Thai Text Coherence Structuring with Coordinating and Subordinating Relations for Text Summarization

Conference Paper · August 2007

DOI: 10.1007/978-3-540-74255-5_34 · Source: DBLP

3 authors, including:

Thana Sukvaree
Sripatum University
3 PUBLICATIONS **9** CITATIONS

SEE PROFILE

Jean Caelen
University of Grenoble
146 PUBLICATIONS **561** CITATIONS

SEE PROFILE

Thai Text Coherence Structuring with Coordinating and Subordinating Relations for Text Summarization

Thana Sukvaree[1], Asanee Kawtrakul[1], and Jean Caelen[2]

[1] Department of Computer Engineering, Kasetsart University, Bangkok, Thailand
thanas_spu@hotmail.com,asanee.kawtrakul@nectec.or.th
[2] Laboratory CLIPS, University of Joseph Fourier, Grenoble Cedex 9, France
jean.caelen@imag.fr

Abstract. Text summarization with the consideration of coherence can be achieved by using discourse processing with the Rhetorical Structure Theory (RST). Additional problems on relational ambiguity may arise, especially in Thai. For example, the use of cue words, i.e. "tae/แต่" (meaning "but"), can be identified as a contrast relation or an elaboration relation. Therefore, we propose the reduction of the ambiguity level by reducing the relation types to two, namely Coordinating and Subordinating relation. Our framework is to concentrate on coherence structuring which requires the following 3 steps: (1) identify an attachment point for an incoming discourse unit by using our Adaptive Right-frontier algorithm; (2) extract Coordinating and Subordinating relations through the identification of linguistic coherence features in the lexical and phrasal level, using Bayesian techniques; (3) construct coherence tree structures, The accuracy is 70.45% for the first step, 77.47% and 79.89% for COR and SUBR extraction respectively in the second step and 64.94% in constructing coherent tree of the third.

1 Introduction

Nowadays, there is a high need for automatic text summarization, which is recognized as one of the solutions to tackle the problem of overwhelming amounts of available information. Research in automatic text summarization still requires further developments. Recent research works on the matter focused on the identification of more relevant information from the text to project in the summary. There are two main approaches in achieving this problem: statistical-based approach and knowledge-based one. The statistical-based approach[1,2] often gives incoherent results, which causes further misunderstandings by humans. By contrast, the knowledge-based approach takes this problem into consideration and uses methods of salience extraction from the structure of text representation which is expressed in the form of discourse relation, but requires strong knowledge in creating the text structure[3,4] with the assumption of the result from the salience extraction remaining the same from the source text.

B. Kokinov et al. (Eds.): CONTEXT 2007, LNAI 4635, pp. 453–466, 2007.

In various research papers in summarization, the Rhetorical Structure Theory (RST)[5], is often used to extract salience through the application of knowledge-based approach at discourse level[3,4]. This theory includes explanations on the occurrence of discourse relations and text generation by using tree structures. However, if the RST is used, problems of relation ambiguity would have to be accounted for[6,7]. One cause of ambiguity is that there are too many rhetorical relations to be classified accurately because the definitions are rather vague and do not provide concrete linguistic criteria to look out for in text. Therefore, in our research, we reduce the number of discourse relations to only two Coordinating relation and Subordinating relation according to the Segmented Discourse Representation Theory (SDRT)[8] . Furthermore, we propose a method for constructing a simplified discourse structure by using the right frontier constraint of Polanyi[9] together with the discourse dependency function called "adaptive right frontier". This should produce a discourse structure sufficient to increase the quality of text summarization.

In Section 2 we give a brief overview of the theoretical concepts which bear on our topic. In section 3, we argue the crucial problems in the construction of the discourse trees. Section 4 presents our solution and we describe experiments and results in Section 5. In section 6, we show that the outcome of this research has positive effects on text summarization. Our conclusions are summarized in Section 7.

2 Preliminary

Mann and Thompson[5] have introduced the RST by classifying rhetorical relations (RR) into two categories, namely a paratactic (multi-nucleus) relation and a hypotactic (mono-nucleus) relation[3]. A paratactic relation is defined to be a relationship that exists between two discourse units (du) which have the same value of interest. A hypotactic relation can be defined as a relationship that exists between discourse units where the value of interest is inequivalent. These two types of relation play a role as nuclearity functions for salience extraction in text summarization. By a Nucleus (N), we mean a discourse unit which is more important or has more value of interest. A Satellite (S) is a discourse unit that contains less value of interest. See example in Text-1 and Fig. 1.

Example-1 [Text-1]

S1: The Brown hopper likes to live in the bottom area of the rice plant
S2: and infest the sap in that area
S3: rice plant shows symptoms of dried leaf as if it has been boiled or burnt
S4: which is called "symptom of inconsistent burning".

Because RST proposes too many possible relation such as Elaboration, Explanation, Cause-result, Conditional, Contrast, Sequence, Consequence, Joint, List, Background etc., it is difficult to specify the relations to generate a rhetorical structure. The root of this problem stems from the vagueness of the definition

Table 1. Rhetorical relations classified as paratactic relations and hypotactic relations

Types	Rhetorical Relations
Paratactic (COR)	List, Joint, Cause-Result, Problem-Solution, Topic-Shift, Contrast, ...
Hypotactic (SUBR)	Elaboration, Background, Justify, Evidence, Condition, Explanation, Consequence, Question-Answer, ...

Fig. 1. Paratactic and hypotactic relations in RST

of RST relations. Thus, in this research, we propose to reduce the number of relations to only two, namely subordinating and coordinating relations. These two relations have been well studied in [9,8], A relation $R(\alpha, \beta)$ between α and β is called subordinating if β adds something to what is said in α so that the information expressed by β is in a sense more granular than the information expressed by α. If $R(\alpha, \beta)$ is not subordinating relation then it is the coordinating relation. In addition to these two relations, we would also take nuclearity into account.

If RR is the main set for this discourse rhetorical relation and if we work in accordance with the RST theory, then we would have the following set:

RR = {Elaboration, Cause-Result, Condition, List, Joint, Contrast, Explanation, Evidence, ...}

When considering the way in which we decide over the subsets of RR to be either the Coordinating relation (COR) or Subordinating relation (SUBR), we can make the following subsets (see Table 1).

We obtain two relations: coordination and subordination (Fig. 2.), including all the rhetorical relations in the RST[**?**]. Thus this concept transforms a space of intentional relations (RST) to the space of coherence relations in the SDRT.

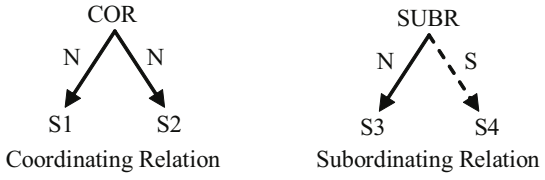

Fig. 2. Coordinating and Subordinating relations with nuclearity

It has the advantage to reduce the complexity of the RR interpretation, e.g. transforming Fig. 1 into Fig. 2. We will elaborate this concept later.

In general, when we take the discourse unit in the tree structure with two discourse relations, following the RST theory, we call this COR&SUBR-tree. The leaves of the tree (the tree end-nodes) comprise of grammatical structures that are elementary discourse units[3] and the internal parts of the tree structure contain relations that exist between the small discourse unit and the complex discourse segment.

3 Problems in the Construction of the COR&SUBR-Tree

The construction of the COR&SUBR-tree generally consists of 3 steps. The first one is to locate the incoming node so that it suits best once attached to the previous COR&SUBR-Tree. The second step is to interpret the relations existing in the text. And the third step is to integrate two previous steps to build up coherent tree. Problems occurring in each step are discussed in the next section.

3.1 Identifying the Connection Between an Incoming Node and the Previous Discourse Tree

This problem is from the point of view of matching, a problem of incoming node (INC) considered to be a part of previous discourse tree (PDT) as in Fig. 3. A possible Attachment Point (AP) that connects the incoming node (INC) with PDT tree consists of AP1, AP2, AP3, AP4 and AP5. Different attachment points will result in different discourse tree structures. This will affect the extract of discourse tree which will cause incorrect text summary.

When considering time complexity of all possible positions, we see that it is in $O(n^2)$. However, this problem can be solved by using Right Frontier algorithm[9] together with linguistic information which decreases time complexity into linear order that we will discuss later.

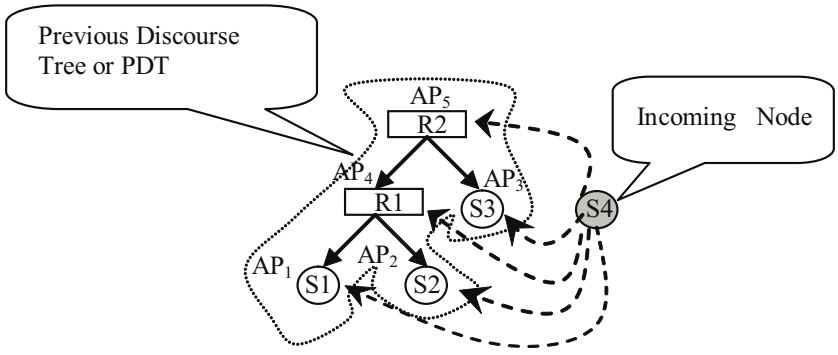

Fig. 3. The possibly APs for S4 attach to PDT

3.2 Problem of Ambiguity in Interpretation of the Discourse Relation

Interpretation of the discourse relation can be in various forms[3,5,6,7]. For example, there are two ways to interpret the discourse in text-1: [S1, S2] is the cause of [S3, S4] which denoted by CR[JT[S1, S], EB[S3, S4]] or the other way is [S3, S4] is a consequence of [S1, S2] which denoted by CSQ[JT[S1,S2], EB[S3,S4]].

Although, T1 and T2 are the same structure but the top label relations of each tree are different in nuclearity, CR (Cause-Result) is multi-nucleus, CSQ (Consequence) is mono-nucleus. If we extract salience from these then will be produced summary in different set.

Our work proposes a solution to those problems by the reduction of relations using a transformation from an n-dimensional space of traditional rhetorical relations into a 2-dimensional space of COR and SUBR. We classify any paratactic relations in RST to coordinating relation in SDRT and classify the hypotactic relations in RST to subordinating relation in SDRT.

The problems of attachment point and ambiguity of discourse relation will affect the extract of salience of text summary from Fig. 3. and Fig. 4. instead of RST tree. The former problem will create different tree structures. The latter will affect nuclearity of RST tree specification which is essential for selecting salience from RST tree. In order to generate the coherent tree, we propose solution to the above problems in the next section.

4 Solution

We describe our solution to discourse tree generation in 2 parts.

4.1 Identifying the Attachment Point of an Appropriate Incoming Node to PDT

This problem is an attachment point problem. The traditional Right Frontier Constraint: RFC[9] only concerns with the anaphoric pronoun. Consequently, RFC cannot solve some of our problems. For example, an incoming EDU (elementary discourse unit[3]) becomes a new topic of content or the segment, and the attachment point does not locate in the right frontier area in PDT[10] Also

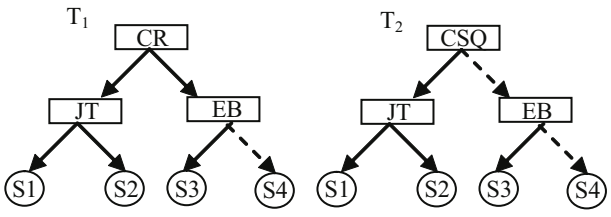

Fig. 4. The problems of multiple interpretations between textual [S1, S2] and [S3, S4]

it could happen that the incoming discourse unit does not have any anaphoric pronoun. Therefore, we propose the algorithm of using Adaptive Right Frontier (ARF) which combines the discourse dependency function (DDF) with the RFC in order to cover these three phenomena.

The DDF is considered having 3 linguistic parameters: Content Relevant (CV), Discourse Marker (DM) and Anaphoric Expression (AE). We use a linear combination model to formulate this function. Let x be a node in PDT, and y be the incoming discourse unit. We define:

$$DDF(x,y) = \alpha CV(x,y) + \beta DM(x,y) + \gamma AE(x,y) \quad (1)$$

where α, β and γ are weighting constants. To determine whether the incoming discourse unit is a new segment or not, the node in PDT that has maximum value is tested against a predefined threshold, δ. Specifically, if $max_{x \in PDT} DDF(x,y) \geq \delta$, the incoming node will be a node in PDT; otherwise, it will becomes a new segment.

How to compute the parameter of the coherence dependency function: CV, DM and AE

CV (Content Relevance)
Content Relevance is a measurement of the relationship between discourse unit pair whose value is computed form discourse context. Conceptually, we want the high value of CV to mean that discourse unit pair describes the same topic and, vice versa, the low value of CV to mean that they describe different topics. Thus, the value of CV is one component to decide whether the new coming discourse unit should create a new segment or not. To compute this, we define

$$CV(x,y) = sim(f_x, f_y)$$

where f_x and f_y are the foci of the discourse unit x in PDT and the incoming discourse unit discourse unit y respectively. According to B.J. Grosz[11], we use the NP (noun phrase) that precedes the main verb or the agent which has the highest potential of the discourse entities as a focus of the discourse unit. The similarity between the foci f_x and f_y is then computed by using the cosine of the angle between the word vectors in the vector space model[12]. For a fixed collection of corpus, a m-dimensional vector is generated for each word, where m is the number of unique noun word in the corpus. The weight associated with each noun word is calculated based on the number of occurrence of word w_i and w_j in the same k-consecutive EDUs. The reason of this is that the consecutive EDUs usually describe the same topic, and, in this paper, k is set to 3 which is the average length of EDU span that has the same discourse topic.

Example-2: [Text-2]

S1: The Brown hopper causes Blast disease,
S2: especially, this disease always occurs in summer.

S3: spay the insecticide only in the morning,
S4: don't spray the insecticide in the evening.

To compute the word similarity in example 2, we firstly create a set of important words of the corpus which are "Brown hopper", "Blast disease", "summer", "Insecticide", "morning", and "evening". Then, the 6-dimentional vector is generated for each word w_i. The value of j-th dimension is generated by counting the number of time w_i appear together with w_j in the same k-consecutive EDUs; for example, the number of time the word 'summer' appears in the same 3-consecutive EDUs with the word 'morning' is 2 (S1:S2:S3 and S2:S3:S4) and the number of time the word 'Brown hopper' appears in the same 3-consecutive EDUs with the word 'evening' is 0. After obtaining the word vector for all words, we can calculate the similarity between word w_i and w_j by using the inner product of the word vector w_i and w_j, which can be computed from the equation:

$$sim(w_i, w_j) = \frac{\sum_{k=1}^{m} w_i[k]\dot{w}_j[k]}{\sqrt{\sum_{k=1}^{m} w_i^2[k]}\sqrt{\sum_{k=1}^{m} w_j^2[k]}}$$

DM (Discourse Marker)
DM is the connective device for discourses which plays two roles: the cohesive device and the inference of semantic relation. In this section, we use DM in the first role as the traffic policeman to point out the attachment point of INC. If DM has the signal for the left hand side attach to PDT, such as "(The Brown hopper likes to live in the bottom area of the rice plant)$_{EDU1}$ (**and** infest the sap in that area)$_{EDU2}$", then its value is +1. On the other hand, if the DM has the signal to create a new segment such as "(**In conclusion**, agriculture commodity and food standards prepared by ACFS can be applied by all stockholders)$_{EDU1}$ (which can benefit all parties concerned in the industry and the economy as a whole)$_{EDU2}$", then its value is -1. When the DM can not be used to guide the direction of the attachment point, the value will be set to zero. To formulate this, we analysis the corpus and create a set of discourse marker that can be used to guide the attachment point direction namely DM_{left} and DM_{right}. The function used to determine the value of the output is defined as in Equation 2.

$$DM(x, y) = \begin{cases} +1 & \text{if discourse unit } y \text{ contain a discourse marker in } \text{DM}_{left}; \\ -1 & \text{if discourse unit } y \text{ contain a discourse marker in } \text{DM}_{right}; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where DM_{left} = {ซึ่ง:which, โดย:by means of, และ:and, หรือ:or, ...}
DM_{right} = {ในที่สุด:finally, สรุปได้ว่า:conclusion, โดยปกติ:normally, ...}

AE (Anaphoric Expression)
AE value describes the strength of the relation between discourse units using anaphoric expression. The intuition of this is that the unit should be related if

they describe the same thing. Specifically, the unit that contains the antecedent of the anaphor of the incoming unit should be more related with the incoming unit than the unit that does not contain any antecedent. To compute this, we use anaphora resolution described in [13] to identify antecedent of y. and the value of AE is compute from the number of antecedent of anaphor in y that exists in x.

$$AE(x, y) = \text{the number of antecedent of anaphor in } y \text{ that exists in } x \quad (3)$$

Identifying the attachment point of an incoming discourse unit
Finally, we propose the Adaptive RFC for computing the coherence value to decide the attachment point of the incoming discourse unit in (4)

$$AdaptiveRFC(x, y) = RFC(x, y) + DDF(x, y) \quad (4)$$

where the RFC is the ranking value of the rightmost node of PDT. The RFC value decreases when it is as long as the distance from the bottom of the rightmost of the PDT. The attachment point of an incoming discourse unit is the unit that has maximum adaptive RFC value.

We tested 200 document files about plant diseases in agricultural domain which has 126 discourse markers (COR/SUBR). We used the threshold ? at 0.057 and the coefficient value of (2) with (0.88, 0.47, and 0.55). Maximum Likelihood Estimation (MLE) was used to compute the coefficient numbers of DDF and we adjust a suitable threshold with trial and error method. The accuracy of result in this section can be seen in the experiment and result section. Then, we follow this algorithm in Fig. 5.

```
Build_Up_Tree(text_seg){
 inc_edu = Get_EDU(text_seg)
 while ( inc_edu <> null) {
  If ( tree == null) then   tree <- inc_edu; exit();
  else
     inc_edu = current;
     RFC_nodes = RFC( PDT );
     RFC_node.area = top;
     while   RFC_node.area  > bottom
    /* ranking the accessible value to the right frontier node */
        RFC_node.accesible++ ;
      if  max(ARFC(RFC_nodes,inc_edu)) > threadhold
        Attach(RFC_nodes,inc_edu)
                    /* add  INC to the new node of PDT */
      else { /* introduce the new segment */
        Tree_Space = current ;  /*save the old tree */
        inc_node = current;  /*set as the new tree */
      } /*  if-max */
 }  /* while-inc_edu */
} /* Build_Up_Tree */
```

Fig. 5. Local Tree construction

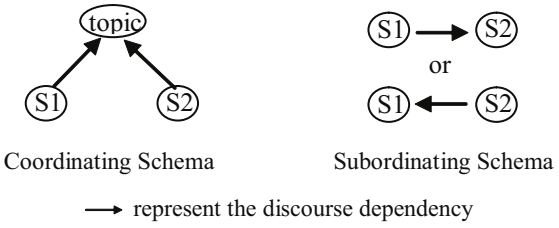

Fig. 6. COR/SUBR relationship model

4.2 Discourse Relations Disambiguation by Reducing Transformations Approach

This approach is to use coordinating relation and subordinating relation which have common nuclearity as a connected relation to generate discourse tree. This gives a transformation from a problem in an n-dimensional space of relations in RST to the 2-dimensional space of Coordinating and Subordinating relations which have common nuclearity.

Let RR^n be the set of rhetorical relations having n relations r_k , k = 1,2, ,n, where r_k is in R. That is

$$PR^n = \{r_k \in R, k = 1, 2, ..., n\}. \text{ Let } RR^2 = \{r_c = COR, r_s = SUBR\}$$

RR^2 is set of 2 dimension space of 2 semantic relations; COR/SUBR relation. Therefore, they are in accordance with Satisfaction-Precedence/Dominant relation[**?**] and Coordinating relation/Subordinating relation[9,**?**] and we denote it with COR and SUBR. To simplify the resolution of the ambiguity problem in n-dimension, we consider just 2 dimensions of COR and SUBR.
We define COR/SUBR properties with nuclearity function of RST.

Definition: Coordinating relation means the relation between discourse topics having the same important interrelated events or objects.

Definition: Subordinating relation means the relation between entity and interrelated proceeding in the form that one proceeding event or object depends on the other. This definition can be explained as shown in Fig. 6.

COR & SUBR recognition
We use Naive Bayes classifier to classify the COR/SUBR relations and Thai discourse cues defined in [14] Semantic noun phrase discourse entities will be used as the features in this learning process. For the discourse cues, considered as the discourse marker in [3], are separated into two groups corresponding with RST's nuclearity as DMCOR , DMSUBR . For example,

DM_{COR} = {"และ:and" , "หรือ:or", "แต่:but", "ถ้า-แล้ว:if-then", ...},
DM_{SUBR} = {"ซึ่ง:which" , "โดย:by", "ดังนั้น:therefore", ...}

However, it is not necessary that there be a discourse marker within a discourse unit. Therefore, we apply DM and the semantic noun phrase discourse entities for COR/SUBR relation recognition. If the discourse entity of the discourse unit pair has hyponym relation in WordNet, then the discourse unit pair has a subordinating relation; otherwise, it is a coordinating relation. For the example, [S1: Blast disease can commonly spread to every parts of Thailand] [S2: It caused by fungus called Pyricularia], the following example, the noun phrase discourse entities are Blast disease, part of Thailand in S1 and fungus in S2. We found that the similarity value of Blast disease - fungus is greater than the other pair. These features computed by supervised learning technique whose annotated data are separated into two sets of COR and SUBR. Each set has 1000 EDU pairs in agricultural domain, and the testing data of each set are 300 EDU pairs. Naive Bayes is applied to calculate the weight of each individual feature. The results of precision and recall are evaluated by an expert. The precision of this experiment is 79/77% for COR/SUB and 76/83% for recall.

4.3 Coherent Tree Construction

As mentioned about local tree construction in section 4.1 and 4.2, we span coherent tree as shown in Fig. 5. We use the bottom-up algorithm to merging local PDT_i where $i = 1$ to n. The local PDT was generated and store into Tree Space in Fig. 5. Then, the discourse relations {COR, SUBR} are identified between local PDT_i and local PDT_j by repeating step 4.2, where the input can be multiple EDU which was selected by nucleus, from leaf to root node of local PDT, to represent the local PDT. For example,

Example-3: [Text-3]

S1: Blast disease can commonly spread to every parts of Thailand
S2: this disease caused by fungus called Pyricularia
S3: which the conidia of this fungus can be blown by the wind
S4: therefore blast disease distribute its through the wind
S5: when the conidia of the fungus settle on various parts, rice that are highly moist
S6: it will sprout in a fiber form, destroying the plant

Fig. 7, illustrates how the representative of local PDT can be derived. The local PDT is the output of algorithm in Fig. 5. From section 2, the nuclearity property has two statuses {N: nucleus denote with line arrow and S: satellite denote with dash line arrow}. We can use nucleus property of discourse relation to decide the representative of individual discourse unit pair. The PDT_j , the discourse unit pair [S2, S3] has relationship with subordinating relation; S2 has a nucleus status and S3 has a satellite status, thus we select S2 as the repres=entative of [S2, S3], denoted with {S2} as parent of [S2, S3]. In the same method, we consider discourse unit pair [S2,S3] and [S4] as subordinating relation, also the {S2} was selected to represent the discourse unit pair [S2, S3] and [S4]. Next, [S1] and [S2, S4] is consecutive discourse unit pair which are

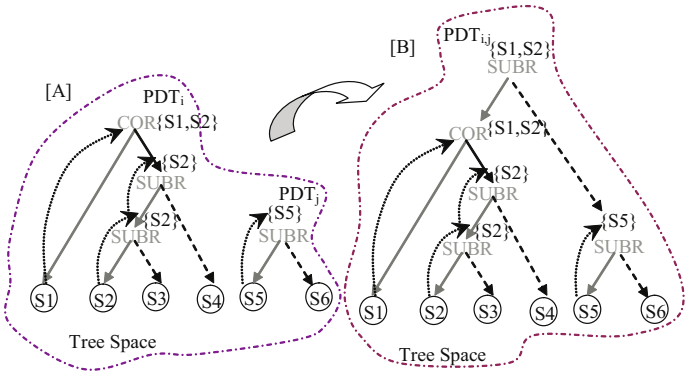

Fig. 7. [A] Representative of local PDT, selected by nucleus property of discourse relations and [B] local PDT Merging

coordinating relation; [S1] and [S2,S4] has a nuclearity property as a nucleus. Therefore, the representative of [S1] and [S2,S4] is {S1,S2}. In the same way, PDT_j , {S5} will be the representative of discourse unit pair [S5,S6]. Finally, we compute the discourse relationship between PDT_i and PDT_j with the method in section 4.2 and the $PDT_{i,j}$ will be generated as a result of this step.

The experiment was maintained under the condition that the input document has one topic which is a plain text. Therefore, this experiment does not cover the case of a long text. And its result will be reported in the next section.

5 Experiment and Result

The experiments were done by testing 200 document files of plant's disease symptoms in agricultural domain. The average edu length is 17.34 words with the total number of 3320 EDUs. The measurement consists three parts. In part-1, we consider the accuracy of the position of an attachment point of incoming EDU. In part-2, we measure the average of accuracy of relation classification between COR and SUBR relations. Finally, we measure accuracy of full coherent tree spanning. In each part, we align the result generated by the system with the result produced by 3 persons, two of which are linguist and another is not. The result is shown in table 2, where PR represents the precision measurement and RC represents the recall measurement. As the F-score of the system in part-1 has an average accuracy of 70.45%, we verify that the causes of almost all errors in this part have been caused by a nonadjacent incoming EDU that crosses over on the left hand side more than 2 distance units. In part-2, there are two numeric results. We interpret this as a result of inadequate features used in the experiment. There are not enough features to distinguish some cases such as cue's ambiguity. Finally, in part-3, we expect that its error results from the propagation error in the previous step. However, the result of the coherent tree spanning is nearly 65% of F-score.

Table 2. Result of experiment in three parts

Level		Part-1: AP	Part-2: COR/SUBR	Part-3: Tree Spanning
System	PR	73	79 / 77	67
	RC	69	76 / 83	63
Human	PR	87	93 / 90	84
	RC	86	90 / 88	82
F-Score of System		70.45	77.47/79.89	64.94

Furthermore, the result of our system is compared with the simple baseline system, based on the answer set from the human judgments. There are three measurements; first is the accuracy of the attachment point, evaluated by counting the newly posited INC at the rightmost branch of the PDT, is 66%, which is below F-Score of the system by 6.74%. Second is the measurement of discourse relations identification (the coordinating relation and the subordinating relation), where any internal nodes are subordinating relation, since this relation mostly occurs more than 79% in corpus. Therefore, only the PR and the RC from subordinating relations are determined as 63% for the PR and 57% for the RC, which is below F-Score of the system by 22.97% and 40.16%, respectively. Third is the determination of the accuracy of the discourse tree spanning process which the baseline system has 49% accuracy, which is below F-Score of the system by 32.53%. In other word, our system have significant outperforms the baseline system in all parts.

6 Application of Coordinating and Subordinating Relations in Text Summarization

After we have generated a coordinating and subordinating structure containing the nuclearity within each a relationship, it is easy to extract the salience unit by using the beam search algorithm based on the nucleus-satellite property and breadth first search policy. For the example, from the nuclearity function we can determine the location of salience in Text-4 as shown in the corresponding leaf nodes of Fig. 8. Then we apply the beam search to extract the salience nodes, as in S1, S6, S7, from the Coherent Tree, T4 (stands for Text-4). By the result of this beam search, the short summary text {S1, S6, and S7} is generated in the first step. The second step: {S2, S3} are added into the previous summary text, following by {S4, S5} and so on, depending on the user desire of the length of summary. However, the coherent characteristic is still in the summary produced.

Example-4 [Text-4]

S1: Soft Rot disease found in almost every growth step,
S2: especially, once lettuces start to bulb.

S3: Initially, softening and water-soaking spots or scales are found.
S4: Afterward, a wound progress widespread
S5: and they cause slimy softening rot with bad smell.
S6: When the disease becomes severe, lettuces are whole-bulb rotten
S7: and their necks become soft when pressed.

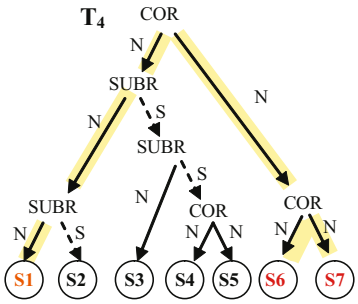

Fig. 8. Salience nodes in the first step of salience extraction on Coherent Tree

7 Conclusion

Our algorithm for the text summarization problem gives a construction of the coherent tree with two COR/SUBR relations. We concentrated in finding a simple method to compute a coherent structure. Consequently, we reduced the n-dimensional space of RST to the 2-dimensional space of COR and SUBR. In this way, we found that evident of features can be increasing more than the individual relation in RST. However, we show how to modify the ARF with forward-backward constrain, which mentions about the closing off status of the previous EDU and a new topic signal of the incoming EDU. These are two important steps before generating the coherent tree. Bayesian techniques are used as tools for recognizing the relations. The results of experiment seems satisfied with our corpus.

The main problem of constructing the coherent tree with COR/SUBR relations in this experiment may result from propagation error in the preceding step. So, we will tune them by adding a temporal feature that relates to positive error significantly.

Acknowledgements

The work described in this paper has been supported by the grant of Franco-Thai project and partially supported by a grant from NECTEC No. NT-B-22-14-12-46-06. The authors would like to present deeply thanks to Jean Caelen to review this work.

References

1. Edmundson, H.P.: New Method in Automatic Extracting. ACM 16(2), 264–285 (1969)
2. Hovy, E., Lin, C.: Automated text summarization in summarist. In: Proceedings of the Workshop on Intelligent Scalable Text Summarization, pp. 18–24 (1977)
3. Marcu, D.: The rhetorical parsing of natural language texts. In: Meeting of the Association for Computational Linguistics, pp. 96–103 (1997)
4. Cristea, D., Postolache, O., Pistol, I.: Summarisation through discourse structure [15], pp. 632–644
5. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Toward a functional theory of text organization. Text 8(3), 243–281 (1998)
6. Moore, J.D., Pollack, M.E.: A problem for RST: The need for multi-level discourse analysis. Computational Linguistics 18(4), 537–544 (1992)
7. Hovy, E., Maier, E.: Parsimonious or profligate: How many and which discourse structure relations. In: Discourse Processes, pp. 18–24 (1977)
8. Asher, N., Lascarides, A.: Logics of Conversation. Studies in Natural Language Processing. Cambridge University Press, Cambridge (2005)
9. Polanyi, L.: A formal model of the structure of discourse. Journal of Pragmatics 12, 601–638 (1988)
10. Sassen, C., Kühnlein, P.: The right frontier constraint as conditional [15], pp. 222–225
11. Grosz, B.J., Joshi, A.K., Weinstein, S.: Centering: A framework for modeling the local coherence of discourse. Computational Linguistics 21(2), 203–225 (1995)
12. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18(11), 613–620 (1975)
13. Kongwa, A., Kawtrakul, A.: Know-what: A development of object-property extraction from thai texts and query system. In: Proceeding of the Sixth Symposium on Natural Language Processing (2005)
14. Wattanamethanont, M., T.S., Kawtrakul, A.: Thai discourse relations recognition by using naive bayes classifier. In: The Proceedings of the Sixth Symposium on Natural Language Processing (2005)
15. Gelbukh, A. (ed.): Computational Linguistics and Intelligent Text Processing. In: Gelbukh, A. (ed.) CICLing 2005. LNCS, vol. 3406, Springer, Heidelberg (2005)