Building an Annotated Corpus
for Text Summarization and Question Answering
Patcharee Varasai, Chaveevan Pechsiri, Thana Sukvari
Vee Satayamas and Asanee Kawtrakul

Specialty Research Unit of Natural Language Processing
and Intelligent Information System Technology (NAiST)
Department of Computer Engineering,
Faculty of Engineering,
Kasetsart University, Thailand

E-mail: {patcha_matsu, itdpu, thanas_sup, libvee}@hotmail.com, asanee.kawtrakul@nectec.or.th

Abstract
We describe ongoing work in semi-automatic annotating corpus, with the goal to answer “why” question in question answering system and give a construction of the coherent tree for text summarization. In this paper we present annotation schemas for identifying the discourse relations that hold between the parts of text as well as the particular textual of span that are related via the discourse relation. Furthermore, we address several tasks in building the annotated corpus in discourse level, namely creating annotated guidelines, ensuring annotation accuracy and evaluating.

1. Introduction
Annotated corpus in discourse level plays an important part in natural language systems performing tasks such as text summarization, question-answering systems and knowledge mining. However, the types of discourse annotation that are relevant vary widely depending on the application. Therefore, empirical analysis is necessary to determine commonalities in the variations of discourse phenomena and to develop general-purpose algorithms for discourse analysis.

In various researches in summarization, the Rhetorical Structure Theory (RST)( Mann, W.C and Thompson, S.A, 1998), are often used to extract salience through the application of knowledge based approach at discourse level(Marcu, D,1997; Cristea, D.et al.,2005). Within this field, annotated corpus is mainly used for machine learning to learn patterns for extracting important information from text as well as for the more complex task of evaluation of summarization methods (Edmundson, H.P. 1969; Kupiec, J., Pederson, J. and Chen, F. 1995; Marcu, D. 1997). To annotate corpora, one (accurate) method is to employ humans to indicate those parts of text to be annotated with whatever information necessary. These human-selected units of text can then be used as a gold standard by which to measure the performance of a system, as well as for discerning which types of units are chosen or discarded by humans during the summarization process. There are semi-automatic (Orasan, C, 2002) and automatic (Jing, H. and McKeown, K., 1999; Marcu, D. 1999) annotation providing the marked information from tags at a discourse level. Despite the fact that they are vital to the field, corpora annotated for summarization are relatively sparse, and those resources, which do exist, do not contain as much information as they could. Lynn Carlson (2003) develops a discourse-annotated corpus in the framework of Rhetorical Structure Theory. However, the semi-annotation is the most appropriate here because the necessary information will be provided.

Automatically extracting causality knowledge to provide a person with explanations in question-answering systems is a very challenging task. Girju and Moldovan (2002) use lexico-syntactic patterns to extract causality within one sentence. Chang and Choi (2004) proposed using word pairs and cue phrase probabilities to classify the causality occurring in one sentence. Marcu and Echihabi (2002) presented the unsupervised approach to recognize the discourse relations by using word pair probabilities between two adjacent sentences. Inui et al, (2004) acquired causal knowledge by using discourse markers or connective marks between two adjacent sentences. Furthermore, Torisawa (2003) extracted rules for reasoning from coordinate verb phrases from two adjacent sentences. In our research, causality extraction is based on multiple EDUs (Elementary Discourse Units, stated by Chareonsuk J.et al., (2005)), and expressed as a simple sentence or a clause).

This paper proposes discourse annotation schemas that account for extracting coordinate and subordinate relations and constructing coherence tree for text summarization (Sukvaree, T, et al.2007) and Verb-pairs rules for mining causality of automatic question answering system (QA) in answering ‘why’ question (Pechsiri, C, 2007). Annotation guidelines are also described. In the section 3, we describe the annotation process and quality control in section 4. Section 5 comprises our results, followed by conclusion.

2. A Design of Annotated Corpus
The corpus consists of 500 articles from the online official technological documents related to plant disease (http://www.doa.go.th/), the native technological journal
Agriculturist should put fertilizer [such as Urea] into plot. The cabbage that was destroyed will have yellow color. [such as Pathumtani]

2.1 Corpus Preparation

The corpus is classified into two sets; narrative full text using for Text summarization and Information extraction and fragmented text for QA. The preparation involves using Thai word segmentation tool to solve a boundary of a Thai word (Sudprasert and Kawtrakul, 2003) and using POS tagger, including Name entity Recognition (Chanlekha and Kawtrakul, 2004), and Word-formation Recognition (Pengphong, et al. 2002) to solve the boundary of Thai Name entity and Noun phrases. After Morphological level segmentation is achieved, EDU segmentation is then to be dealt with. According to Charoensuk et al. (2005), the principle of EDU segmentation is a clause or a simple sentence. EDUs must have unequivocally the nucleus or satellite of a rhetorical relation that holds between two EDUs or adjacent text spans and non-overlapping spans (Mann and Thompson, 1988). In Thai language, EDU is classified into two types: Basic EDU and Embedded EDU. Basic EDU is an EDU that has clause structure or simple sentence. Embedded EDU also has clause or phrase structure in the middle of the basic EDU. The examples of each EDU type are shown in below.

Basic EDU
1) [กะหล่ําปลีมีสีเขียว] /The cabbage has green color
2) โรคระบาดพบในภาคกลาง […] /Epidemics was found in the middle region [such as Pathumtani]

Embedded EDU
1) [กะหล่ําปลีมีสีเหลือง] จะมีสีเหลือง […] /The cabbage that was destroyed will have yellow color.
2) เกษตรกรควรใส้ปุยยูเรียลงในแปลงด้วย […] /Agriculturist should put fertilizer [such as Urea] into plot.

2.2 Annotation schema Design

In this section, the annotation type used in corpus is collaborative designed by language engineers and knowledge engineers. They classified discourse annotation into two sets; Cause-effect tag for causality extraction and Coordinate & Subordinate Tag for coherent tree construction.

2.2.1 Causality Tags

Causality extraction is based on multiple EDUs. Most of causality expressions are realized in two main forms: an inter-causal EDU and an intra-causal EDU. The inter-causal EDU is defined as a causality expression of more than one simple EDU and intra-causal EDU as a causality expression occurring within one EDU. Moreover, there are 20% of inter-causal EDU and 7% of intra-causal EDU from the annotated-causality corpora are found. Then, causality extraction focuses only on the inter-causal EDU. To extract a causative unit and its effective information unit in the form of the inter-causal EDU, these tags are defined for verb-pair rules learning for extracting the inter-causal EDU. Table 1 shows causality tag schema.

| Tag type          | Description                                                                 |
|-------------------|-----------------------------------------------------------------------------|
| Cause-Boundary tag| -To annotate the starting point of the cause unit by “<C id=num type=cause/noncause> EDU1 EDU2…….EDUn</C>” tag (where the attribute id is a cause number and the attribute type is type of EDUs in this boundary) |
| Effect-Boundary tag| -To annotate the starting point of the effect unit by “<R id=num type=effect/noneffect> EDU1 EDU2…….EDUn</R>” tag (where the attribute id is an effect number corresponding to the cause id and the attribute type is type of EDUs in this effect unit) |
| Embedded cause tag| -To annotate the embedded cause by “<EmC id=num type=cause/noncause>…”</EmC>” tag (where the attribute id is a cause number and the attribute type is type of the EDU that is embedded into the result EDU) |
| Effect-Boundary tag of the effect EDU containing an embedded-cause EDU| -To annotate the starting point of the effect unit by “<EmR id=num type=effect/noneffect>…”</EmR>” tag (where the attribute id is an effect number corresponding to the cause id and the attribute type is type of EDUs in this effect unit) |

Table 1: Annotation schema for causality extraction
Figure 1: Example of annotated causality relation

From the figure 1, we manually annotated the features specifying 'causality/non causality', and also annotating only NP1, NP2, and NP3 with their semantic from Wordnet (Miller G., 1995) through mapping Thai to English terms tool. This allows us to solve the word-sense ambiguity and the variety of surface forms of a word with the same concept. If the NP has a modifier, only the head noun will be annotated with the concept provided in the Thai plant encyclopedia. For example, if the Head noun is a ‘ผิว ผุ/dry/mottled mark’, ‘จุดสีน้ำตาล/brown spot’, etc., we label it as a ‘symptom’. The step after annotation is to extract the verb features and their information (NPs) from the annotated corpora, which are the trained data, for the inter-causal EDU extraction (see Table 2). Each verb concept will be learned for learning the verb-pair rules by NB and SVM techniques.

2.2.2 Coordinate & Subordinate Tag
The Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) is often used to extract salience through the application of knowledge based approach at discourse level. This theory includes explanations on the occurrence of discourse relations and text generation by using tree structures. There are 78 relations such as Elaboration, Explanation, Cause-result, Conditional, Contrast, Sequence, Consequence, Joint, List, Background etc. that are complex and vague, especially in Thai text. 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, in our applications, we reduce the number of relations to only two, namely Coordinating and Subordinating relation. The example of COR&SUBR-tree are shown in Figure 2.

Figure 2 : Coordinating(COR) and Subordinating (SUBR) relations with nuclearity

From the figure 2, N is nucleus and S is satellite.

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. So, these tags are developed as shown in Table 3.
Tag Description
1) <focus type={0,1,2}>..</focus>
Type { 0=introduce
l=continuous 2=interrupt}

2) <coref ref=antecedent value dist= type={0,1} > ..</coref>
dist ( distance value of anaphor far from antecedent ) ;type
0=zero anaphora
1=explicit reference

3) <rel name={co ,sub} ;dm= (w);co-dm=; dm-pos =
{1,2,3} ;dm-type={s,w} ; kp=; rel=001,002; ns=n,s >....
name=  {co,sub}
co ;coordinate relation
sub;subordinate relation
dm= discourse marker
codm= correlative discourse marker
dm-pos( position of dm)
1; at the beginning of the first edu
2; at the beginning of the second edu
3; at the beginning of the first edu and at the begin of second edu
dm-type = {s=strong,
w=weak}
kp( key phrase value) ={..}
rel(relevant value)={id1, id2..}
ns(nucleus, satellite)= {n,s}

Table 3 : COR&SUBR Tag and Attribute coding

The segmented EDU will be annotated with COR&SUBR tag which could be represented as a coherent tree as shown in Figure 4.

Example-1 [Text-1]
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.

From the example of COR & SUB tagged text in figure 3 , we can span coherent tree as in Figure 4

Figure 3 : The example of COR & SUB Tag text

The tagged corpus will be the input for machine learning to classify the COR/SUBR relations and compute a coherent structure (Sukvaree, T., 2007).

Although the tagging task is completed, but errors and inconsistencies still exist in the corpus. Furthermore, it may become more apparent with time. How to reduce errors, inconsistencies and time is a challenging problem. So, guidelines are developed for solving these problems.

2.2.3 Guidelines
Guidelines are essential for consistent and reliable annotation of texts. They contain examples for aiding the annotator in this process, and include detailed information to ensure strict adherence to them. By developing guidelines to which annotators can adhere, we can reduce the amount of discrepancies between annotators. In an ideal situation, guidelines would be available before annotation begins. However, real data from a corpus are far more complicated and subtle than examples discussed in the linguistics literature and many problems do not appear until sufficient data have been annotated.

To create annotated corpus in discourse level, we need to prepare five sets of guidelines; word segmenting, POS...
tagging, Name entity tagging, causality relation and COR&SUB relations tagging.

- **Word segmentation guideline**
  Unlike Western writing systems, Thai writing does not have a natural delimiter between words, such as space, for word segmentation. Therefore, it is usually required to segment Thai texts prior to annotate. And the notion of word is very hard to define.

- **POS Tagging guideline**
  This guideline is to describe the grammatical tagger. This tagset contains 48 POS tags and other tags (for punctuation and currency symbols).

- **Name entity guideline**
  This guideline explains about the category of NE that will be annotated and the general rules that are commonly applied to the annotation process for all of the entity types.

- **Causality relation guideline**
  This guideline explains about meaning of causal tag and cue set that is used for boundary identification of causative/effective unit (see the detail in 2.1).

- **COR&SUB relations guideline**
  This guideline explains a construction of the coherent tree with co-ordinate/sub-ordinate relations and terms (see the detail in 2.2).

During the annotation process, annotators can account the errors and inter-annotator inconsistencies. Sometimes linguistic problems posed by corpora are much more diverse and complicated than those discussed in theoretical linguistics or grammar books, and new problems surface as we annotate more data. Hence, our guidelines have been revised, updated and enriched incrementally as the annotation process progressed. In cases of disagreement among several alternatives, the one most consistent with the overall guidelines was chosen.

3. Annotation Process

The annotation components can be decomposed into two components: metadata annotation and linguistic information annotation. The first one establishes the information of text for text retrieval. The second is linguistic information for studying and modeling the language phenomena.

3.1 Metadata annotation

The annotation metadata framework is used to describe additional information about resources or information not directly found in text. This information can include, but is not exclusive too, comments, ideas for use, contextual explanations and other summary information. The fifteen Dublin Metadata Core Element Set (or simply, Dublin Core) and AgMES, Agricultural Metadata Element Set by FAO (http://www.fao.org/aims/intro_meta.jsp) have been employed with corpus. Figure 5 shows the graphic user interface (GUI) of this component.

3.2 Linguistic information annotation

A semi-automatic tool is provided to annotate linguistic information. This tool allows users to browse information in three independent layers.
Figure 6 shows the annotation process and related tools. They are packaged tools which consist of word segmentation, EDU segmentation, POS Tagging and NE Recognition. The tool supports wide range of applications that require language behavior analysis; for example, machine translation, information retrieval, information extraction, summarization, etc.

**NE Recognition (NER)**

Named entity (NE) extraction is very important in many NLP tasks. We will focus on extracting person, location, and organization name. To extract Thai NE, we proposed the approach by applying Maximum Entropy model and incorporate knowledge, which are rules and dictionary to NE extraction system. Our NER system divide into 2 modules: training module by conditional random field model, and NE extraction module. Both module needs same preprocesses, word segmentation and feature extraction. Feature that we use include orthographic, lexicon, and dictionary. We also use gazetteer and some heuristic rules that produce expert who observe our corpus to help in NE extraction. Result F-measurement for each class is 90.10 for person, 74.68 for location, and 83.55 for organization. Based on current NER, we utilized it as a tool for annotating NE.

**EDU segmentation**

Elementary discourse unit (EDU) segmentation is an important process, since it separates full text into minimal discourse units that are used as an input of many applications such as text summarization and question-answering. This tool used a hybrid approach for Thai EDU segmentation by using the decision-tree learning and rules. In additional, the important problem of this process is EDUs boundary ambiguity because Thai does not have punctuation marks or special symbols to signal EDU boundary and Embedded EDU usually occurring in the middle of another EDU. The precision and recall of the system are 0.80 and 0.81. As NER, we utilized this research as a tool for corpus annotating.

After the annotating in each level, annotator will check errors and validation by manual. At this current state, annotators annotate text in discourse level by using a general-purpose text editor or word processor.

### 4. Quality Control

The quality of annotated corpus involved two tasks; checking the validation of segmented EDUs including words and POS tags and measuring inter-annotator consistency.

#### 4.1 EDUs segmentation Validation Procedures

After completing annotation process automatically by tools, annotators will review each EDU for the correctness of EDU segmentation, word boundary, and name entity by using guidelines. All EDUs were checked for the errors and validated by manually. In discourse level annotation, our process is as follows: first, annotated files are uploaded on Wiki-like-website. Next, annotators will recheck the annotated files on Wiki. Then, other annotators are randomly selected file for re-editing. By this way, we monitor our consistency for improving the guidelines.
4.2 Inter-annotator consistency

The annotators have to identify those EDUs containing discourse level information. They annotate 75% and 25% of corpus for learning and evaluation, respectively. In order to assess the quality of the annotation we computed the inter-annotator agreement using Cohen’s Kappa coefficient (Cohen, J., 1960), it measures pairwise agreement among annotator and takes into account the possibility of their agreement by chance. Kappa takes values between 0 and 1, and it is considered that a value over 0.81 indicates high agreement between annotators, whereas values between 0.41 and 0.6 indicate moderate agreement. Usually values below 0.41 are taken to indicate little or no agreement. In this paper, we considered the agreement for causality or non-causality and co-ordinate or sub-ordinate EDUs.

Table 5 shows average kappa coefficient reflecting the agreement of two annotators. The statistics measure annotation reliability at two levels; causality or non-causality and co-ordinate or sub-ordinate.

| Annotation List | Avg. 1st | Avg. 2nd |
|-----------------|----------|----------|
| causality or non-causality | 0.591    | 1        |
| co-ordinate or sub-ordinate    | 0.54     | 0.78     |

Table 5: The score of annotator consistency by Kappa

Table 5 shows two sets of score. The first column represent agreement on annotated text without using guideline and second column represents agreement on annotated corpus with guidelines. The first set of scores 0.591 and 0.54 indicates moderate agreement. These scores set reflect disagreement and represent significance for improving those guidelines. While the second set of score is 1 and 0.78, they indicate high quality and almost agreement.

5. Results

The results of discourse annotation in 8,000 EDU of plant disease (http://www.doa.go.th/), the native technological journal (mostly in agriculture), and the online Bird flu and health news both causality relation and COR&SUBR relation annotation show the language phenomena and the statistic of language behavior. The first part of Table 6 shows the causative-verb concepts that can be classified into 2 main groups, a regular causative verb group and a compound causative verb group, where the compound causative verb is a general verb as shown in Table 6 and the second part shows the V-effective set. In Table 6, we can use it for statistical analysis.

From the corpus observation, we have found 5 discourse relations frequently used which are Elaboration, condition, Cause-Result, Joint, and Consequence as shown in Table 7. Table 8 shows the statistical occurrence of discourse markers that appear in text.

| Features | Verb concepts |
|----------|---------------|
| V (causative verb set with concept) | Regular causative verb group |
| Surface form | Concept |
| ติ่ง/suck, งิ้ว/suck, ติ่ง/ea | consume |
| ติ่ง/destuct, งิ้ว/eliminate, ติ่ง/spread out, งิ้ว/diffuse | destroy, spread |
| ติ่ง/ | force |
| ติ่ง/ | consume |

| Compound causative verb group |
| Surface form | Concept |
| ติ่ง+ฝูน/ | be+ get disease |
| ติ่ง+ฝูน/ | get pathogen |
| ติ่ง+สูบน/ | infect |
| ติ่ง+ | force |
| ติ่ง+บาย+ | food consume |

| V (effective verb set with concept) | Regular effective verb group |
| Surface form | Concept |
| ติ่ง+ข่ม | shrink, | bend, | change shape |
| ติ่ง+ดี | dry, | ไหม้ | blast, | be symptom |
| ติ่ง+รู | stunt, | be symptom |
| ติ่ง+ไส | rot, | ไส้ | spoil |
| ติ่ง | die |

| Compound effective verb group |
| Surface form | Concept |
| ติ่ง+จุด | be+ spot |
| ติ่ง+แผล | be+ scar |
| ติ่ง+จุด | have+ spot |
| ติ่ง+แผล | have  scar |

Table 6: The verb features and their concepts from the cause-effect annotation corpus

Table 7: Types of discourse relation with frequency of occurrences
Table 8: The phenomena of using Discourse Markers

6. Conclusion & Future work
In this paper, we build an annotated corpus that accounts for extracting coordinate and subordinate relations and construct coherence tree for text summarization and extracting Verb-pairs rules for mining causality to provide answer ‘why’ question. From the lesson learned in annotating, annotation guidelines become necessary and need continuous updating. The annotated corpus can be used to account for linguistic phenomena such as verb pairs in causal/non causal EDU, Discourse phenomena and types of discourse relations. Furthermore, the guidelines provided for annotators also have an influence on the inter-annotator agreement.

In future work, this annotation will be extended in QA system for tagging answers related to Know-How and Know-what. The lesson learned in annotating will be used for incrementally improving the guidelines. To cope with components (e.g. EDU segmentation, syntactic parser etc.) integration problem, common data format based on S-SSTC as being used in our tree editor, are going to be applied.

7. Acknowledgements
The work describes in this paper has been supported by the grant of NECTEC.

8. Reference
Carlson L., Marcu, D., Okurowski, M. E., “Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory”, In Current Directions in Discourse and Dialogue, pp.85-112, 2003.

Chang D.S, Choi, K.S., “Causal Relation Extraction Using Cue Phrase and Lexical Pair Probabilities”, LCNLP, pp. 61 - 70, 2004.

Chanlekh, H. Kawtrakul A., “Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information”, LCNLP, 2004.

Chareonsuk J., Sukvakree, T., Kawtrakul, A., “Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information”, NCSEC 2005.

Cohen, J., “A coefficient of agreement for nominal scales”, Educational and Psychological Measurement 20: 37–46, 1960.

Cristea, D, Postolache, O, Pistol, L., Summarisation through discourse structure. Computational Linguistics and Intelligent Text processing. Lecture Notes in Computer Science, 3406, pp. 632-644, 2005.

Edmundson, H.P. New methods in automatic abstracting, Journal of the Association for Computing Machinery, 16 (2): 264-285, 1969.

Girju R, Moldovan D., “Mining answers for causation questions”, In AAAI symposium on mining answers from texts and knowledge bases, 2002.

Imsonbuth, A., Kawtrakul, A., “Automatic building of an ontology on the basis of text corpora in Thai”, Language Resources and Evaluation Journal special issue on Asian Language technology, December, 2007.

Inui, T., Inui, K, Matsumoto, Y., “Acquiring causal knowledge from text using the connective markers”, Journal of the information processing society of Japan 45(3), pp. 919-993, 2004.

Jing, H., McKeown, K. The Decomposition of Human-Written Summary Sentences. In Proc SIGIR’99, Berkeley, pp129-136, 1999.

Kupiec, J., Pederson, J. and Chen, F. 1995 A trainable document summarizer. In Proc. ACM/SIGIR Annual Conference on Research and Development in Information Retrieval, Seattle, pp 68-73.

Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Toward a functional theory of text organization. Text 8(3), 243–281 (1998)

Marcu, D., Echihabi A. “An Unsupervised Approach to Recognizing Discourse Relations”, in Proc. ACL Conference, pp. 368 – 375, 2002

Marcu, D.: The rhetorical parsing of natural language texts. In: Meeting of the Association for Computational Linguistics, pp. 96–103,1997.

Miller, G. “WordNet: A lexical database”. Communications of the ACM, 38(11), pp. 39 – 41, 1995

Mosleh, H. Al-Adhaileh, Tang, E. K. and Yusoff, Z. “A synchronization structure of SSTC and its application in machine translation”. COLING 2002 Work-shop on Machine Translation in Asia, Taipei, Taiwan. 2002.

Orasan, C. Building annotated resources for automatic text summarisation. In Proc LREC-2002, Las Palmas de Gran Canaria, pp 1780-1786, 2002.

Pechsiri, C., Kawtrakul, A., Mining Causality from Texts for Question Answering System, The Institute of Electronics, Information and Communication Engineers(IEICE), IEICE Transactions .Oxford University Press, Volume 90-D. pp.1523-1533 October, 2007

Pengphnon, N., Kawtrakul A., Suktarachan M, “Word Formation Approach to Noun Phrase Analysis for Thai”, SNLP, 2002.

 Sudprasert S., Kawtrakul A., “Thai Word Segmentation based on Global and Local Unsupervised Learning”, NCSEC’2003.

Sukvaree, T, Kawtrakul, A., Caelen, J., Thai Text Coherence Structuring with Coordinating and Subordinating Relations for Text Summarization/CONTEXT 2007, Lecture Notes in Artificial Intelligence(LNAI), August, 2007.

Torisawa, K., “Automatic Extraction of Commonsense Inference Rules from Corpora”, In Proc. Of The Association for Natural Language Processing, pp. 318-321,2003.