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 (mostly in agriculture), and the online Bird flu and health news, representing 8,000 EDUs. Each document ranges in size from 13 to 24 EDUs, with an average of 120 words per document.

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 (Pengphon, 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 [เช่น ปทุมธานี]
Epidemics was found in the middle region [such as Pathumtani]

Embedded EDU 1) กะหล่ำปลี [ที่ถูกทำลาย] จะมีสีเหลือง
The cabbage [that was destroyed] will have yellow color.
2) เกษตรกรควรใส่ปุ๋ย[เช่น ปุ๋ยยูเรีย] ลงในแปลงด้วย
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 <C id=*num* type=*cause/noncause*> EDU1 EDU2…..EDUn </C>	-To annotate the starting point of the **cause** unit by "<C id=*num* type=*cause/noncause*>" tag (where the attribute **id** is a cause number and the attribute **type** is type of EDUs in this boundary) -ending boundary with" </C>" tag
Effect-Boundary tag <R id=*num* type=*effect/noneffect*> EDU1 EDU2…..EDUn </R>	-To annotate the starting point of the **effect** unit by "<R id=*num* type=*effect/noneffect*>" tag (where the attribute **id** is a effect number corresponding to the cause **id** and the attribute **type** is type of EDUs in this **effect** unit) -ending boundary with" </R>" tag
Embedded cause tag inside the result EDU <EmC id=*num* type=*cause/noncause*> …</EmC>	-To annotate the embedded cause by "<EmC id=*num* type= *cause/noncause*>" 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 <EmR id=*num* type=*effect/noneffect*> EDU1 EDU2…..EDUn </EmR>	-To annotate the starting point of the **effect** unit by "<EmR id=*num* type=*effect/noneffect*>" 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) -ending boundary with "</EmR>" tag
<np1 concept= '…..'>noun-phrase</np1>	To annotate the concept of an agent. This concept will be useful for selecting word sense of verb to be causative.
<VC concept= '….'>verb</VC>	To annotate the concept of verb for solving the word form variety
<np2 concept= '…..'>noun-phrase</np2>	To annotate the concept of a patient. This concept will be useful for selecting word sense of verb to be causative.
<VE concept= '….'>verb</VE>	To annotate the concept of verb for solving the word form variety

Table 1: Annotation schema for causality extraction

The cause-effect schema is used for annotating in an EDU corpus for machine learning in cause-effect relation between causative events and effective events. These relation can be expressed by a combination of the causative verb (vc) and the effective verb or result verb (ve) in the verb pairs from different EDUs or by a lexical pair from a lexico syntactic pattern within one EDU (in case of intra-causality). The example of annotated sentence is shown in Figure 1.

“เพลี้ย<u>ดูดกิน</u>น้ำเลี้ยงจากช่อดอกทำให้ดอก<u>แห้ง</u> <u>ร่วง</u>และ<u>ติดผล</u>น้อย”
(“‘Aphids <u>suck</u> sap from flower. [It] makes flower <u>dry</u> , <u>come off</u> and <u>yield</u> less.”)
<C id=1 type=cause>
<EDU><np1 concept='plant louse#1'> *เพลี้ย/aphids*
</np1>
<VC concept='consume#2'>*ดูดกิน/suck* </VC>
<np2 concept='sap/solution#1'>*น้ำเลี้ยง/sap*
</np2> <Preposition=’from’>จาก/from
</Preposition>
<np3 concept='plant organ/solution#1'> ช่อดอก/flower </np3>
</EDU>
</C>
<R id=1 type=effect>
<EDU> ทำให้/make ช่อดอก/flowers
**<VE concept='dry/be symptom'>*แห้ง/dry*
</VE>**
</EDU>
<EDU>
**<VE concept=‘come off/be symptom'>
ร่วง/come off</VE>**
</EDU>
<EDU>และ/and
<VE concept='yield#3'>*ติด ผลyield*</VE>
น้อย/less
</EDU>
</R>

Figure 1: Example of annotated causality relation

Ex.#	NP1/ concept	Vc/ *concept*	NP2/ concept	Ve/ concept	class
1	เด็ก/ person#2	เป็นโรค/ get disease.	-	อาเจียน/ be symptom	yes
2	อากาศ weather#1	เปลี่ยน แปลง Change#1	-	ไม่สบาย /be sick#1	yes
3	เด็ก person#1	ไอ symptom	-	คิดถึง/ think of#2	no
4	ลูก person#1	ได้รับวัคซีน get vaccine	-	มีไข้/have symptom	yes
5	ดอก/ plant organ #1	บาน bloom#1	-	ลดลง/ decrease#1	no
6	มะม่วง Plant organ#1	แทงช่อดอก Sprout#1	-	เพิ่มขึ้น /increase#1	no
7	เพลี้ยจักจั่น Plant louse#1	ดูด Consume #2	น้ำเลี้ยง	แห้ง dry/be symptom	yes
				ร่วง fall off/be symptom	yes
8	เพลี้ย	ระบาด	-	ไม่สมบูรณ์ /incomplete /symptom	yes
				ติดผล yield#1	yes
….	…..	….	…	…	…

Table 2 : Extracted Features for the inter-causal EDU

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 ‘*รอยด่าง/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.

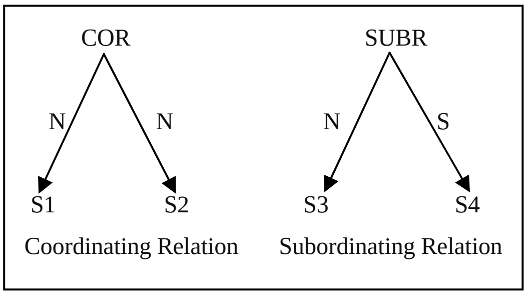


Figure 2 : Coordinating(COR) and Subordinating (SUBR) relations with nuclearity

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 as in Table 3

Tag	Description
1)< focus type={0,1,2} >..< /focus >	Type{ 0=introduce 1=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=(w) ;dm-pos = {1,2 ,3} ;dm-type={s,w} ; kp=(w) ;rel=(w); ns={n,s} >....	name= {co,sub}
	co ;coordinate relation
	sub;subordinate relation
	dm= discourse marker
	co-dm= correlative discourse marker
	dm-pos(position of dm){
	1; at the beginning of the first edu
	2; at the begin of the second edu
	3; at the begin of the first edu and 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.

```
<para id=001>
<topic id=001>ผักกาดหอม(letture)<\topic >
<topic id=002 name=โรคเน่าเละ parent=001>โรคเน่าเละ(Soft rot disease)" <\topic>
<text>
<edu id=001><focus>โรคนี้<coref ref= โรคเน่าเละ dist=1 type=1></coref></focus>พบมากเกือบทุกระยะการเจริญเติบโต</edu>
<edu id=002>โดยเฉพาะอย่างยิ่ง<focus type=0> <coref ref=โรคเน่าเละ dist=-2 type=0></coref></focus>พบมากในระยะห่อหัว</edu>
<rel name=sub; dm=; co-dm=; dm-pos=; dm-type=; kp=; relev=001,002; ns=n,s>
```

```
<edu id=003>ในระยะแรก<focus><coref ref=ผักกาดหอม dist=-1 type=0></coref></focus> พบเป็นจุดมีลักษณะฉ่ำน้ำ และรอยช้ำ</edu>
<edu id=004>หลังจากนั้น<focus>แผล</focus>จะขยายลุกลามออกไป</edu>
<edu id=005>และ<focus><coref ref=ผักกาดหอม dist=-1 type=0></coref></focus> เป็นเมือกเยิ้มมีกลิ่นเหม็นจัด</edu>
<rel name=sub; dm=; co-dm=; dm-pos=; dm-type=s; kp=; relev=003,[004,005]; ns=n,s>
<rel name=co; dm=; co-dm=; dm-pos=; dm-type=; kp=; relev=002,003; ns=n,n>
```

Figure 3 : The example of COR & SUB Tag text

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

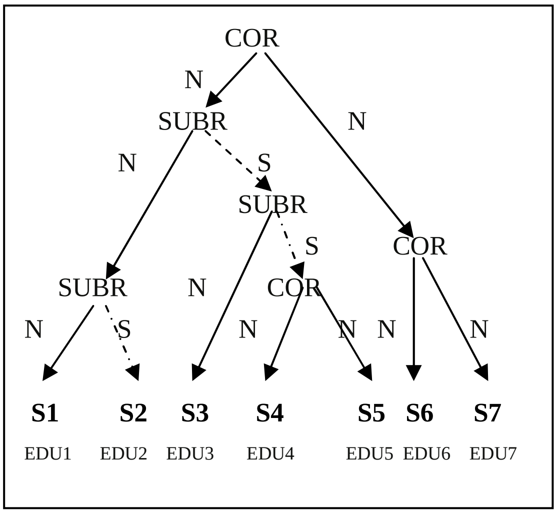


Figure 4 : Salience extraction on Coherent Tree

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

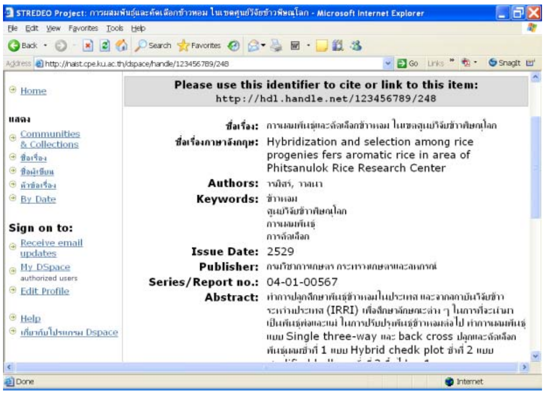
The annotation components can be decomposed into two components: metadata annotation and linguistic information annotation. The first one establishes the information of text for text retrieval. The second is linguistic information for studying and modeling the language phenomena.

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.


STREDEO Project: การผสมพันธุ์และคัดเลือกข้าวหอม ในเขตศูนย์วิจัยข้าวพิษณุโลก - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://naist.cpe.ku.ac.th/dspace/handle/123456789/248
Go Links SnagIt
Home
แสดง
Communities & Collections
ชื่อเรื่อง
ชื่อผู้เขียน
หัวข้อเรื่อง
By Date
Sign on to:
Receive email updates
My DSpace
authorized users
Edit Profile
Help
เกี่ยวกับโปรแกรม Dspace
Done
Internet

Please use this identifier to cite or link to this item:
http://hdl.handle.net/123456789/248

ชื่อเรื่อง: การผสมพันธุ์และคัดเลือกข้าวหอม ในเขตศูนย์วิจัยข้าวพิษณุโลก
ชื่อเรื่องภาษาอังกฤษ: Hybridization and selection among rice progenies fers aromatic rice in area of Phitsanulok Rice Research Center
Authors: วาณิสร์, วาลนา
Keywords: ข้าวหอม
ศูนย์วิจัยข้าวพิษณุโลก
การผสมพันธุ์
การคัดเลือก
Issue Date: 2529
Publisher: กรมวิชาการเกษตร กระทรวงเกษตรและสหกรณ์
Series/Report no.: 04-01-00567
Abstract: ทำการปลูกศึกษาพันธุ์ข้าวหอมในประเทศ และจากสถาบันวิจัยข้าวระหว่างประเทศ (IRRI) เพื่อศึกษาลักษณะต่าง ๆ ในการที่จะนำมาเป็นพันธุ์พ่อและแม่ ในการปรับปรุงพันธุ์ข้าวหอมต่อไป ทำการผสมพันธุ์แบบ Single three-way และ back cross ปลูกและคัดเลือกพันธุ์ผสมชั่วที่ 1 แบบ Hybrid chedk plot ชั่วที่ 2 แบบ

Figure 5 : Metadata annotation tool

Each layer has specific purpose of observation as shown in Table 4.

Layer	Tag/ mark up	Observation purpose
Morphological layer	POS, semantic, NE (name entity)	1. Grammar rules for parser, Machine Translation (MT) 2. Heuristic rules for word cut, NE Recognition 3. word collocation
Syntactic layer	Sentence, phrase, EDU	Sentences and phrase pattern for parser, MT
Discourse layer	Anaphora, co-referential , co-occurrence and discourse -marker	1. anaphora resolution for text summarization, MT, know-How, Know-why 3. Discourse Relation Recognition

Table 4 : Observations for language modeling

In each layer, the semi-automatic tool is provided to annotate linguistic information. This tool allows users to handle information in three independent layers; morphological, syntactic and discourse layer.

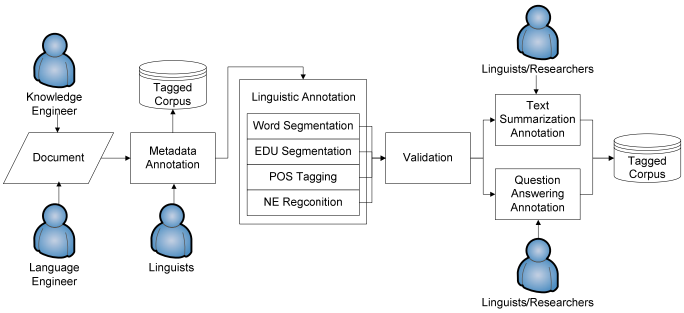


Figure 6 : Annotation Process

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.

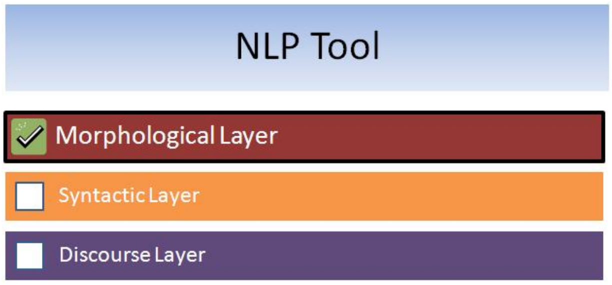
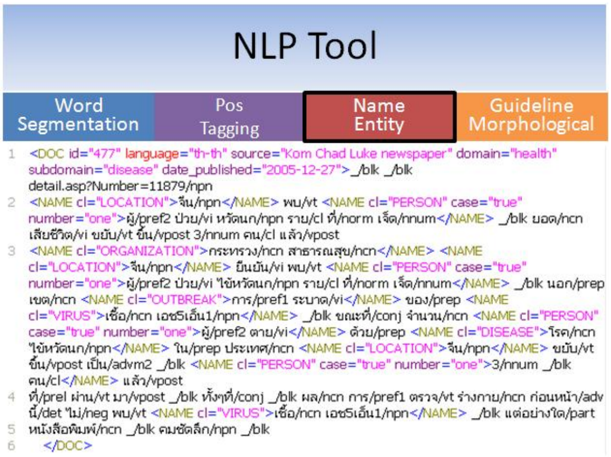


Figure 7 : Three layers of NLP Tool

Figure 7 shows three layers of NLP tools. Each tool will provide a set of features or tags for the user to select. And Figure 8 shows the process of tool in morphological layer.



```
<DOC id="477" language="th-th" source="Kom Chad Luke newspaper" domain="health"
subdomain="disease" date_published="2005-12-27">_/blk _/blk
detail.asp?Number=11879/npn
<NAME cl="LOCATION">จีน/npn</NAME> พบ/vt <NAME cl="PERSON" case="true"
number="one">ผู้/pref2 ป่วย/vi หวัดนก/npn ราย/cl ที่/norm เจ็ด/nnum</NAME> _/blk ยอด/ncn
เสียชีวิต/vi ขยับ/vt ขึ้น/vpost 3/nnum คน/cl แล้ว/vpost
<NAME cl="ORGANIZATION">กระทรวง/ncn สาธารณสุข/ncn</NAME> <NAME
cl="LOCATION">จีน/npn</NAME> ยืนยัน/vi พบ/vt <NAME cl="PERSON" case="true"
number="one">ผู้/pref2 ป่วย/vi ไข้หวัดนก/npn ราย/cl ที่/norm เจ็ด/nnum</NAME> _/blk นอก/prep
เขต/ncn <NAME cl="OUTBREAK">การ/pref1 ระบาด/vi</NAME> ของ/prep <NAME
cl="VIRUS">เชื้อ/ncn เอช5เอ็น1/npn</NAME> _/blk ขณะที่/conj จำนวน/ncn <NAME cl="PERSON"
case="true" number="one">ผู้/pref2 ตาย/vi</NAME> ด้วย/prep <NAME cl="DISEASE">โรค/ncn
ไข้หวัดนก/npn</NAME> ใน/prep ประเทศ/ncn <NAME cl="LOCATION">จีน/npn</NAME> ขยับ/vt
ขึ้น/vpost เป็น/advm2 _/blk <NAME cl="PERSON" case="true" number="one">3/nnum _/blk
คน/cl</NAME> แล้ว/vpost
ที่/prel ผ่าน/vt มา/vpost _/blk ทั้งๆที่/conj _/blk ผล/ncn การ/pref1 ตรวจ/vt ร่างกาย/ncn ก่อนหน้า/adv
นี้/det ไม่/neg พบ/vt <NAME cl="VIRUS">เชื้อ/ncn เอช5เอ็น1/npn</NAME> _/blk แต่อย่างใด/part
หนังสือพิมพ์/ncn _/blk คมชัดลึก/npn _/blk
   </DOC>
```

Figure 8 : The process for NE tagging

Word segmentation

Thai sentences are similar to the Japanese's and Chinese's in term of having no blank space to mark each word within the same sentences, the difficulty of identifying unknown words then includes the clarification of ambiguous segmentations. Additionally, most of a Thai multisyllabic word contains more than one monosyllabic word as parts of its components. As a result, Thai unknown words are generally, classified into two groups: explicit unknown and hidden unknown words. Explicit unknown words are those words which are not available in the dictionary. Hidden unknown words are those words which are composed of one or more known words

POS tagging

The POS tagger, based on the trigram model, will distinguish reliable from suspected assignments. Suspected assignments are highlighted and prompted for user confirmation or correction.

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

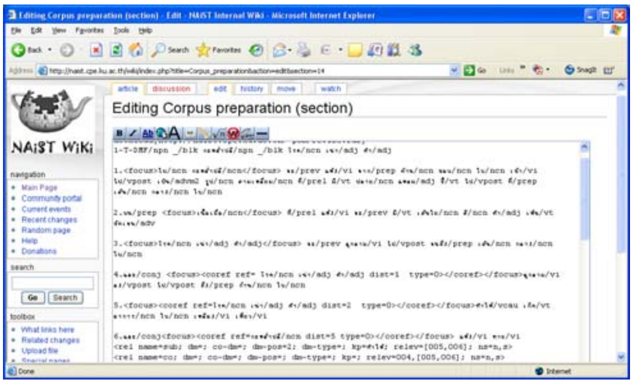
After the annotating in each level, annotator will check errors and validation by manual. At this current state, annotators annotate text in discourse level by using a general-purpose text editor or word processor.

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.



Editing Corpus preparation (section)

```
1-T-DBF/npn _/blk [illegible]/npn _/blk [illegible]/ncn [illegible]/adj [illegible]/adj

1.<focus>[illegible]/ncn [illegible]/ncn</focus> [illegible]/prev [illegible]/vi [illegible]/prep [illegible]/ncn [illegible]/ncn [illegible]/ncn [illegible]/vi
[illegible]/vpost [illegible]/advm2 [illegible]/ncn [illegible]/ncn [illegible]/prel [illegible]/vt [illegible]/ncn [illegible]/adj [illegible]/vt [illegible]/vpost [illegible]/prep
[illegible]/ncn [illegible]/ncn [illegible]/ncn

2.[illegible]/prep <focus>[illegible]/ncn</focus> [illegible]/prel [illegible]/vi [illegible]/prev [illegible]/vt [illegible]/ncn [illegible]/ncn [illegible]/adj [illegible]/vt
[illegible]/adv

3.<focus>[illegible]/ncn [illegible]/adj [illegible]/adj</focus> [illegible]/prev [illegible]/vi [illegible]/vpost [illegible]/prep [illegible]/ncn [illegible]/ncn
[illegible]/ncn

4.[illegible]/conj <focus><coref ref= [illegible]/ncn [illegible]/adj [illegible]/adj dist=1  type=0></coref></focus>[illegible]/vi
[illegible]/vpost [illegible]/vpost [illegible]/prep [illegible]/ncn [illegible]/ncn

5.<focus><coref ref=[illegible]/ncn [illegible]/adj [illegible]/adj dist=2  type=0></coref></focus>[illegible]/vcau [illegible]/vt
[illegible]/ncn [illegible]/ncn [illegible]/vi [illegible]/vi

6.[illegible]/conj<focus><coref ref=[illegible]/ncn dist=5 type=0></coref></focus> [illegible]/vi [illegible]/vi
<rel name=sub; dm=; co-dm=; dm-pos=2; dm-type=; kp=[illegible]; relev=[005,006]; ns=n,s>
<rel name=co; dm=; co-dm=; dm-pos=; dm-type=; kp=; relev=004,[005,006]; ns=n,s>
```

Figure 9 : Validating annotated corpus on Wiki

4.2 Inter-annotator consistency

The annotators have to identify those EDUs contained 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 as in Table 6 and the second part shows the V-effective set. In Table 6, we can use it for statistical analysis.

Features	Verb concepts	
V$_c$ (causative verb set with concept)	**Regular causative verb group**	
	Surface form	**Concept**
	ดูด/***suck,*** ดูดกิน/***suck.*** กิน/***ea,***	*consume*
	ทำลาย/***destruct,*** กำจัด/***eliminate,***	*destroy*
	ระบาด/*spread out,* แพร่กระจาย/*diffuse*	***spread***
	……………..	……………………
	Compound causative verb group	
	Surface form	**Concept**
	เป็น+โรค/ be+ disease,	***get disease***
	ได้รับ+เชื้อโรค/get+ pathogen,	***get pathogen***
	ติดเชื้อ/contract	***infect***
	เป็น+ แรงกดดัน/get pressure	***force***
	ได้รับ+อาหาร/get+food	***consume***
	……………..	……………………
V$_e$ (effective verb set with concept)	**Regular effective verb group**	
	Surface form	**Concept**
	หงิก /shrink, งอ/bend,	*change shape*
	แห้ง/dry, ไหม้/blast, แคระแกรน/stunt,	*be symptom*
	เน่าเละ/*rot,* เน่าบูด/*spoil*	*decay*
	ตาย/die	*die*
	……………..	……………………
	Compound effective verb group	
	Surface form	**Concept**
	เป็น+จุด /be+spot,, เป็น+แผล /be+ scar	*be symptom*
	มี+จุด /have+spot, มี+แผล /have scar	*have symptom*
	……………..	……………………

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

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.

Num	Discourse Marker				
8000	Elaborate	Condition	Cause-Result	Joint	Consequence
EDUs	27.77%	26.19%	19.84%	11.90%	8.73%

Table 7 : Types of discourse relation with frequency of occurrences

Type of relation	Discourse Marker				
Elaborate	โดย40.91%	แต่31.82%	ซึ่ง18.19%	ในกรณี 4.55%	นอกจากนี้ 4.55%
Condition	เมื่อ...จ.. 80%	ถ้า..จะ.. 12%	หากจะ..8%		
Cause-Result	ทำให้ 66.66%	เพราะว่า %13.33	เนื่องจาก %13.33	เป็นสาเหตุ 6.66%	
Joint	และ100%				
Consequence	ต่อมา %86.67	หลังจากนั้น 13.33%			

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", *IJCNLP*, pp. 61 - 70, 2004.

Chanlekha, H. Kawtrakul A., "Thai Named Entity Extraction by incorporating Maximum Entropy Modelwith Simple Heuristic Information", *IJCNLP'* ,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.

Imsombut, 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 Prof SIGIR'99*, Berkeley, pp129-136. 1999.

Kupiec, J., Pederson, J. and Chen, F. 1995 A trainable document summarizer. *In Prof. 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 Prof. 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 Prof 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

Pengphon, 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.