the same topic is maintained through them. Therefore, the occurrence of identical/synonymous words/phrases (the word/phrase chain) in topic parts of two sentences supports this relation. In the case of topic-dominant chaining relation, a dominant constituent introduced in a non-topic part of a prior sentence becomes a topic in a succeeding sentence. So, the word/phrase chain from a non-topic part of a prior sentence to a topic part of a succeeding sentence supports this relation.

However, since there are many clues for an NS supporting other relations to some CSs, we must not only find such word/phrase chains but also give some reliable score to topic chaining or topic-dominant chaining relation. In order to do this, we give scores to words/phrases in topic and non-topic parts according to the degree of their importance in sentences; we also give scores to the matching of identical/synonymous words/phrases according to the degree of their agreement. Then we give these relations the sum of the scores of two chained words/phrases and the score of their matching (Figure 3).

All of these are done by applying rules consisting of a pattern for a partial dependency structure and a score. For example, by Rule-a and b in Table 2, words in a phrase whose head word is followed by a topic marking postposition "WA" are given some scores as topic

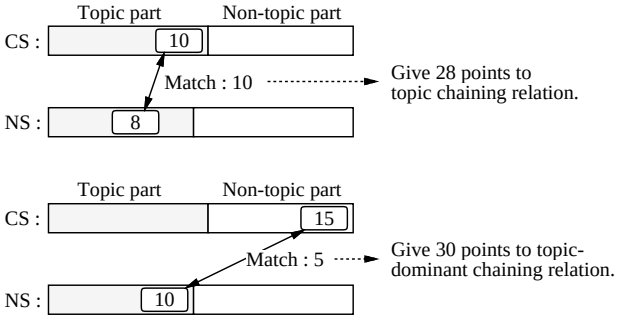


Figure 3: Scores for topic/topic-dominant chaining.

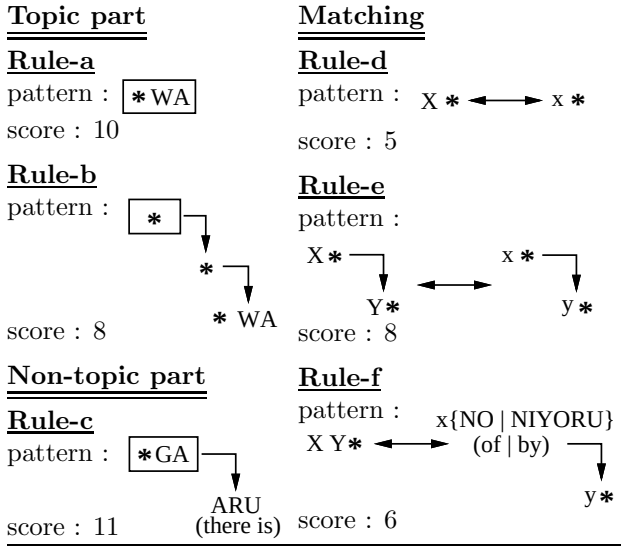
parts. A word in a non-topic part in the sentential style, ".. GA ARU(there is ...)" is given a large score by Rule-c in Table 2 because this word is an important new information in this sentence and topic-dominant chaining relation involving it often occur. Matching of phrases like "A of B" is given a larger score than that of word like "A" alone by Rule-d and e in Table 2. [3]

3.4 Calculation of Similarity between Sentences

When two sentences have list or contrast relation, they have a certain similarity. However, their similarity cannot be detected by rules like the above which see relatively small blocks in sentences, because it is not the simple similarity but the similarity in the sequence of words and their grammatical structures as a whole. In order to measure such a similarity, we extended our dynamic programming method for detecting the scope of a coordination in a sentence (Kurohashi and Nagao 1992b). This method can calculate the overall similarity value between two word-strings of arbitrary lengths. First, the similarity value between two words are calculated according to exact matching, matching of their parts of speech, and their closeness in a thesaurus dictionary. Then, the similarity value between two word-strings are calculated roughly by combining the similarity values between words in the two word-

[3] One difficult problem is that authors often use subtly different expressions, not identical words/phrases, for such chains. While some of them can be caught by using a thesaurus and by rules like Rule-f in Table 2, there is a wide range of variety in their differences. Their complete treatment will be a target of our future work.

Table 2: Examples of rules for topic/non-topic parts and matching.



As for rules for topic/non-topic parts, the score is given to the bunsetsu marked by a square. As for rules for matching, "X" and "x" denote identical words or synonymous words from this Japanese thesaurus, "Bunrui Goi Hyou". So do "Y" and "y".

strings.

While originally we calculated the similarity value between possible conjuncts in a sentence, here we calculate the similarity value between two sentences, a CS and an NS, by this method. This can be done simply by connecting two sentences and calculating the similarity value between two imitative conjuncts consisting of the two sentences. We give the normalized similarity score between a CS and an NS (divided by their average length) to their list and contrast relations as a reliable score.

4 Experiments and Discussion

Experiments of detecting DS were done for nine sections of an article of the popular science journal, "Science", translated into Japanese (Vol.17,No.12 "Advanced Computing for Science", the original is "Scientific American" Vol.257,No.4). For the first three sections, we wrote rules for clue expressions and word/phrase chains, and adjusted their parameters through experimentation. Then we analyzed the remaining six sections by adding rules only for the clue expressions. The analysis results are shown in Table 3. Here the NSs in the text were classified according to their correct relations in connecting to proper CSs. "Success" means that the correct relation and CS were detected for an NS (correct relations and CSs were judged by authors).

Table 3 shows that many clues exist in a text so that much of the DS can be guessed without detailed knowledge. In order to construct rules for clue expressions with broad coverage, we need to consult and analyze a large volume of texts. However, in most cases rules for clue expressions can be written exclusively so that they scarcely interfere with each other. In our experiments, added rules for the remaining six sections had no influence on the analysis of the first three sections.

Table 3: Analysis results.

Relation	Training text (3 sections)		Test Text (6 sections)	
	Success	Failure	Success	Failure
Start	7	1	6	2
List	10	1	15	2
Contrast	6	1	2	2
Topic chaining	13	1	21	5
Topic-dominant c.	10	4	37	14
Elaboration	9	1	9	1
Reason	3	0	1	0
Cause	2	0	6	0
Change	3	0	0	0
Exemp.-present	1	0	0	0
Exemp.-explain	3	0	2	0
Question-answer	1	0	1	0
Total (Success ratio)	68	9 (88%)	100	26 (79%)

The text from s1-13 to s1-19 in Appendix was transformed to the structure in Figure 1-a as follows.

s1-14: the clue expression, "-DAKARA-DEARU" which means "this is because".
s1-15: the clue expression, "-WAKE-DEARU".
s1-16: the clue expression "example of X".
s1-17: the heuristic rule supporting elaboration relation after exemplification-present relation.
s1-18: the clue expression "(SONO)KEKKA(-WA), (the result is)" which corresponds to "lead" in semantics.
s1-19: the chain of "synthetic approach".

The text from s4-1 to s4-7 in Appendix was also transformed to the structure in Figure 1-b as follows.

s4-2: the clue expressions: "-KA" (a suffix indicating an interrogative sentence) in s4-1 and "(the) answer" in s4-2.
s4-3: the chain of "double star".
s4-4: the chain from "shrink" in s4-3 to "this process" in s4-4 (some expressions like "this process" are regarded as matching any verb in a previous sentence).
s4-5: the chain of "nuclear fusion".
s4-6: the large similarity value between s4-3 and s4-6 and the clue expression "similarly".
s4-7: this NS could not be analyzed correctly. List relation with s4-6 was detected incorrectly because of their similarity value.

In s4-6 and s4-7, while the same word "heat" is used in English, the prior "heat" was translated into "ONDO(temperature)-GA JOUSHOU-SURU(rise)" in Japanese. In order to detect the chain for their topic-dominant chaining relation, we must infer that the rising of temperature produce a heat. Such a problem is ignored in this research.

5 Conclusion

We have proposed a method of detecting DS automatically using surface information in sentences: clue expressions, word/phrase chains, and similarity between sentences. In the case of scientific and technical texts, considerable part of the DS can be estimated by incorporating the three types of clue information, without performing sentence understanding processes which requires giving knowledge to computers. This approach can be smoothly integrated with the current NLP systems dealing with large amounts of texts.

References

Cohen, R. (1984). "A Computational Theory of the Function of Clue Words in Argument Understanding." In *Proceedings of 10th COLING*.

Dahlgren, K. (1988). *Naive Semantics for Natural Language Understanding*. Kluwer Academic Publishers.

Grosz, B. J. and Sidner, C. L. (1986). "Attention, Intentions, and the Structures of Discourse." *Computational Linguistics*, 12-3.

Hobbs, J. R. (1985). *On the Coherence and Structure of Discourse*. Technical Report No. CSLI-85-37.

Kurohashi, S. and Nagao, M. (1992a), "A Syntactic Analysis Method of Long Japanese Sentences based on Conjunctive Structures' Detection." (in Japanese), *IPSJ-WGNL* 88-1.

Kurohashi, S. and Nagao, M. (1992b), "Dynamic Programming Method for Analyzing Conjunctive Structures in Japanese." In *Proceedings of 14th COLING*.

Mann, W. C. (1984). "Discourse Structures for Text Generation." In *Proceedings of 10th COLING*.

Murata, M. and Nagao, M. (1993). "Determination of referential property and number of nouns in Japanese sentences for machine translation into English." In *Proceedings of TMI '93*.

Polanyi, L. and Scha, R. (1984). "A syntactic Approach to Discourse Semantics." In *Proceedings of 10th COLING*.

Reichman, R. (1985). *Getting Computers to Talk Like You and Me*. Cambridge, MA, The MIT Press.

Zadrozny, W. and Jensen, K. (1991). "Semantics of Paragraphs." *Computational Linguistics*, 17-2.

Appendix: Sample Text

Title: Advanced Computing for Science

("i" and "j" in si-j denote the section number and the sentence number respectively)

. . .

s1-13: 合成法は、「あるシステムの各部分の間の相互作用の基本過程はわかっているが、当のシステムの細かな構成はわからない」という場合に使われる (The synthetic approach is called for when the fundamental processes of the interactions among the parts of a system are known, but the detailed configuration of the system is not.)

s1-14: それにより、未知の構成を合成によって決定したり、可能な構成を考えて、その結果を試してみることも可能だからである (One can attempt to determine the unknown configuration by synthesis: one can survey the possible configurations and work out the consequences of each.)

s1-15: そうした結果を実験から得られる細かなデータとつき合わせてみれば、観察の結果を最もよく説明できる構成を選ぶことができるわけである (By carefully matching the observable details of the experimental situation with these consequences, one can choose the configuration that best accounts for the observations.)

s1-16: １９世紀以来の合成法の有名な例として、天王星の軌道に見られる不可解な摂動を理解しようとした試みをあげることができる (A famous example of the synthetic approach from the 19th century is the attempt that was made to understand the observed but unexplained perturbations in the orbit of Uranus.)

s1-17: 研究者たちは太陽系に仮想の惑星を加え、満足のいく摂動が得られるまで、その軌道のパラメーターを変化させていった (Investigators added a hypothetical planet to the solar system and varied the parameters of its orbit until a satisfactory reconstruction of the perturbation was found.)

s1-18: その結果は、予想された位置の近くでの海王星の発見という成果に直接結びついたのである (The work led directly to the discovery of Neptune, found near the predicted position.)

s1-19: この合成法が適用できるのは、過去には、比較的単純な場合に限られていた (In the past the synthetic approach was limited to comparatively simple situations.)

. . .

s4-1: 天文学者はこの種の衝突になぜ興味をもつのだろうか (Why are astronomers interested in this kind of collision?)

s4-2: その答えは、"熱"を発生させるのに連星が演じている役割にある (The answer lies in the role of double stars in generating "heat.")

s4-3: 連星と単星が衝突する際、連星は縮んで小さくなり、単星にエネルギーを与え、その周囲の星の集団を温めることがある (In a collision between a double star and a single star, the double star can shrink, transferring energy to the single star and thereby heating the pool of stars around them.)

s4-4: この過程は、原子核が衝突して融合し、より重い原子核になる際、エネルギーを放出する核融合とよく似ている (This process is analogous to nuclear fusion, wherein atomic nuclei collide and fuse into heavier nuclei, releasing energy.)

s4-5: 核融合は、太陽を含む恒星を光らせるメカニズムである (Nuclear fusion is the same phenomenon that makes the stars, including the sun, shine.)

s4-6: また、遭遇によって連星の軌道が縮小し、そのために高密度の星団の中核の温度が上昇することも考えられる (Similarly, orbital shrinkage of double stars induced by encounters can heat the core of dense star clusters.)

s4-7: この熱は、星が絶えず沸騰している星団の表面における熱損失と釣り合うことのできるものである (This heat can balance the losses at the surface of star clusters, where stars boil off continuously.)