รายงานวิจัยฉบับสมบูรณ์

โครงการ การแปลภาษาด้วยเครื่องแบบสร้างและแก้ไข

โดย กัลยา นฤดมกุล

พฤศจิกายน 2547

สัญญาเลขที่ TRG4580108

รายงานวิจัยฉบับสมบูรณ์

โครงการ การแปลภาษาด้วยเครื่องแบบสร้างและแก้ไข

ผู้วิจัย

กัลยา นฤดมกุล
คณะวิทยาศาสตร์ มหาวิทยาลัยมหิดล

สนับสนุนโดยสำนักงานกองทุนสนับสนุนการวิจัย

(ความเห็นในรายงานนี้เป็นของผู้วิจัย สกว.ไม่จำเป็นต้องเห็นด้วยเสมอไป)

Acknowledgments

I would like to acknowledge The Thailand Research Fund (TRF) and Faculty of Science, Mahidol University for the financial support which makes this research possible.

My deepest gratitude and appreciation goes to my mentors: Prof. Nick Cercone and Assoc. Prof. Booncharoen Sirinoavakul for their invaluable advice and constructive criticism over the years.

I am thankful to the department of Mathematics, Faculty of Science, Mahidol University for computer resources, to Ms. Sompis Soodsakorn and Ms. Chaovanee Tamyong for all the support they have given me, to administrative and research staff in TRF especially Ms. Sangpetch Israpanitchakit for providing me with the information and answering all the questions I have with patience.

Lastly, I would like to express my deepest thanks and appreciation to Prof. Amaret Bhumiratana, Dean of Faculty of Science for encouraging and supporting me to continue research in machine translation.

บทคัดย่อ

รหัสโครงการ: TRG4580108

ชื่อโครงการ: การแปลภาษาด้วยเครื่องแบบสร้างและแก้ไข

ชื่อนักวิจัย: กัลยา นฤดมกุล คณะวิทยาศาสตร์ มหาวิทยาลัยมหิดล

E-mail Address: scknr@mahidol.ac.th

ระยะเวลาโครงการ: 1 ธันวาคม 2545 – 30 พฤศจิกายน 2547

โครงงานวิจัยนี้เสนอแนวทางปรับปรุงระบบการแปลภาษาด้วยเครื่องแบบสร้างและแก้ไข เพื่อใช้ในการพัฒนาระบบการแปลภาษาแบบสองทิศทางได้ดีและเหมาะสมยิ่งขึ้น (bi-directional translation system) หน่วยประมวลผลหลักทั้งสามของระบบซึ่งรวมถึง Analysis Lite Machine Translation (ALMT) Translation Candidate Evaluation (TCE) และ Repair and Iterate (RI) ได้ถูกทบทวนแก้ไขเพื่อให้มีประสิทธิภาพยิ่งขึ้น

โมดูล Word segmentation ได้ถูกเพิ่มเข้าใน ALMT เพื่อตัดคำในวลีหรือในประโยคอินพุท โมดูลนี้จะทำงานเมื่อประโยคอินพุทเป็นภาษาที่ไม่มีช่องว่างหรือเครื่องหมายใดๆระหว่างคำ เช่น ภาษาไทย จีน และ ญี่ปุ่น เซตของข้อบังคับ (constraints) สำหรับภาษาอังกฤษและภาษาไทยที่ใช้ใน ALMT ถูกเพิ่มเติมแก้ไขให้ใช้ได้ทั้งในกรณีที่เป็นภาษาต้นทางและภาษาปลายทาง ความสัมพันธ์ระหว่างคำลักษณะนามกับคำนามซึ่งแสดงด้วย word association number ถูกทบทวนแก้ไขเพื่อให้ระบบเลือกคำลักษณะนามได้ถูกต้องยิ่งขึ้น กฎ “not-head schema” ถูกเพิ่มเข้ามาใน TCE เพื่อใช้วิเคราะห์วลีหรือประโยคที่มีคำปฏิเสธอยู่ด้วย ระบบสามารถวิเคราะห์วลีหรือประโยคที่มีตัวเลขบ่งปริมาณและจำแนกค่าของตัวเลขได้โดยใช้ข้อมูลของตัวเลขที่ระบบสร้างขึ้นเองด้วยกฎ number lexical rule ทันทีที่มีการเรียกใช้ พจนานุกรมสองภาษา (ไทย-อังกฤษ) ที่ใช้ในขั้นตอนการแปลภาษาถูกสร้างขึ้นจากพจนานุกรมภาษาไทยและพจนานุกรมภาษาอังกฤษที่มีอยู่ในระบบ พจนานุกรมสองภาษานี้จะถูกสร้างขึ้นเมื่อระบบถูกเรียกใช้และสร้างเฉพาะข้อมูลของคำที่ปรากฏในวลีหรือประโยคอินพุทเท่านั้น

เพื่อแสดงประสิทธิภาพของการแปลภาษาด้วยเครื่องแบบสร้างและแก้ไข ระบบต้นแบบแปลภาษาอังกฤษ-ไทยได้ถูกพัฒนาขึ้นโดยใช้ SWI-Prolog 5.4 กฎไวยากรณ์ของคู่ภาษาได้ถูกพัฒนาตามรูปแบบของ Head-Driven Phrase Structure Grammar และ Attribute Logic Engine และส่วนการติดต่อกับผู้ใช้ได้ถูกพัฒนาโดยใช้ XPCE ผลจากการทดสอบพบว่าระบบสามารถสร้างผลการแปลที่สื่อความหมายของประโยคต้นทางได้ถูกต้องเป็นส่วนใหญ่โดยไม่ต้องมีการแก้ไข โครงสร้างของประโยคถูกต้องตามกฎไวยากรณ์ การเลือกใช้คำถูกต้อง ส่วนผลการแปลบางส่วนที่ต้องการการแก้ไขนั้น TCE และ RI ทำการแก้ไขโดยใช้กฎไวยากรณ์และพจนานุกรมที่พัฒนาขึ้นในโครงการวิจัยนี้

คำหลัก: การแปลภาษาด้วยเครื่อง การแปลภาษาด้วยเครื่องแบบสร้างและแก้ไข Head-Driven Phrase Structure Grammar

Abstract

Project Code: TRG4580108

Project Title: Generate and Repair Machine Translation

Investigator: Kanlaya Naruedomkul, Faculty of Science, Mahidol University

E-mail Address: scknr@mahidol.ac.th

Project: Period: 1 December 2002 – 30 November 2004

A new version of Generate and Repair Machine Translation (GRMT) is proposed to make it more suitable for developing bi-directional translation system. Each processing module in GRMT including Analysis Lite Machine Translation (ALMT), Translation Candidate Evaluation (TCE) and Repair and Iterate (RI) was revised.

The Word segmentation, which segments the input phrase/sentence into units that can be translated into other languages, is augmented to ALMT. The word segmentation module is activated only if the input string is in the language that has no explicit word boundary delimiters e.g., Thai, Chinese and Japanese. The sets of English and Thai constraints were revised so that the constraints can be applied via ALMT whether they are SL or TL. The classifier relation was re-designed for better selecting the appropriate classifier for each noun. A “not- head schema” is added to TCE to analyze a phrases/sentences containing negation. The system is able to analyze a string, with a quantity specified by a number, and to recognize its value by using information automatically generated by the “number lexical rule” when the system is activated. A bi-lingual dictionary is generated from the built-in SL and TL dictionaries once the translation is requested. Only the information of the words in the input string is generated for a better performance in a larger system.

The English-Thai MT prototype has been implemented to illustrate the performance of GRMT. The prototype has been developed and run under SWI-Prolog 5.4. The grammars were developed based on *Head-Driven Phrase Structure Grammar* and implemented on *Attribute Logic Engine*. The user interface was developed by using XPCE. This English-Thai MT system was evaluated and it performs in the way we intended. ALMT generated a large number of acceptable translations (grammatically correct, correct word usage and convey the original meaning) without repair. For some translations which require repairing, they are repaired by TCE and RI using the current HPSG based grammars and lexicons developed in this project.

Keywords: machine translation, Generate and Repair Machine Translation, Head-Driven Phrase Structure Grammar

Executive Summary

1. **Project Title** การแปลภาษาด้วยเครื่องแบบสร้างและแก้ไข
   Generate and Repair Machine Translation

2. **Project Leader**

Name	Asst. Prof. Kanlaya Naruedomkul
Education	Ph.D. (Computer Science)
Professional Address	Department of Mathematics Faculty of Science, Mahidol University Rama 6, Bangkok, Thailand 10400
Telephone Number	02-2015444
Fax Number	02-2015343
E-mail Address	scknr@mahidol.ac.th

3. **Mentors**

3.1

Name	Prof. Nick Cercone
Education	Ph.D. (Computer Science)
Professional Address	Faculty of Computer Science Dalhousie University 6050 University Avenue Halifax, Nova Scotia B3H 1W5 Canada
Telephone Number	902-494-2832
Fax Number	902-494-3962
E-mail Address	nick@cs.dal.ca

3.2

Name	Assoc. Prof. Booncharoen Sirinoavakul
Education	Ph.D. (Computer Engineering)
Professional Address	Computer Engineering Department King Mongkut's University of Technology Thonburi 91 Suksawasd 48, Bangkok 10140 Thailand
Telephone Number	02-4708002
Fax Number	02-8725050
E-mail Address	boon@kmutt.ac.th

4. **Field of Research** Natural Language Processing
5. **Project Grant** 480,000 Baht
6. **Project Duration** 2 years (1 December 2002 - 30 November 2004)

7. **Objectives**

Our primary goals are to continue the research on GRMT approach to improve its efficiency in order to develop bi-directional translation and to enhance multilingual translation capabilities. Our goals also are to encourage others to adopt our approach as a usable tool and to produce bilingual and multilingual MT system prototypes.

8. **Why Machine translation?**

Country to country exchanges in trade, technologies, politics, telecommunications, etc. and continues to grow rapidly. Language plays a significant role in the communication process between nations. In order to understand what is communicated, it is necessary to understand the language used in the communication process, its nuances and subtleties. Hence, machine translation (MT) assumes importance in order to facilitate this communication process from one language into another. When exchanges become more globalized, the requirement for machine translation (or machine aided translation) increases.

In multilingual countries the need for translation exists apart from increasingly globalized communications, e.g., in Canada (French-English), Switzerland (German, Italian, French, Swiss), India (32 official languages), etc. Translation enables people to express themselves in the way in which they wish to express themselves and to obtain the type of information they desire. Choice of language for communication and other aspects is important in order for people to retain and enhance distinctive cultures and distinctive ways of thinking. Language loss should matter to everyone.

The advance in computer technologies and the explosive growth of the World Wide Web have brought people closer together. Multlingual text has reached nearly to everyone with a computer. Information is readily accessible in growing numbers of languages. Machine translation has been integrated into man-machine communication systems which include electronic mail, information retrieval and the internet. The demand for machine translation systems has, therefore, undoubtedly increased.

For over a half century, MT research has drawn attention from people in different fields: cognitive psychology, linguistics, philosophy, cultural studies, computer science, computational linguistics. Several paradigms have developed including *Generate and Repair Machine Translation* (GRMT) [Naruedomkul and Cercone 2000]. GRMT is a constraint-based approach to MT that focuses on accurate translation output. GRMT is designed to be highly modular and extendible which enables it to have a great potential for a multilingual MT system. GRMT is applied in developing the English-Thai translation system. The developed English-Thai translation system generates acceptable translations (grammatically correct, correct word usage and convey the original meaning) for the sentences in the test corpus, some with repair and some without repair. However, a few translation sentences face with the problems of adding linking words and classifiers in Thai.

It is for these reasons that we decide to further study GRMT approach to improve the translation process, to strengthen its efficiency and make it attractive to a multilingual MT system.

Successful results from this research should prove beneficial to machine translation studies currently underway around the world. GRMT's application should minimize the language barrier in communication between anyone who speaks different languages.

9. Research Methodologies

To fulfill our objectives, the study is divided into 4 phases. However, sponsorship is initially requested to carry out the first and the second phases of the project.

Phase 1 -- Revise and improve the efficiency and performance of GRMT

9.1 Revise each processing module in GRMT including *Analysis Lite Machine Translation*, *Translation Candidate Evaluation* and *Repair and Iterate*.

9.2 Study on constraints used in *word treatment* and *word addition* modules for source and target languages respectively. For example, to find out when and which connecting word, plurality indicator or classifier (in Thai) is required, when and which preposition (in English) is required.

9.3 Explore semantic representation of some language structures, e.g., structure containing logical connection or negation.

9.4 Re-design the knowledge-bases e.g., dictionaries, grammars and lexicons to satisfy requirements for larger systems.

9.5 Evaluate GRMT by examining the developed English-Thai MT prototype. This prototype will be developed and run under SWI-Prolog. The grammars will be developed based on *Head-Driven Phrase Structure Grammar* [Pollard and Sag 1987; Pollard and Sag 1994] and implemented on *Attribute Logic Engine* [Carpenter and Penn 1999]. The user interface will be developed by using XPCE.

Timing

It is expected to take 12 months to complete this phase.

Phase 2 -- Develop a bi-directional translation system based on GRMT approach

9.6 Design and develop all necessary knowledge-bases required in a bi-directional translation system including:

- sets of constraints which are specific to resolve the syntactic differences between the source and the target languages.
- a SL dictionary to be used together with a set of constraints to refine the scope of the translation choices.
- a bilingual dictionary to be used in relating SL and TL.

9.7 Classify words into appropriate categories.

9.8 Develop a semantic relationship between words based on word classification to be used in selecting an appropriate translation for each input word.

9.9 Develop the ordering rules. These rules are based on the syntactic structure difference between the SL and TL.

9.10 Design and developing the parsers for SL and TL. The grammars used are developed based on Head-Driven Phrase Structure Grammar, and implemented on Attribute Logic Engine.

9.11 Develop an English ↔Thai MT prototype. This prototype will be developed and run under SWI-Prolog. The user interface will be developed by using XPCE.

9.12 Test and evaluate.

Timing

It is expected to take 12 months to complete this phase.

10. Plan of Activities and Expected Output

Month / Phase	Activity	1	2	3	4	5	6	7	8	9	10	11	12
Phase 1	9.1												
	9.2												
	9.3												
	9.4												
	9.5												
Month / Phase	Activity	13	14	15	16	17	18	19	20	21	22	23	24
Phase 2	9.6												
	9.7												
	9.8												
	9.9												
	9.10												
	9.11												
	9.12												

Phase 1 -- Revise and improve the efficiency and performance of GRMT

- Revise each processing module in GRMT including *Analysis Lite Machine Translation*, *Translation Candidate Evaluation* and *Repair and Iterate*.

1. GRMT: The 3-step translation process

GRMT comprises three modules as shown in Figure 1: Analysis Lite Machine Translation (ALMT), Translation Candidate Evaluation (TCE) and Repair and Iterate (RI). ALMT generates the translation which we call translation candidate (TC), next TCE verifies the TC to see whether it retains the meaning of the original sentence. If the TC does retain the original meaning, the TC is then redeemed as the translation. If the TC does not retain the original meaning, the TC will then be repaired by RI. Basically, the translation process generates the translation candidate and then repairs it when necessary.

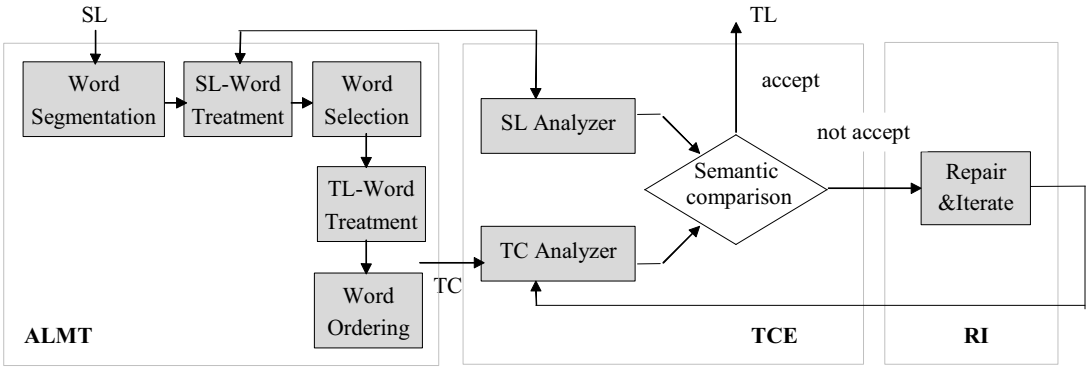


Figure 1 GRMT architecture

1.1. Analysis Lite Machine Translation

Analysis Lite Machine Translation (ALMT) was re-designed to generate an appropriate translation candidate from the source language. ALMT performs generation in five phases: word segmentation, SL_word treatment, word selection, TL_word treatment and word ordering.

Word segmentation segments the input (string) sentence into units that can be translated into other languages. The word segmentation module is activated only if the input string is in the language that has no explicit word boundary delimiters e.g., Thai, Chinese and Japanese. Inappropriate segmentations can lead to a wrong translation result. In this research, we have developed the Thai Word Segmentation algorithm which we call ***Thai Word Segmentation for Accurate Translation*** (TSAT). TSAT performs the segmentation in two phases: *Word boundary tagging* and *Word boundary selection*. Word boundary

tagging is to find all possible word boundaries while word boundary selection is to find the most appropriate word boundary.

SL_word Treatment is composed of two steps: source language constraints application and dictionary look-up. The SL constraints are applied to narrow the scope of possible TL words that correspond to each SL word. Dictionary look-up maps all corresponding words in the TL to each SL word.

SL constraints are characteristics of the SL which differ from those of the TL. Some constraints, e.g., plurality, continuous tense, passive voice, adjective and negation are considered in the English ↔Thai translation system.

Word selection selects the most appropriate word if there is more than one possible meaning. The selection process is performed by considering the semantic relationship between words. This semantic relationship is based on the Word Association number (WordAsso). WordAsso number is assigned to a word class. We classify words according to their meaning and usage. The first meaning appearing on the list of meanings of each word is selected in the case where the semantic relationship fails.

TL_word treatment adds (back) any syntactic information which was removed previously in the *SL_word treatment module* into the string in the form of the TL to maintain the meaning of the original sentence. This process is performed by applying the TL constraints. Syntactic differences between the TL and correct SL constructs are guided by application of TL constraints.

Word ordering rearranges all selected words in the TL grammatical order to complete the sentence in the TL by consulting the ordering rules. These ordering rules are generated from the syntactic level differences between languages. The structure of the SL which is similar to that of the TL remains the same, only the sentence fragments that are different will be rearranged into the grammatical order of the TL.

1.2. Translation Candidate Evaluation

Translation Candidate Evaluation (TCE) verifies the accuracy and correctness of the TC in terms of both syntax and semantics. TCE performs the verification in two phases: SL-TL Analyzer and semantic comparison.

SL-TL Analyzer performs the evaluation in two steps: Parsing and Semantic Extraction. Parsing parses both the TC and the SL in parallel to examine their syntax and semantics. Semantic Extraction extracts semantic information of each parse. Only their semantic results are considered since there are syntactic level differences between languages.

In our English ↔Thai translation system, we have developed grammars for English and Thai based on Head-Driven Phrase Structure Grammar [Pollard and Sag 1994]. The grammars we developed have been implemented using the Attribute Logic Engine (ALE) version 3.2 Beta. ALE is an integrated phrase structure parsing and definite clause logic programming system in which the terms are typed feature structures [Carpenter and Penn 1999].

Semantic Comparison cross-examines the meaning of the TC with that of the SL once semantic information of the SL and the TC are extracted successfully. If their semantic results are the same, that TC will be deemed an acceptable translation. Otherwise, TCE will report the differences to the next phase, RI, for further corrections.

1.3. Repair and Iterate

Repair and Iterate (RI) uses the information provided by TCE to repair the TC if the semantics of the TC is different from that of the SL. TCE identifies the different parts and RI determines whether that part should be removed and/or replaced. The differences might result from the word selection or word ordering process. Therefore, during the repair process, RI may request a re-segmentation, re-selection or re-ordering process if necessary.

In the case that the different part is removed and/or replaced, the transitional TC will be put through the word ordering module to have its syntax revised. Once the revision is completed, the repaired TC is returned to TCE for re-evaluation.

- Study on constraints used in *word treatment* and *word addition* modules for source and target languages respectively. For example, to find out when and which connecting word, plurality indicator or classifier (in Thai) is required, when and which preposition (in English) is required.

Constraint Application in English → Thai Translation

The set of constraints in Table 1 are applied via ALMT to simplify the structure of the English input sentence and to refine the scope of translation choices of each input (English) word. The second column shows the input/output of the constraint application process.

Table 1: Some SL (English) Constraints

SL-Constraints	Descriptions	Examples
Plurality	noun_(e)s → noun + plural	The books … →The book+ plural
Continuous tense	V to be + V-ing → ing + V	I am swimming → I + ing + swim
Passive voice	be + V3 → passive + V	He was arrested → He + passive + arrest
Adjective	V to be + adj → adj	I am glad → I + glad
Negative	V to do + not + V → not + V	He does not eat → He + not + eat

The set of constraints in Table 2 are applied via ALMT to complete the syntax of the translation language. Again, the second column shows the input/output of the constraint application process.

Table 2: Some TL (Thai) Constraints

SL-Constraints	Descriptions	Examples
Continuous tense	V + ing → กำลัง + V	ฉัน+ ing + ว่ายน้ำ → ฉัน + กำลัง + ว่ายน้ำ
Passive voice	passive + V → ถูก + V	เขา + passive + จับ → เขา + ถูก + จับ

When translating from English into Thai, GRMT begins the process by applying the SL (English) constraint. Some inflections which are removed from the input string, their corresponding features will be added to preserve their semantics. For example, once "-ing" is removed from the word "swimming", the feature "-ing" will be added to preserve their "continuous tense". The SL Constraint output is put through the *Word Selection Module* and all the selected words are forwarded to the *TL-word treatment module*. In the *TL-word treatment module*, the morphological and syntactic characteristics of the SL which are removed in the SL constraint application step will be replaced with the appropriate corresponding words by applying the TL (Thai) constraint application. For example, the features "-ing" is replaced by the word "กำลัง". The feature "กำลัง" indicates that the event is being carried on at the moment.

In the case that the SL (English) contains words expressing a quantity of countable nouns, e.g., “many”, “some” or numbers, *classifiers* are required in its translation. In Thai, a classifier indicates the unit of a noun, each countable noun relates to a specific classifier. Therefore, the classifier relation was designed in the form of WordAsso numbers to be used to select the appropriate classifier for each noun in *TL-word treatment module*.

A noun and its classifier were stipulated by the Thai Royal Institute. To date, there are approximately 3000 different classifiers [The Thai Royal Institute 1995]. From our studies, we classified the classifiers into classes as shown in Table 3. With this classification together with the specification of the Thai Royal Institute, we then developed the relation between nouns and their classifiers, some examples are illustrated in Figure 2. The noun with the WordAsso number in the first argument is compatible with a classifier with a WordAsso number shown in the second argument.

Table 3 A small fraction of classifier classes

Class	Sample Word
2-4-2 Classifier	
2-4-2-1 Classifier of object	
2-4-2-1-1 Classifier of living thing	
2-4-2-1-1-1 Classifier of human	คน
2-4-2-1-1-2 Classifier of animal	ตัว
2-4-2-1-1-3 Classifier of plant	ต้น
2-4-2-1-2 Classifier of non-living thing	
2-4-2-1-2-1 Classifier of tool	อัน
2-4-2-1-2-2 Classifier of room	ห้อง
2-4-2-1-2-3 Classifier of external body part	ขา หัว มือ
2-4-2-1-2-5 Classifier of root vegie	หัว
2-4-2-2 Classifier of collection	ชนิด ประเภท
2-4-2-2-1 Classifier of human	กลุ่ม
2-4-2-2-2 Classifier of animal	ฝูง
2-4-2-3 Classifier of time period	day วัน hour ชั่วโมง
2-4-2-5 Classifier of frequency	ครั้ง

Example 1 illustrates how to select the appropriate classifiers for the words *woman*, *cat* and *hen*. The selected words in Thai corresponding to the words *woman*, *cat* and *hen* in this example are shown in the second column of Table 4. The indefinite determiners *a* and *an* in this expression, corresponding to the word หนึ่ง in Thai, indicate the need for classifiers for the words ผู้หญิง, แมว and ไก่ respectively. The word ผู้หญิง belongs to the class Female (1-1-1-1-1-2), a subclass of Human (1-1-1-1). A noun which

belongs to the class 1-1-1-1 is compatible with a classifier with the WordAsso number 2-4-2-1-1-1 based on the classifier relation illustrated in Figure 2. Therefore, the classifier คน with 2-4-2-1-1-1 in Table 3 is selected for the word ผู้หญิง.

```
clf_rel('1-1-1-1',[[wasso('คน',['2-4-2-1-1-1'])]]).
clf_rel('1-1-1-2',[[wasso('ตัว',['2-4-2-1-1-2'])]]).
clf_rel('1-1-2-1-6-1',[[wasso('แห่ง',['2-4-2-1-2-2'])]]).
clf_rel('2-4-3-1-2',[[wasso('วัน',['2-4-2-3'])]]).
clf_rel('1-1-2-1-1-2-2',[[wasso('หลัง',['2-4-2-1-2-10'])]]).
```

Figure 2 Examples of classifier relations.

Table 4. The selected words in Thai for the words *woman*, *cat* and *hen*

Example 1: An old woman lived in the cottage, with a fat black cat and a plump brown hen.

English	Selected Word in Thai	WordAsso	Class	Classifier
woman	ผู้หญิง	1-1-1-1-1-2	Female	คน
cottage	กระท่อม	1-1-2-1-1-2-2	Housing	หลัง
cat	แมว	1-1-1-2-1-1	Mammal	ตัว
hen	ไก่	1-1-1-2-1-2-2.	Fowl	ตัว

The words แมว (cat) and ไก่ (hen) belong to the classes mammal (1-1-1-2-1-1) and fowl (1-1-1-2-1-2-2) respectively. Both classes are subclasses of animal (1-1-1-2). Since a noun with the WordAsso number 1-1-1-2 relates to a classifier with 2-4-2-1-1-2 according to the classifier relation shown in Figure 2, the classifier ตัว with 2-4-2-1-1-2 is selected for the words แมว (cat) and ไก่ (hen).

The definite determiner *the* corresponding to the word นั้น in Thai indicates the need for a classifier for the word กระท่อม (cottage). The word กระท่อม belongs to the class Housing (1-1-2-1-1-2-2) which is compatible with a classifier with the WordAsso number 2-4-2-1-2-10 based on the classifier relation illustrated in Figure 2. Therefore, the classifier หลัง with the WordAsso 2-4-2-1-2-10 is selected for the word กระท่อม.

Further researches on connecting words and prepositions are still required before we are able to make any conclusion.

- Explore semantic representation of some language structures, e.g., structure containing logical connection or negation.

To analyze the phrases/sentences containing negation, the “not- head schema” (Figure 3) is required. This schema combines negation “ไม่ (not)” and verb.

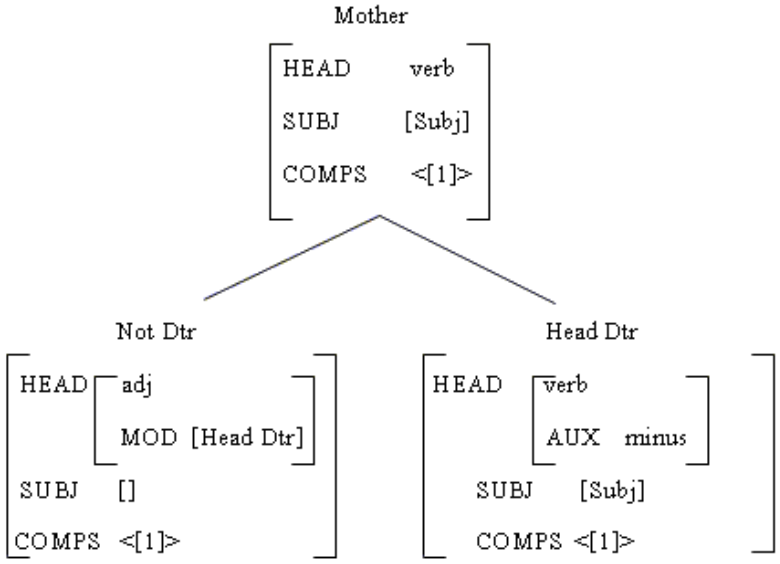


Figure 3 not- head schema

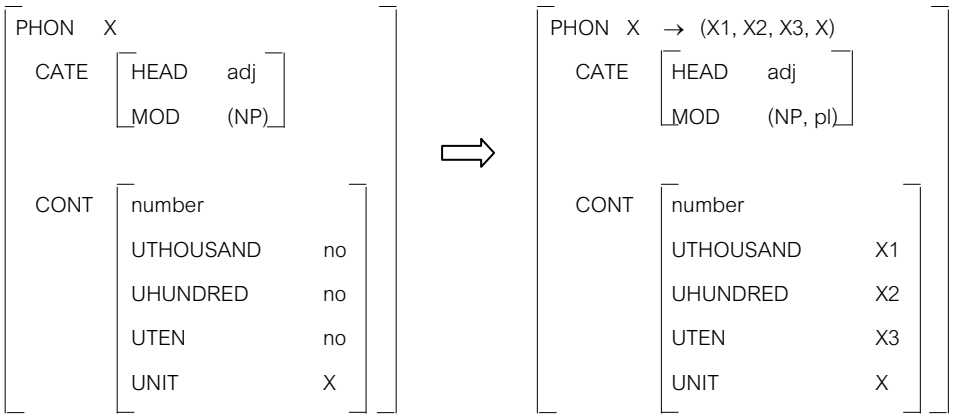


Figure 4 Number lexical rule (X, X1, X2, X3 ∈ {0, 1, 2, 3, 4, 5, 6, 7, 8, 9})

The lexicon used in analyzing English and Thai phrases/sentences in TCE module contains English and Thai words together with their syntax and semantic information. To make the system to be able to

parse a string with a quantity specified by a number (e.g., 5 ducks, 50 ducks) and to distinguish one number from others (e.g., distinguish 5 from 50), the syntax and semantic information of each number is required. However, storing all numbers and their information in TCE lexicon is space consuming. Therefore, only ten digit number 0, 1, … , 9 are included in the lexicon. The system will generate numbers from 10-9999 by the “number lexical rule” (Figure 4) when the system is activated.

The semantic representation of the structure containing logical connection “and” and “or” is under study still.

- Re-design the knowledge-bases e.g., dictionaries, grammars and lexicons to satisfy requirements for larger systems.

In the previous version of GRMT, four types of lexicons: SL dictionary, bi-lingual dictionary, TL dictionary and TCE lexicon were built for the system. The SL dictionary contains English word entries and their syntactic categories. Bilingual dictionary contains English word entries and all of their possible corresponding words in target language together with their word association number [Naruedomkul and Cercone, 1999]. Word association number is assigned to a word class. Words are classified according to their meaning and usage. TL dictionary contains Thai word entries and their syntactic categories. TCE lexicon contains both syntactic and semantic information of the word entries in terms of feature structure. The SL, TL and bilingual dictionaries are represented in Prolog. TCE lexicon is encoded using the Attribute Logic Engine (ALE) version 3.2 beta. ALE is an integrated phrase structure parsing and definite clause logic programming system in which the terms are typed feature structures [Carpenter and Penn 1999].

In a current version of GRMT, SL dictionary, TL dictionary and TCE lexicon were built for the system. A bi-lingual dictionary will be generated once the system is activated. For a better performance in a larger system, only the information of the words in the input string is generated. For example, to translate the input string in Example 1, a bi-lingual containing 16 words (an, old, woman, live, in the, cottage, with, a, fat, black, cat, and, plump, brown, hen) is generated to be used in the translation process.

- Evaluate GRMT by examining the developed English-Thai MT prototype. This prototype will be developed and run under SWI-Prolog. The grammars will be developed based on *Head-Driven Phrase Structure Grammar* [Pollard and Sag 1987], [Pollard and Sag 1994] and implemented on *Attribute Logic Engine* [Carpenter and Penn 1999]. The user interface will be developed by using XPCE.

The English-Thai MT prototype has been developed and run under SWI-Prolog. The grammars were developed based on *Head-Driven Phrase Structure Grammar* [Pollard and Sag 1987], [Pollard and Sag 1994] and implemented on *Attribute Logic Engine* [Carpenter and Penn 1999]. The user interface was developed by using XPCE.

This English-Thai MT system was evaluated and it performs in the way we intended. ALMT generated a number of acceptable translations (grammatically correct, correct word usage and convey the original meaning) without repair. TCE and RI improved a few sentences using the current our HPSG based grammars and lexicons.

Phase 2 -- Develop a bi-directional translation system based on GRMT approach

- Design and develop all necessary knowledge-bases required in a bi-directional English-Thai translation system including:
    1. Sets of constraints which are specific to resolve the syntactic differences between English and Thai.
    2. English and Thai dictionaries to be used together with a set of constraints to refine the scope of the translation choices and to complete the syntax of translation.
    3. TCE lexicon for the parser in the evaluation process.

- Classify words in the developed dictionaries into appropriate categories for being used in word selection and classifier selection processes.

- Develop a semantic relationship between words based on word classification to be used in selecting an appropriate translation for each input word.

- Develop the ordering rules. These rules are based on the syntactic structure difference between the English and Thai.

- Design and developing the parsers for SL and TL. The grammars used are developed based on Head-Driven Phrase Structure Grammar, and implemented on Attribute Logic Engine.

- Develop an English ⇔ Thai MT prototype. This prototype is developed and run under SWI-Prolog 5.4. The user interface is developed by using XPCE.

- Test and evaluate.

Discussion

The central theme of this research is to rework on GRMT approach to improve its efficiency in order to develop bi-directional translation. GRMT has endured remarkably well, including the translation process and translation's accuracy ensuring.

GRMT comprises three modules: Analysis Lite Machine Translation (ALMT), Translation Candidate Evaluation (TCE), and Repair and Iterate (RI). The first phase, ALMT, generates translation candidates for the source language without performing any sophisticated analysis. This process ensures that the translation candidate can be generated quickly and simply. Next, TCE, the second phase analyzes the generated TC to determine accuracy of the translation. Then, RI, the third phase repairs the generated TC until repair is no longer necessary. The TCE and RI stages ensure the accuracy of the translation result.

We revised some parts of GRMT to support the need of bi-directional translation. In the previous version, there are four phases in ALMT: SL_word treatment, word selection, TL_word treatment and word ordering. In this version, the Word segmentation, which segments the input phrase/sentence into units that can be translated into other languages, is augmented. The word segmentation module is activated only if the input string is in the language that has no explicit word boundary delimiters e.g., Thai, Chinese and Japanese.

The sets of English and Thai constraints were revised so that it can be applied via ALMT to simplify the structure of the input sentence and to refine the scope of translation choices of each input word when English is a source language. The same set of constraints can also be applied to complete the syntax of the translation language when English is a target language (TL). The set of Thai constraints was revised for the same purpose as well. Another significant part of completing the syntax of TL is to select the most appropriate classifier for each noun when TL is Thai and its SL contains words expressing a quantity of countable nouns. Therefore, the classifier relation was re-worked to provide a better selection.

Translation Candidate Evaluation (TCE) was revised to provide more coverage. A "not- head schema" is added to analyze the phrases/sentences containing negation. The syntax and semantic information of each number is required for parsing a string with a quantity specified by a number and to distinguish one number from others.

To reduce the storage space, the TCE lexicon includes only ten digit number 0, 1, … , 9. The numbers from 10- 9999 will be automatically generated by the “number lexical rule” when the system is activated. A bi-lingual dictionary will be generated from the built-in SL and TL dictionaries once the system is activated. Only the information of the words in the input string is generated, for a better performance in a larger system.

We have proved the idea of GRMT by constructing the English-Thai MT system. The English-Thai MT system translates isolated sentences (sentence by sentence). This English-Thai translation system has been developed and run under SWI-Prolog 5.4. The English and Thai grammars have been developed based on the Head-Driven Phrase Structure Grammar formalism [Pollard and Sag 1987; Pollard and Sag 1994] and implemented on the Attribute Logic Engine (ALE) [Carpenter and Penn 1999]. The user interface was developed by using XPCE.

This English-Thai MT system was evaluated and it performs in the way we intended. ALMT generated a number of acceptable translations (grammatically correct, correct word usage and convey the original meaning) without repair. TCE and RI improved a few sentences using the current our HPSG based grammars and lexicons.

I believe the results of this research will contribute to current attempts to develop accurate and reliable bi-directional machine translation systems and to produce quality translations from one language to another. This accurate and reliable translation methodology should enhance the effectiveness of communication among people. Nevertheless, I hope to have shown that pursuing further research in this direction is a worthwhile aim, and one likely to result in commercial machine translation systems in the not too distant future.

Related Activities

June 24, 2003 I was invited to give a presentation on "Machine Translation" for the Ph.D. students at the Institute for Innovation and Development of Learning Process., Mahidol University (see Appendix D).

December 18, 2003 Prof. Nick Cercone, Dr. Booncharoen Sirinaovakul and I set up the seminar on Machine Translation during 15-16 March 2004.

September 15, 2004 I was invited to be a local organizing committee of MT SUMMIT X which will be held at Phuket, Thailand in September 2005 (see Appendix E).

September 15, 2004 I was invited to be a conference committee of PACLING'05 which will be held in Tokyo, Japan in August 2005 (see Appendix F).

Research Achievements

1. An improved GRMT approach.
2. English-Thai MT system prototype (see Appendix A for User's guide).
3. Two research papers:
   3.1. Nuntadilok, J. Naruedomkul, K. and N. Cercone. (2003) Thai Word Segmentation For Accurate Translation *In* Proceedings of the Conference Pacific Association for Computational linguistics (PACLING'03), Halifax, Nova Scotia, Canada, p 97-107 (see appendix B).
   3.2. Pluempitiwiriyawej, C., Naruedomkul, K. And N. Cercone. (2003) Towards Mulex – A Multilingual Lexical Database System For Machine Translation *In* Proceedings Of The Conference Pacific Association For Computational Linguistics (Pacling'03), Halifax, Nova Scotia, Canada, p 181-189 (see appendix C).
4. One journal paper (submitted).

Expected Benefits

- The research result should prove beneficial to researches in MT.
- It could be used for Artificial Intelligence course.
- GRMT's applications should minimize the language barrier in communication between anyone who speaks different languages.
- With further study, it could lead to the machine translation system for a commercial use.

References

CARPENTER, B. and G. PENN. 1999 ALE: The Attribute Logic Engine User's Guide Version 3.2 Beta, available on line at "http://www.sfs.nphil.uni-tuebingen.de/~gpenn/ale.html#Obtain," May.

Naruedomkul, K. 2000. Machine Translation, PhD thesis, Department of Computer Science, University of Regina, Canada.

Naruedomkul, K. and N. Cercone. 1999 The Role for Word Association Numbers in Machine Translation *In* Proceedings of the Conference Pacific Association for Computational linguistics (PACLING'99), Waterloo, Ontario, Canada, p.379-392.

POLLARD, C. and I.A. SAG. 1987. Information-Based Syntax and semantics, Lecture notes No. 13, Stanford, California: CSLI Publication.

POLLARD, C. and I.A. SAG. 1994. Head-Driven Phrase Structure Grammar. Center for the Study of Language and Information, Stanford, The University of Chicago Press, Chicago & London.

Appendix A

GRMT User's Guide

1. How to run GRMT

- Double click on "GRMT.exe" icon 

to activate the translation program as illustrated in Figure A1.

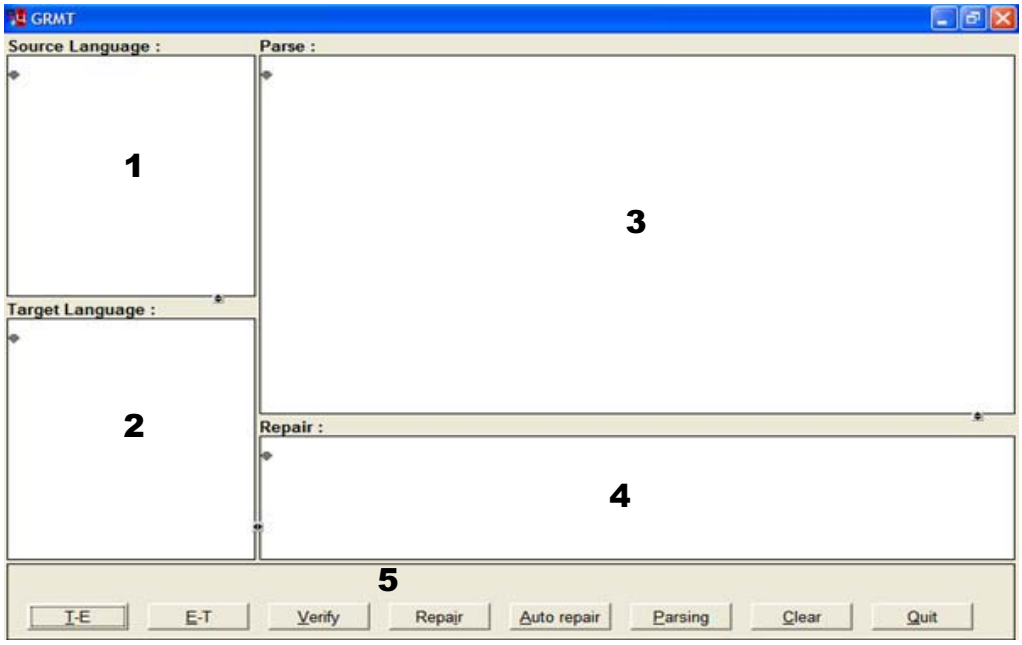


Figure A1 GRMT

#	Item	Description
1	Source Language panel	Panel for input phrase/sentence (Source Language: SL).
2	Target Language panel	Display a translation output (Target Language: TL).
3	Parse panel	Display both parses of input and output.
4	Repair panel	Display corrected TL (if any).
5	Activity Buttons	Button to control the activity.

2. How to translate?

- Type English/Thai input in 1 .

- Click E-T to translate from English into Thai. Click T-E to translate from Thai into English.
- The translation will be displayed in 2 .

3. How to verify?

- Once the translation output appears in 2 , click Verify to verify TL.
- The verification result will be displayed in 3 .

4. How to repair?

- If the verification result shows any incorrect part, click Repair to repair the translation.
- The repaired translation will be displayed in 4 .

5. How to auto-repair?

- Once the translation output appears in 2 , click Auto repair to directly go to the repair process without displaying the verification result.
- The repaired translation will be displayed in 4 .

6. Input Example

6.1 Translate “A mother duck hatches an egg”.

- Type “A mother duck hatches an egg”.
- Click E-T.
- Click Verify, the translation result is illustrated in Figure A2.

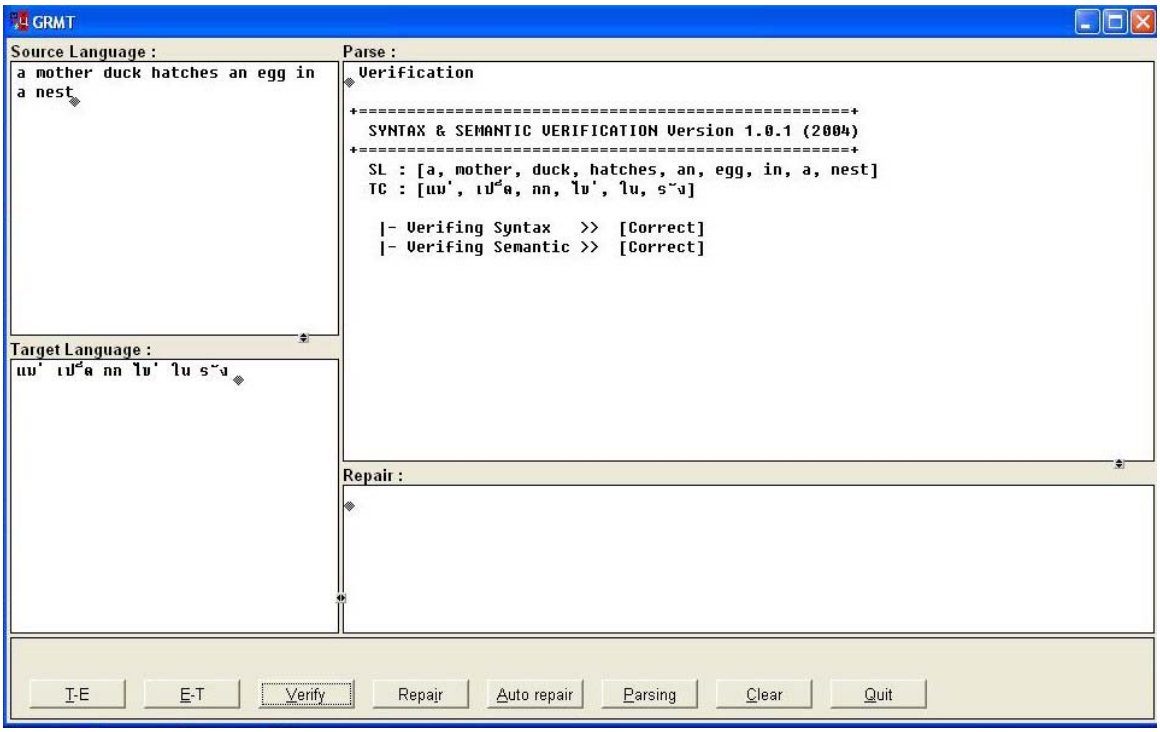


Figure A2 Translate and verify the input sentence: “A mother duck hatches an egg”.

6.2 Verify the translation pairs in Figure A3. (To demonstrate the repair ability of GRMT, we replaced the word “รัง” in the translation string in 2 with the word “กระท่อม” to make GRMT generate an incorrect translation.)

- Click Verify.
- Click Repair, the translation result is illustrated in Figure A4.

6.3 Repair the translation pairs in Figure A3.

- Click Auto repair, the translation result is shown in Figure A5.

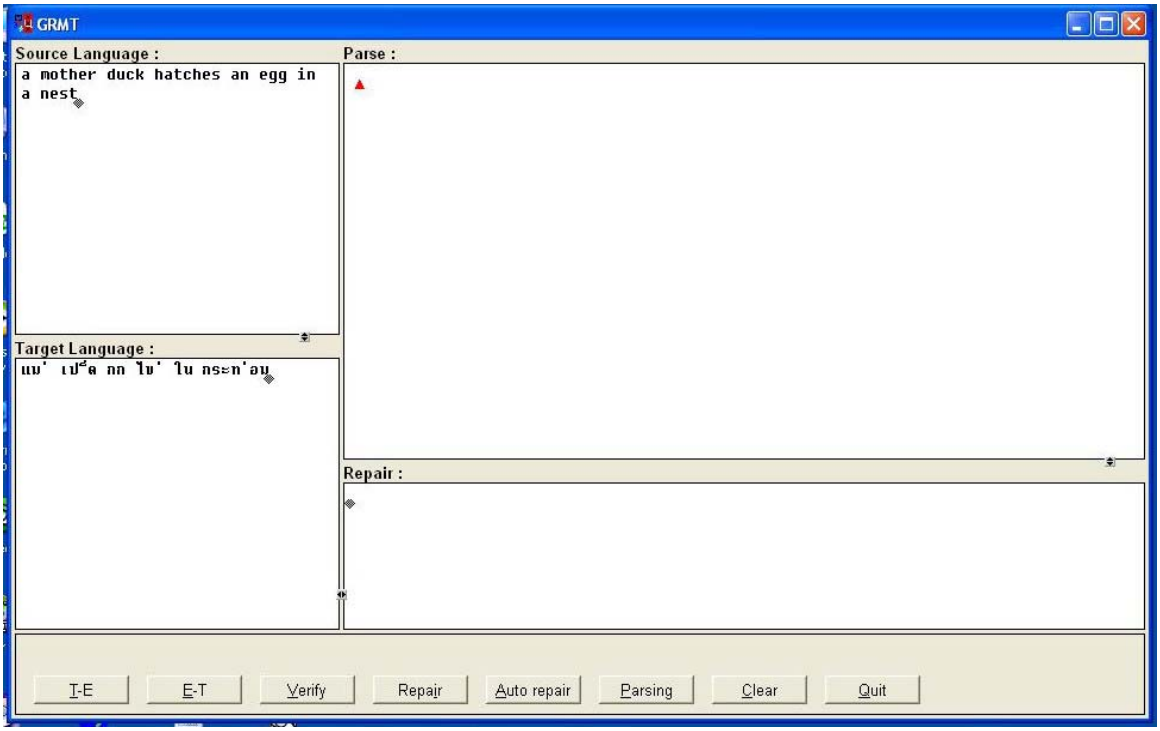


Figure A3 Translation of “A mother duck hatches an egg”.

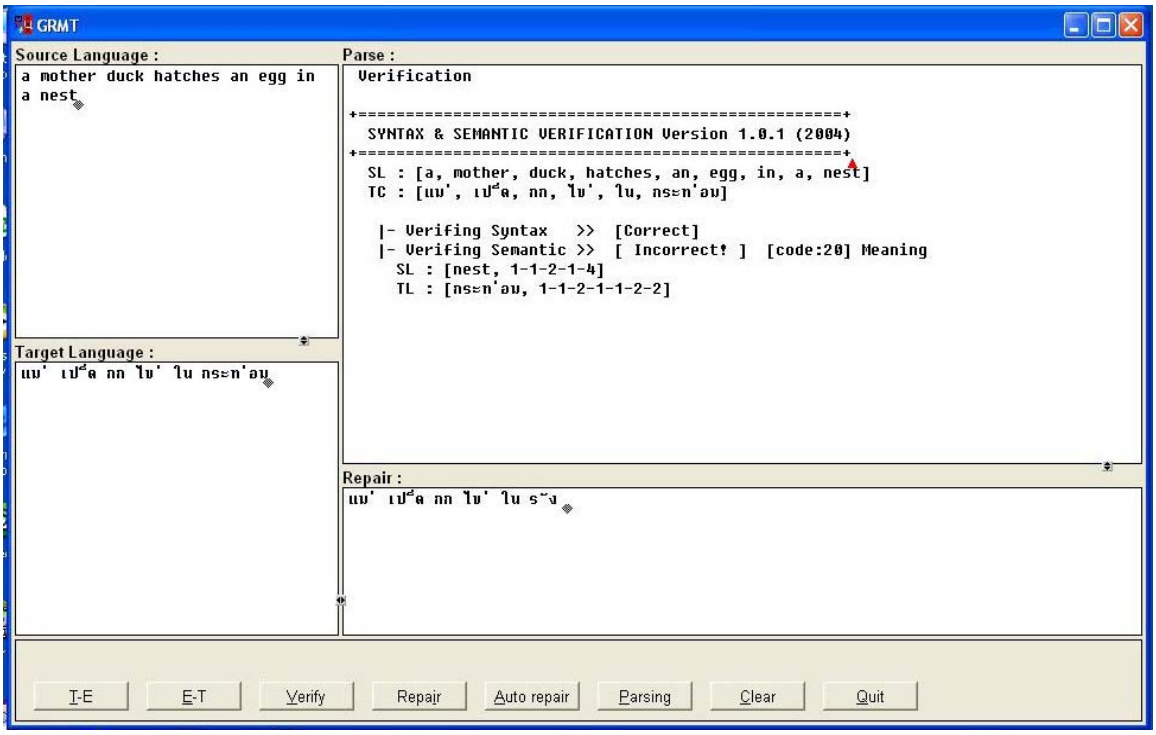


Figure A4 Translation of “A mother duck hatches an egg”, its verification and repaired translation.

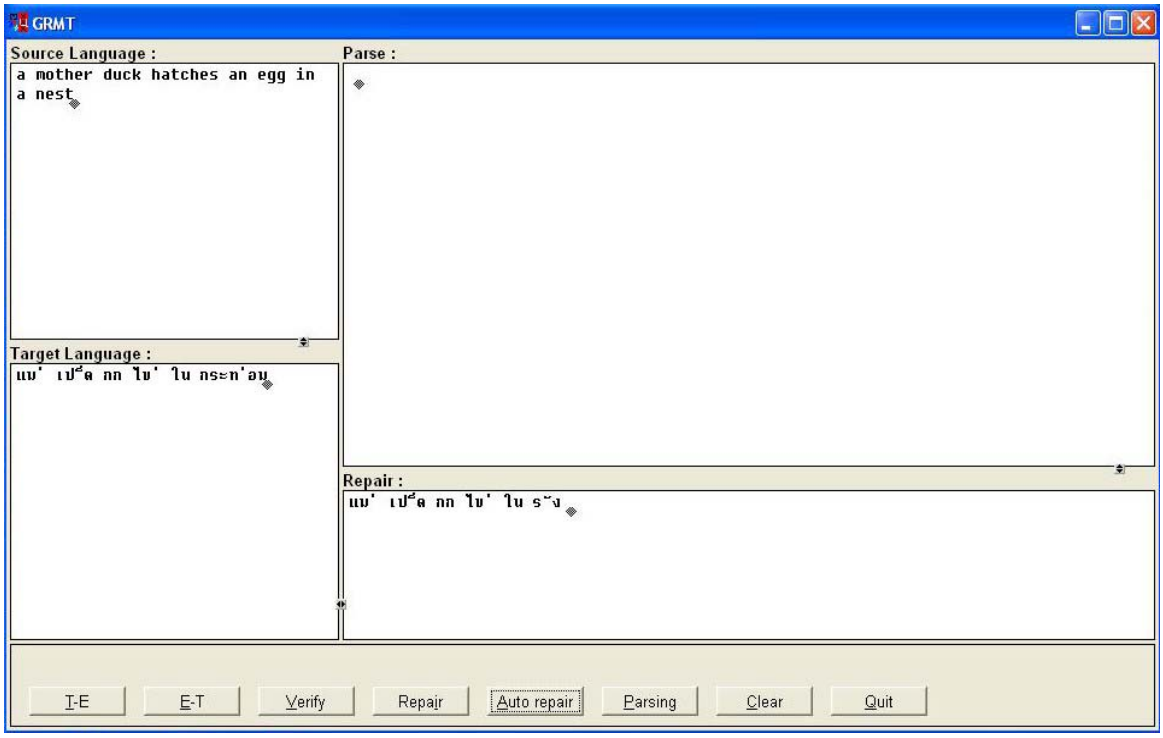


Figure A5 Translation of “A mother duck hatches an egg” and its repaired translation.

Appendix B

Pacific Association for Computational Linguistics 97

THAI WORD SEGMENTATION FOR ACCURATE TRANSLATION

JAKKREE NUNTADILOK
Department of Mathematics, Mahidol University, Bangkok, Thailand, jopal_jopal@hotmail.com
KANLAYA NARUEDOMKUL*
Department of Mathematics, Mahidol University, Bangkok, Thailand, scknr@mahidol.ac.th
NICK CERCONE
School of Computer Science, University of Dalhousie, Nova Scotia, Canada, nick@cs.dal.ca

The preliminary idea of a Thai Word Segmentation for Accurate Translation (TSAT) is proposed. TSAT is designed especially to increase the efficiency of a machine translation. The outstanding features of TSAT include simplicity, efficiency, and multilinguality.

To illustrate the performance of TSAT, we have tested the system by using it to segment an article in Thai. In this paper, the design of TSAT is discussed and some examples are presented.

Keywords: machine translation, word segmentation, Longest Matching, Maximum Matching, Feature based approach, Thai Character Cluster.

1. Introduction

Word segmentation is an important problem in Asian language, e.g. Chinese, Japanese, Korean and Thai, because there is no explicit word boundary delimiters. To overcome this problem, several segmentation algorithms were proposed. Some researchers attempt to increase an accuracy of their algorithms by using following techniques; *Longest Matching* [Aroonmanakun, 1997], *Maximum Matching* [Chuleerat, 1997], *Feature based approach* [Meknavin, 1997], and *Thai Character Cluster with statistical knowledge* [Theeramunkong et al., 2000]. Table 1 and 2 illustrates the comparison of these algorithms.

Word segmentation can provide more than one result as shown in Example 1 and 2. In Example 1, the sentence can be segmented into two different forms. However, only the first segmentation conveys the proper meaning. The second segmentation is grammatically incorrect and conveys no meaning. To make it simple for the reader, who is not familiar with Thai language, we use English sentence with no word boundary as examples. The second segmentation in Example 2 does not convey

* The author wish to acknowledge the support of the Thailand Research Fund (New Research Grant TRG4580108)

any meaning either. These examples illustrate that different segmentation means different interpretation. Therefore, segmenting word for translation must be accurate. Word segmentation is needed in a translation process of any language with no word boundary. Inappropriate segmentations can lead to a wrong translation result. For these reasons, we decided to find an alternative method, which can provide the correct segmentation for translation.

In this paper we propose the idea behind *Thai Word Segmentation for Accurate Translation* (TSAT) including some examples.

TABLE 1 Comparison of some existing word segmentation methods.

Method	Learning Algorithm	Dictionary Using	Statistical Computing	Accuracy of Segmentation
1.Longest matching	-	Use	Not use	54-97%
2.Maximal matching	-	Use	Not use	49-97%
3.Feature based approach	RIPPER and Winnow	Use	Use	91-99%
4.Thai Character Cluster	Decision tree (C4.5)	Not use	Use	~ 87%

TABLE 2 Advantages/Disadvantages of some word segmentation methods.

Method	Advantages	Disadvantages
1.Longest matching	Easy to build the program to segment the sentence. Not use statistical method.	It fails to find the corrected segmentation, because of its greedy characteristic.
2.Maximal matching	Easy to build the program. It can segment some words which the Longest matching cannot.	If the number of alternated word segmentations is the same, it cannot determine the best candidate.
3.Feature based approach	91-99% accurate.	The huge number of words in a dictionary and the training sets are required.
4.Thai Character Cluster	~ 87% accurate. Less memory space is required since there is no dictionary.	Less accuracy than Feature based approach.

Example 1	Therearemanyimportantthingsinhislife.
1st segmentation	There are many important things in his life.
2nd segmentation	The re ar em any im port ant things in his life.
Example 2	ไปหามเหสี
1st segmentation	ไป หา มเหสี
2nd segmentation	ไป หาม เห สี

2. Our Algorithm: Thai Word Segmentation for Accurate Translation

Thai Word Segmentation for Accurate Translation (TSAT) performs the segmentation in two phases: Word boundary tagging and Word boundary selection. *Word boundary tagging* is to find all possible word boundaries while *Word selection* is to find the most appropriate word boundary.

2.1 Word boundary tagging

The word tagging process begins at the end of the phrase/sentence. For example, Figure 1 presents the case that an input string which contains 30 characters. The tagging process begins at "X_{30}" and then searches for "X_{30}" in the available dictionary. If "X_{30}" is found, then TSAT tags "1" at "X_{30}". But if not, TSAT will search for "$X_{29}X_{30}$". Figure 2 illustrates the first tagging since "$X_{28}X_{29}X_{30}$" is found in the dictionary. The next step, TSAT looks for "X_{27}" and "$X_{27}X_{28}X_{29}X_{30}$". The system repeats the process until it hits the starting point as shown in Figure 3. The result of tagging process is shown in Figure 4.

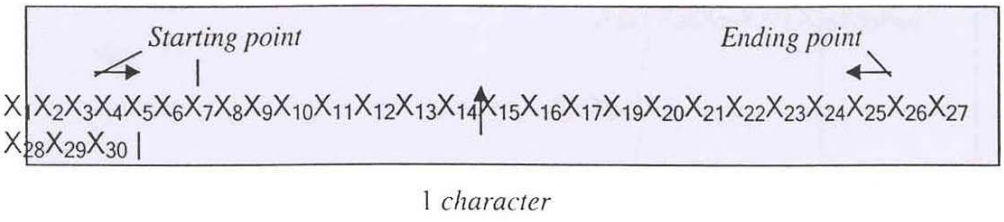
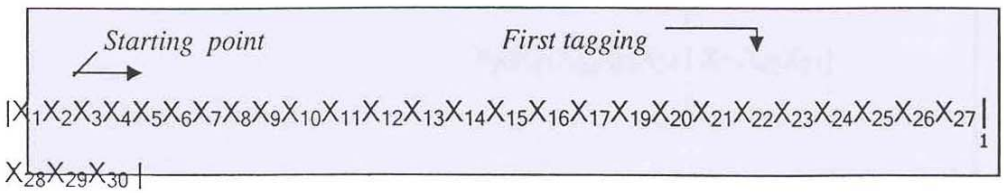


FIGURE 1. The segmenting phrase/sentence



1 *block: meaning* ⟷

FIGURE 2. The first tagging

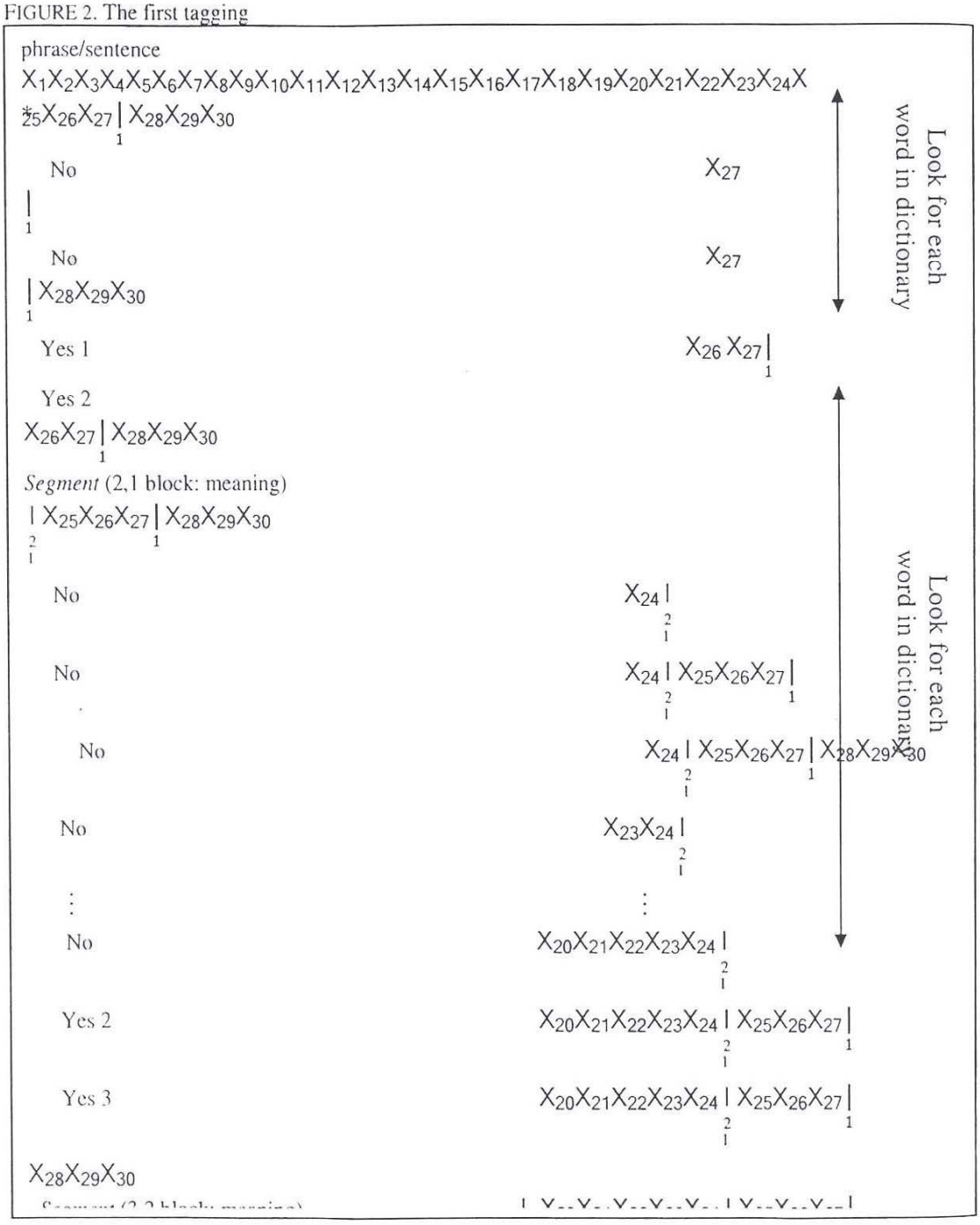


FIGURE 3. The tagging process (* No – The considered word cannot be found in the dictionary. Yes – The considered word can be found in the dictionary.)

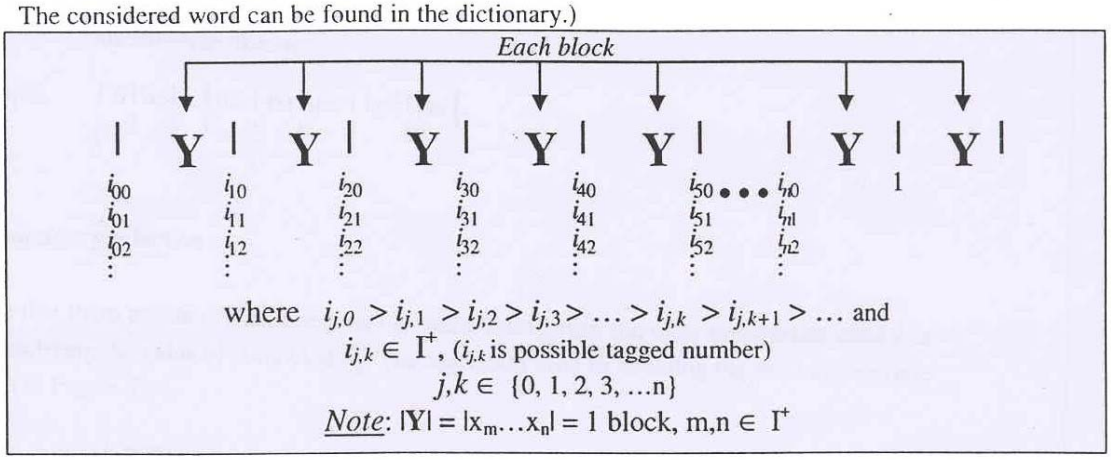


FIGURE 4. The tagged result

Input: $i_{j,k}$ is possible tagged number.

Output: A_j is boundary selection number.

Step 1. Consider at $i_{j,k}$ ($j = 0$, $k = 0$)

If it has only number $i_{0,0} = n$, then go to Step 3.

Else there are more than one number of $i_{j,k}$ (j-constant, k-running)

if $i_{j,k} = 1$, then go to Step 3.

Else if $i_{j,k} = m > 1$, compute value A_k by

$$A_j = \frac{[i_{j,k} - i_{j+1.0}] + [(i_{j,k} - 1) - i_{j+2,0}] + [(i_{j,k} - 2) - i_{j+3,0}] + \ldots + [2 - i_{j+m-2,0}]}{i_{j,k} - 1}$$

If $A_k > 0$, go to Step 3.

Else $A_k \leq 0$, go to Step 2.

Step 2.

Let $k = k + 1$.

If $i_{j,k} = 1$, then go to Step 3.

Else if $i_{j,k} = p > 1$, compute value A_i by

$$A_j = \frac{[i_{j,k} - i_{j+1.0}] + [(i_{j,k} - 1) - i_{j+2,0}] + [(i_{j,k} - 2) - i_{j+3.0}] + \ldots + [I - i_{j+p-2}]}{i_{j,k} - 1}$$

If $A_k > 0$, go to Step 3.

Else $A_k \leq 0$, repeat Step 2.

where $I = i_{j,0} + 2 - i_{j,k}$

Step 3.

Count $i_{j,k}$ block from left to right and then segment it to a segmented word.

Let $j = j + 1$.

Repeat step 1.

FIGURE 5 Word boundary selection algorithm ($i_{j,k}$, A_j)

Example 3 Sheisthehousekeeper.

Segmented output | S | he | is | the | ho | use | kee | per |.
2 1 1 1 4 1 2 1
1 2
1

2.2 Word boundary selection

In the case that there are more than one possible word boundaries, the most appropriate choice is selected by considering the value of computed A_j. The algorithm used in selecting the most appropriate choice is shown in Figure 5.

Example 4: onepitfallofthissystem

Word boundary tagging

onepitfallofthissyst | em :1 block:meaning
1
('em' has meaning in English)
onepitfallofthissy | st | em :2 block:meaning
2 1
('stem' has meaning in English)
onepitfallofthis | sy | st | em :3 block:meaning
3 2 1
('system' has meaning in English)
onepitfallofth | is | sy | st | em :1 block:meaning
1 3 2 1
('is' has meaning in English)
onepitfalloft | h | is | sy | st | em :2 block:meaning
2 1 3 2 1
('his' has meaning in English)
onepitfallof | t | h | is | sy | st | em :3 block:meaning
3 2 1 3 2 1
('this' has meaning in English)
onepitfall | of | t | h | is | sy | st | em :1 block:meaning
1 3 2 1 3 2 1
('of' has meaning in English)
onepitfal | l | of | t | h | is | sy | st | em :3 block:meaning
3 1 3 2 1 3 2 1
('loft' has meaning in English)
onepitf | al | l | of | t | h | is | sy | st | em :2 block:meaning
2 3 1 3 2 1 3 2 1
('all' has meaning in English)

onepit | f | al | l | of | t | h | is | sy | st | em :3 block:meaning
3 2 3 1 3 2 1 3 2 1

('fall' has meaning in English)

onep | it | f | al | l | of | t | h | is | sy | st | em :1 block:meaning
1 3 2 3 1 3 2 1 3 2 1

('it' has meaning in English)

one | p | it | f | al | l | of | t | h | is | sy | st | em :5,2 block:meaning
5 1 3 2 3 1 3 2 1 3 2 1
2

('pit' and 'pitfall' have meaning in English)

| one | p | it | f | al | l | of | t | h | is | sy | st | em :1 block:meaning
1 5 1 3 2 3 1 3 2 1 3 2 1
2

('one' has meaning in English)

Word boundary selection

| one,
1

(*use* ***step 1***, $i_{0.0} = 1$, *there is only number* "1", *then count* 1 *block from left to right and segment it*)

[(5-1)+(4-3)+(3-2)+(2-3)]/4=1.25>0 | pitfall,
5

(*use* ***step 1***, *there are two number* "5 *and* 2", $i_{1.0} = 5$ *then compute* $A_1 = 1.25>0$, *count* 5 *block from left to right and segment it*)

| of, (*use* ***step 1***, $i_{7.0} = 1$, *there is only number "1"* , *then count* 1 *block from left to*
1

right and segment it)

[(3-2)+(2-1)]/2=1>0 | this,
3

(*use* ***step 1***, $i_{8.0} = 3$, *there is only number* "3", *then count* 3 *block from left to right and segment it*)

[(3-2)+(2-1)]=1>0 | system
3

(*use* ***step 1***, $i_{11.0} = 3$, *there is only number* "3", *then count* 3 *block from left to right and segment it*)

The above step can write in short form :

| one, [(5-1)+(4-3)+(3-2)+(2-3)]/4=1.25>0 | pitfall, | of, [(3-2)+(2-1)]/2=1>0 | this,
1 5 1 3

[(3-2)+(2-1)]=1>0 | system.
3

Segmentation : one pitfall of this system

Example 5: thelasthomeworkofthisweek

Word boundary tagging

| t | hc | last | ho | me | work | of | t | h | is | week
2 1 1 3 1 1 1 3 2 1 1
2
1

Word boundary selection

(2-1)=1>0 | the, | last, [(3-1)+(2-1)]/2=1.5>0 | homework, | of, [(3-2)+(2-1)]/2=1>0 | this,
2 1 3 1 3

| week
1

Segmentation : the last homework of this week

Example 6: เขาใช้ยานอกรักษาอาการ

Word boundary tagging

| เขา | ใช้ | ยา | น | อก | รักษา | อา | การ
1 1 2 2 1 1 2 1 1
1

Word boundary selection

| เขา
1

(*use* ***step 1***, $i_{0,0} = 1$, *there is only number* "1", *then count* 1 *block from left to right and segment it*)

| ใช้
1

(*use* ***step 1***, $i_{1,0} = 1$, *there is only number* "1", *then count* 1 *block from left to right and segment it*)

(2-2)=0≤0 → | ยา
1

(*use* ***step 1***, *there are two number* "2 *and* 1", $i_{2,0} = 2$ *then compute* $A_2 = 0 \leq 0$, *go to* ***step2***, , $i_{2,1} = 1$, *there is only number* "1", *then count* 1 *block from left to right and segment it*)

| นอก
2

(*use* ***step 1***, $i_{3,0} = 2$, *there is only number* "2", *then count* 2 *block from left to right and segment it*)

| รักษา
1

(*use* ***step 1***, $i_{4,0} = 1$, *there is only number* "1", *then count* 1 *block from left to right and segment it*)

(2-1)=1>| อาการ
2

(*use* ***step 1***, *there are two number "2 and 1", $i_{5,0}$ =2 then compute* $A_5 = 1 > 0$
count 2 block from left to right and segment it)

The above step can write in short form :

| เขา, | ใช้, (2-2)=0≤0 → | ยา, (2-1)=1>0 | นอก, | รักษา, (2-1)=1>| อาการ
1 1 1 2 1 2

Segmentation : เขา ใช้ ยา นอก รักษา อาการ

Example 7: การเก็บภาษีในประเทศไทย

Word boundary tagging

| การ | เก็บ | ภาษี | ใน | ประ | เทศ | ไทย
1 1 1 1 3 1 1

Word boundary selection

| การ, | เก็บ, | ภาษี, | ใน, [(3-1)+(2-1)]/2=1.5>0 | ประเทศไทย
1 1 1 1 3

Segmentation : การ เก็บ ภาษี ใน ประเทศไทย

Example 8: ขอเวลาทำความรู้จักกับตัวเองก่อน

Word boundary tagging

| ข | อ | เว | ลา | ทำ | ความ | รู้ | จัก | กับ | ตัว | เอง | ก่อน
2 1 2 1 1 2 2 1 1 1 1 1
1 1

Word boundary selection

(2-1)=1>0 | ขอ, (2-1)=1>0 | เวลา, | ทำ, (2-2)=0≤0→ | ความ, (2-1)=1>0 | รู้จัก, | กับ, | ตัว, | เอง,
2 2 1 1 2 1 1 1

(2-1)=1>0 | ก่อน
1

Segmentation : ขอ เวลา ทำ ความ รู้จัก กับ ตัว เอง ก่อน

Example 9: เขายืนตากลมอยู่ที่หน้าบ้าน

Word boundary tagging

| เ | ขา | ยืน | ตา | ก | ลม | อยู่ | ที่ | ห | น้า | บ้าน
2 1 1 2 2 1 1 1 2 1 1
1 1

Word boundary selection

(2-1)=1>0 | เขา, | ยืน, (2-2)=0≤0 → | ตา, (2-1)=1>0 | กลม, | อยู่, | ที่, (2-1)=1>0 | หน้า, | บ้าน

Segmentation : เขา ยืน ตา กลม อยู่ ที่ หน้า บ้าน

Example 10: ฉันเอาผ้าขนหนูไปตากลมที่หน้าบ้าน

Word boundary tagging

| ฉัน | เ | อา | ผ้า | ขน | หนู | ไป | ตา | ก | ลม | ที่ | ห | น้า | บ้าน

Word boundary selection

| ฉัน, | (2-1)=1>0 | เอา, [(3-1)+(2-1)]/2=1.5>0 | ผ้าขนหนู, | ไป, (2-2)=0≤0→ | ตา, (2-1)=1>0 | กลม, | ที่, (2-1)=1>0 | หน้า, | บ้าน

Segmentation : ฉัน เอา ผ้าขนหนู ไป ตา กลม ที่ หน้า บ้าน (incorrect segmentations)*

3. Conclusion

TSAT is an alternative word segmentation method that can efficiently serve machine translation. The word segmentation is required in machine translation which translates from any Asian language (with no word boundary). TSAT comprises of two phrases: word boundary tagging and word boundary selection.

TABLE 3. Comparisons of segmentation results

TSAT		Longest Matching		Maximal matching	
she is the housekeeper	/	she is the ho use keeper	-	she is the housekeeper	/
one pitfall of this system	/	one pit fall of this system	-	one pitfall of this system	/
last homework of this week	/	last ho me work of this week	-	last homework of this week	/
re pair the building	/	rep air the building	-	*Cannot determine the best candidate*	* *(1)*
เขา ใช้ ยา นอก รักษา อาการ	/	เขา ใช้ ยาน อก รักษา อาการ	-	*Cannot determine the best candidate*	* *(2)*
การ เก็บ ภาษี ใน ประเทศไทย	/	การ เก็บ ภาษี ใน ประ เทศ ไทย	-	การ เก็บ ภาษี ใน ประเทศไทย	/
ขอ เวลา ทำ ความ รู้จัก กับ ตัว เอง ก่อน	/	ขอ เวลา ทำ ความรู้ จัก กับ ตัว เอง ก่อน	-	*Cannot determine the best candidate*	* *(3)*
เขา ยืน ตา กลม อยู่ ที่ หน้า บ้าน	/	เขา ยืน ตาก ลม อยู่ ที่ หน้า บ้าน	/	*Cannot determine the best candidate*	* *(4)*

เอา ผ้าขนหนู ไป ตา กลม ที่ หน้า บ้าน	-	เอา ผ้า ขน หนู ไป ตาก ลม ที่ หน้า บ้าน	-	*Cannot determine the best candidate*	* (5)

Note:

/ = grammatically correct

- = grammatically incorrect

*(1) = It can be "re pair the building" or "rep air the building".

*(2) = It can be "เขา ใช้ ยา นอก รักษา อาการ" or "เขา ใช้ ยาน อก รักษา อาการ".

*(3) = It can be "ขอ เวลา ทำ ความ รู้จัก กับ ตัว เอง ก่อน" or "ขอ เวลา ทำ ความรู้ จัก กับ ตัว เอง ก่อน".

*(4) = It can be "เขา ยืน ตา กลม อยู่ ที่ หน้า บ้าน" or "เขา ยืน ตาก ลม อยู่ ที่ หน้า บ้าน".

*(5) = It can be "เอา ผ้าขนหนู ไป ตา กลม ที่ หน้า บ้าน" or "เอา ผ้าขนหนู ไป ตาก ลม ที่ หน้า บ้าน".

The designed TSAT algorithm is tested to segment Thai articles. Table 3 presents some segmented results of the three different algorithms. The TSAT generated the better results than the others.

4. References

Aroonmanakun, W., 1997. Collocation and Thai Word Segmentation. Department of Linguistics, Faculty of Arts, Chulalongkorn University.

Chuleerat, J., 1997, Dictionary-based Thai CLIR: Experimental Survey of Thai CLIR, Department of Computer Science, Faculty of Science, Kasetsart University.

Meknavin, S., Charoenpornsawat, P. and Kijsirikul, B., 1997, Feature-based Thai Word Segmentation,

Theeramunkong, T., Usanavin, S., Machomsomboon, T. and Opasanont. B. 2000. Thai Word Segmentation Without a Dictionary by Using Decision Tree, SNLP, The Fourth Symposium on Natural Language Processing.

Appendix C

TOWARDS MULEX – A MULTILINGUAL LEXICAL DATABASE SYSTEM FOR MACHINE TRANSLATION

CHARNYOTE PLUEMPITIWIRIYAWEJ*

Department of Computer Science, Mahidol University, Bangkok, Thailand, cccpt@mahidol.ac.th

KANLAYA NARUEDOMKUL+

Department of Mathematics, Mahidol University, Bangkok, Thailand, scknr@mahidol.ac.th

NICK CERCONE

Faculty of Computer Science, Dalhousie University, Nova Scotia, Canada, nick@cs.dal.ca

The preliminary idea of a multilingual lexical database system for machine translation (MULEX) is proposed. MULEX is designed especially to increase the efficiency of multilingual machine translation. The outstanding features of MULEX include simplicity, efficiency, extendibility and multilinguality. MULEX can be used as a dictionary and a thesaurus via a web-based interface as well.

To illustrate the performance of MULEX, we have integrated it with Generate-and-Repair machine translation system (GRMT). In this paper, the building of MULEX database is discussed and the integration between MULEX and GRMT is presented.

Key words: multilingual lexical database, machine translation, Generate-and-Repair machine translation.

1. INTRODUCTION

Lexical databases have become very important information resources for a variety of applications such as language parsing, information retrieval (IR) and machine translation (MT). For language parsing, the syntactic information in the databases is used to explain the relationships between words (structure) in a sentence. For IR, the semantic information is used to disambiguate the meaning of words in texts and the synonymy is used for query expansion to improve the recall. For MT, both the syntactic and the semantic information are used to translate sentences from one language into another or others in multilingual machine translation system (MMT).

One of the most important components of MT systems is the lexical database. The size of a lexical database for MT, especially MMT, is relatively large. In the past, the development of the lexicon was limited due to the lack of powerful computer and that of a storage space. Presently, with the advance in computer technologies, those issues are no longer problems. A current problem is an ability to manage a large amount of information in the lexical database.

The design of the lexicon for each MT system basically depends on an approach used in developing that system. In Generate-and-Repair Machine Translation (GRMT) approach [Naruedomkul and Cercone 2000] (see section 2), the designed lexicons contain both syntactic and semantic information of word entries. Research in GRMT is divided into 3 phases: single-directional translation, bi-directional translation and multi-directional translation. In the first two phases, the lexicons were developed in Prolog representation. In the third phase, the lexicons became larger and more complicated. Therefore, the lexicons used in the original system were re-designed and, whenever possible coalesced. Proper design is required once we step from bi-directional translation to a true multilingual MT system.

* Supported by the Faculty of Science, Mahidol University, under the Young Researcher Fund.
+ Supported by the Thailand Research Fund (TRF), under the New Research Grant (TRG-4580108).

Several projects of lexical database have been proposed, for example, Papillon [Mangeot 2000; Serasset and Mangeot 2001], SAIKAM [Ampornnaramveth and Methapisit 2000], MatsLex [Tiedemann 2002], PARAX [Blanc 1999], LOLA [Blaser et al. 1992]. PARAX database stores lexical data in hypertextual form, while the others use relational database to store lexical data. MatLex and SAIKAM databases contain English-Swedish lexicons and Japanese-Thai lexicons, respectively. They also provide a Web-based interface to allow user to access the databases conveniently. After we have studied and compared different lexical database projects, we have concluded in favor of our MUltilingual LEXical database that is designed to provide a simple and extendible environment for storing and managing multilingual lexical data. Our approach uses the existing relational database management system (RDBMS) technology to manage a large amount of lexical data. The process of creating the lexical database can be accelerated by reusing the existing lexicons when it is possible.

In this paper, we present our preliminary idea of designing and building a lexical database via MULEX framework to support machine translation system. The implementation of MULEX uses RDBMS technology as a tool for simplicity, extendibility and efficiency. In addition to support machine translation, our lexical database has several added values. It can be used as a regular dictionary which provides meaning and syntax of words. MULEX also can be used as a wordbook which provides synonym, antonym and hypernym[1] of words.

The rest of this paper is organized as follows. In section 2, we present the existing lexical databases and machine translation systems which are the motivations of this research. We introduce our preliminary idea of MULEX system in section 3. An initial model and an implementation of MULEX database are illustrated in section 4 and 5, respectively. The last section summarizes our contribution.

2. GRMT AND ITS LEXICON

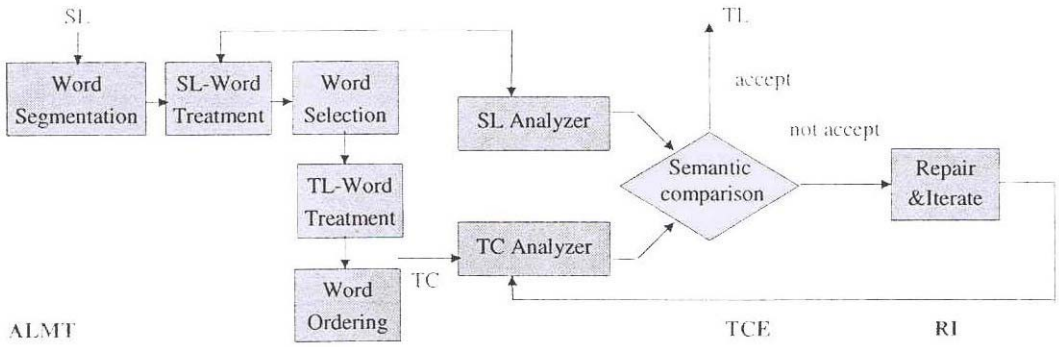


FIGURE 1. GRMT Architecture

Generate and Repair Machine Translation (GRMT) was first proposed in 1996 as an alternative and novel approach to machine translation [Naruedomkul and Cercone 1997]. It is a constraint-based approach to machine translation that focuses on accurate translation output [Naruedomkul and Cercone 2000]. GRMT comprises three modules as shown in Figure 1: Analysis Lite Machine Translation (ALMT), Translation Candidate Evaluation (TCE) and Repair and Iterate (RI). ALMT generates the translation which we call translation candidate (TC), next

[1] Hypernym is the generic term used to designate a whole class of specific instances [Fellbaum 1999]. Y is a hypernym of X if X is a (kind of) Y. For example: bird is a hypernym of robin.

TCE verifies the TC to see whether it retains the meaning of the original sentence. If the TC does retain the original meaning, the TC is then redeemed as the translation. But if the TC does not retain the original meaning, the TC will be repaired by RI. Basically, the translation process generates the translation candidate and then repairs it when necessary.

In the first phase of GRMT, single-directional translation, the English-Thai MT (ETMT) system was developed to translate from English into Thai. Three types of lexicons: SL dictionary, bi-lingual dictionary and TCE lexicon were built for this system. The SL dictionary contains English word entries and their syntactic categories. Bilingual dictionary contains English word entries and all of their possible corresponding words in target language together with their word association number [Naruedomkul and Cercone 1999]. Word association number is assigned to a word class. Words are classified according to their meaning and usage. TCE lexicon contains both syntactic and semantic information of the word entries in terms of feature structure. The SL dictionary and bilingual dictionary are represented in Prolog. TCE lexicon is encoded in using the Attribute Logic Engine (ALE) version 3.2 Beta. ALE is an integrated phrase structure parsing and definite clause logic programming system in which the terms are typed feature structures [Carpenter and Penn 1999].

In the second phase of GRMT, the Thai-English MT (TEMT) system was developed. Four types of lexicons and five templates were added to the system. The SL dictionary contains Thai word entries and their syntactic categories. Bilingual dictionary contains Thai word entries and all of their possible corresponding words in English together with their word association number. The TL dictionary contains English word entries and their syntactic categories. TCE lexicon contains the same information as the TCE lexicon of ETMT. Five templates include uncountable noun, past simple form of verb, past participle form of verb, Comparative and Superlative KB.

To develop the bi-directional translation, English⇔Thai MT (E⇔TMT) system, we integrated ETMT with TEMT. The E⇔TMT is able to translate back and forth between English and Thai. After integration, some lexicons in E⇔TMT are redundant. We need a new lexicon databases design to make the system more efficient and ready for the third phase, multi-directional translation. To re-design a lexicon databases, we have borrowed the technique used in Princton's WordNet [Fellbaum 1998], EuroWordNet [Vossen 1998] and MultiWordNet [Pianta et al. 2002] Projects to classify and relate words according to their meanings.

3. THE PRELIMINARY IDEA OF MULEX

MULEX is a multilingual lexical database system that is designed to provide a simple and extendible environment for storing and managing lexical data that is from multiple languages. It provides a central lexical database that stores all lexical data and makes them available for MT systems. The primary goal of MULEX is to support the translation process in any MT system. MULEX can also be used as an online dictionary that can be searched for meaning of words and for related words via a web-based interface. In MULEX, the database is designed to be well organized and efficiently accessed. The fundamental principle of building MULEX database is to reuse lexical data that is stored in existing lexicons when it is possible.

Figure 2 illustrates a conceptual overview of the MULEX system which consists of the *database*, the *adapters*, the *manager* and the *web-based interface* (WBI). The MULEX database contains phonological, morphological, syntactic, and semantic information of words and appropriate relations between words. The database structure is modeled according to the requirements of MT systems. The database model design and the database creation are discussed in the next two sections.

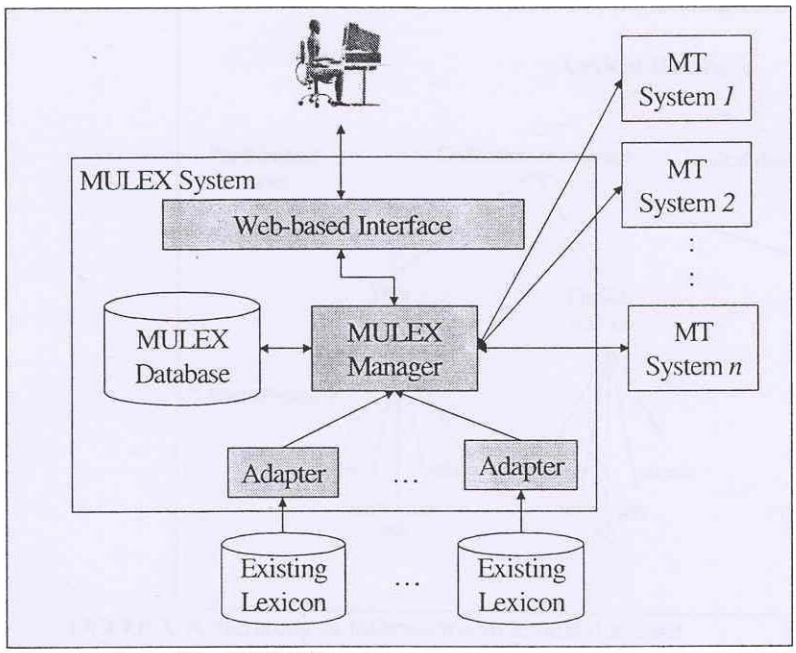


FIGURE 2. An overview of the MULEX system

The MULEX adapters perform all operations associated with the extraction and conversion of lexical data from existing lexicons. Some existing lexicons may contain lexical data that is irrelevant to information needed for MT because they are designed for other purposes. Only relevant lexical data is extracted from the existing lexicons. In addition to the extraction, the adapter may have to perform the data conversions. The data conversions are captured in the form of mappings, which are generated when the adapter is configured. The mappings are performed via the operations that may include simple transformations of data to prepare the data for entry into the database. The complexity of the adapter varies between existing lexicons.

The MULEX manager performs all the operations associated with the management of the data in the database and the user queries. It deals with two main tasks. First, it performs analysis, loading, reconciling and cleansing of related lexical data from the external lexicons. Second, it supports querying from the front-end users and the MT systems.

The WBI provides a graphical web-based user interface to allow front-end users to query and browse the lexical data and shows it in easy-to-understand form. The user queries can be seen as requests of searching for word meanings and related words.

4. MULEX DATABASE MODEL

Our *lexical database* is designed to meet the requirements of machine translation systems. Figure 3 outlines our lexical database model. We model the lexical database as a collection of *lexicons* (e.g., Thai lexicon, English lexicon) and a *lexical database schema* which is a description of the lexicons. Each lexicon contains sets of words which are organized hierarchically according to their part-of-speech.

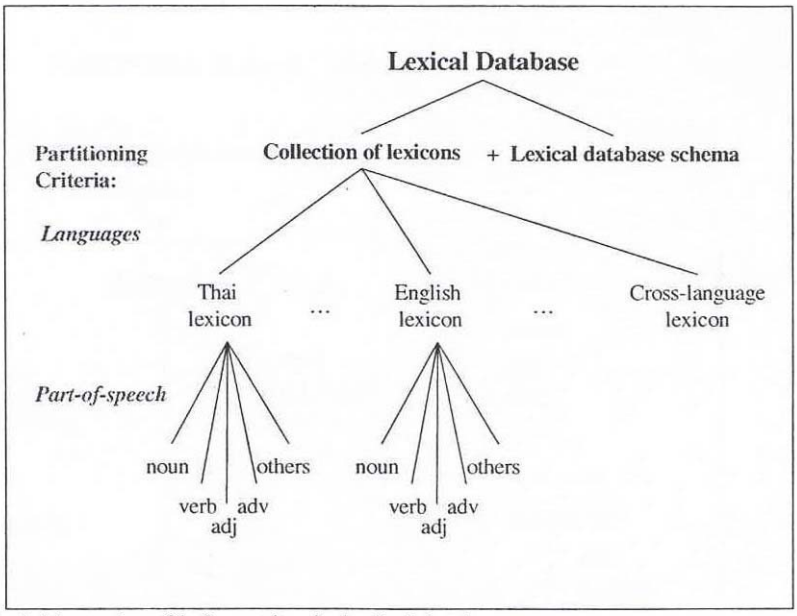


FIGURE 3. A hierarchy of information in lexical database

4.1 Lexicon

Each lexicon contains an ordered list of *lexical entries*. Each lexical entry represents an occurrence of word or an occurrence of relationship between words. We refer to an occurrence of word as a *lexical item* and an occurrence of relationship between words as a *lexical association*. The lexical items that have the same characteristics are said to be of the same *lexical item type*. Likewise, the lexical associations that have the same characteristics are said to be of the same *lexical association type*.

Based on language types, a lexicon can be divided into two types: a *monolingual* type and a *cross-language* type. The monolingual lexicons will provide information that describes lexical items of a particular language and their lexical associations, called *internal associations*. The cross-language lexicons will provide information that describes relationships among lexical items or groups of lexical items from different languages. The occurrences of such relationships are referred to as the *external associations*.

In the monolingual lexicon, the information that describes lexical items includes a phonological description (e.g., pronunciation), a syntactic description (e.g., part-of-speech), a semantic description of lexical items, and sample sentences in which the lexical item is used. Internal associations are defined upon the requirements of each language. For example, in English, the word "ate" is a past simple of the word "eat"; hence, the IsaPastOf association is required to represent their relationship. Such association may not occur in other languages such as Thai.

In the cross-language lexicon, the information that describes an external association includes an association type, an association description and a mapping between lexical items that represent words from different languages. As in [Naruedomkul and Cercone 2000], a hierarchical structure is used to represent relations of association types. Each association type has a description and an attached unique number which is used as an important feature of mappings between lexical items. A part of the hierarchical structure of association types is shown in Figure 4.

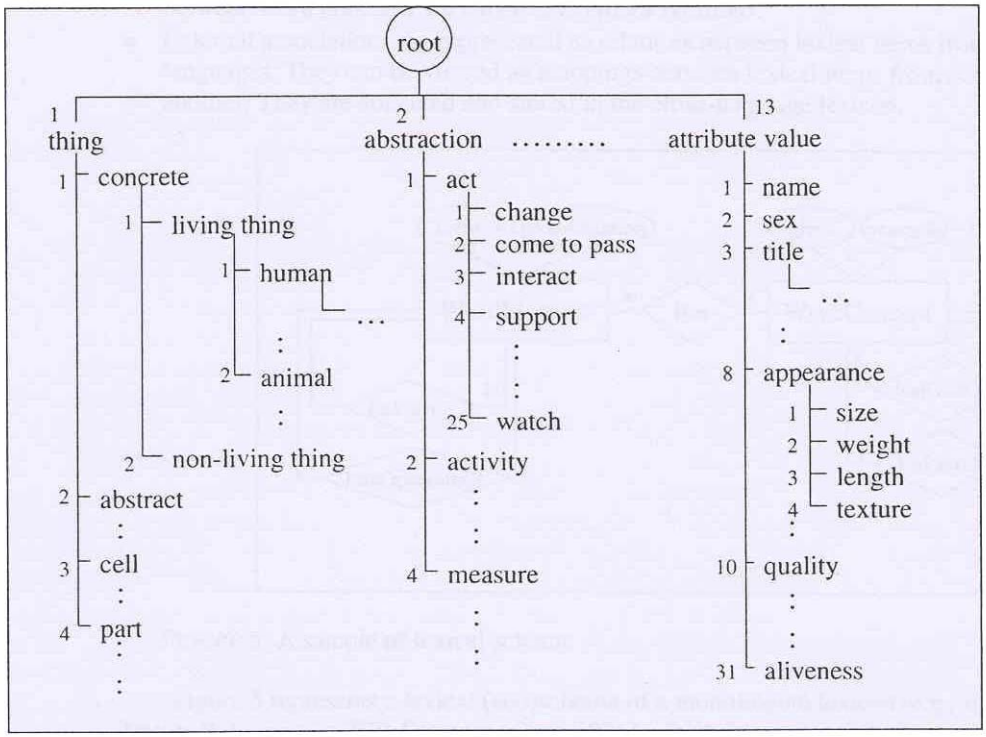


FIGURE 4: A sample of association type hierarchy

In our current lexical database, we create two monolingual lexicons–English lexicon and Thai lexicon–and a cross-language lexicon to support the translation between Thai and English. We can extend the database by including lexicons of other languages and updating cross-language lexicon to support translations of other languages.

4.2 Lexical database schema

A *lexical database schema* is a description of lexical database which contains a collection of lexicons. The term *lexical schema* is used to refer to a part of lexical database schema that describes the structure of a particular lexicon which can be outlined as follows:

- Each lexicon is partitioned into several parts according to lexical item types or *word category*. As shown in Figure 3, we use the part-of-speech, which is a kind of grammatical information, to classify the type of lexical item.
- Each lexical item has two main components: *word element* and *word concept*. The word element represents the physical existing form of word and its phonological information whereas the word concept represents the semantic information of word.
- Some word elements may have several different word concepts, and some word concepts may be expressed by several different word elements. Therefore, mappings between word elements and word concepts are many-to-many. A word element is *polysemous* if it can be mapped to

different word concepts (i.e., it is has multiple meanings). Two or more word elements are *synonymous* if they are mapped to the same word concept (i.e., one is a synonym of another).

- Internal associations can be represented as relations between lexical items in monolingual lexicon. The relations between word elements are called *elemental relations*, and the relations between word concepts are called *conceptual relations*.
- External associations are represented as relations between lexical items from different languages. They can be viewed as mappings between lexical items from one language to another. They are collected and stored in the cross-language lexicon.

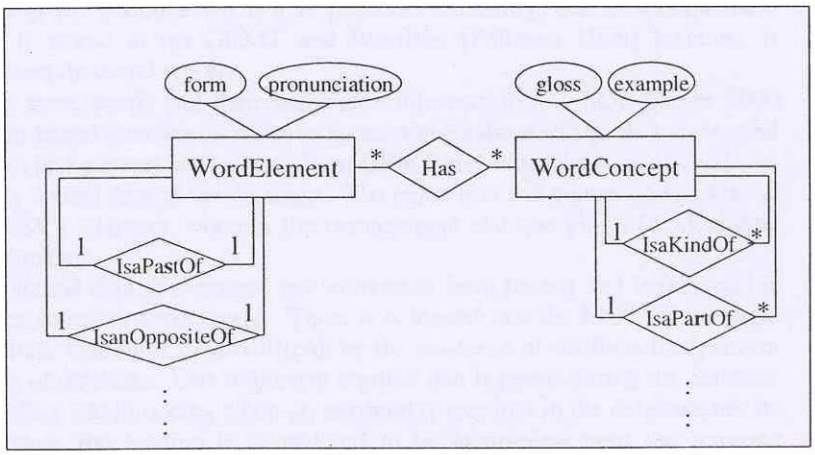


FIGURE 5: A sample of lexical schema

Figure 5 represents a lexical (sub)schema of a monolingual lexicon (e.g., English) by using Entity-Relationship (ER) Diagram [Chen 1976], which uses rectangular boxes to represent entity types, diamond-shaped boxes to represent relationship types, and ovals to represent attributes of entities or relationships. In Figure 5 two rectangular boxes represent a set of word elements and a set of word concepts, respectively. Each word element has two attributes, namely the 'form' which describes the physical utterance of word, and the 'pronunciation' which describes a kind of phonological information. Other kinds of information (e.g., word sound) which are not shown in the figure are attached as attributes to describe properties of word elements. Each word concept has two attributes, namely the 'gloss' which describe the meaning of word, and the 'example' which provides sample sentences in which the word is used. Mappings between word elements and word concepts are represented as the 'Has' relationship whose cardinality ratio is many-to-many as shown via the lines labeled with the '*' symbols (i.e., the '*' symbol means 'many').

In Figure 5, only two elemental relations, namely 'IsaPastof' and 'IsanOppositeOf', are showed. Other relations can be defined upon the specification of each language. Each of those relations takes two word forms as parameters and describes the relationships between word elements. For example, IsaPastOf(rose, rise) indicates that the word "rose" is a past of the word "rise" and IsanOppositeOf(rise, fall) indicates that the word "rise" is an opposite of the word "fall".

Unlike elemental relations, conceptual relations describe relationships between word concepts. The 'IsaKindOf' relation in Figure 5 describes class/sub-class relationship. For example, an eye which is a sense organ is a kind of organ. The 'IsaPartOf' relation in Figure 5 describes part-whole relationship. For example, an eye is a part of the whole face.

5. BUILDING MULEX DATABASE

Regarding our database model explained in previous section, we have preliminarily built our large-scale lexical database, namely MULEX database. We start building the database that is designed to meet the requirements of the Generate-and-Repair Machine Translation (GRMT) System [Naruedomkul and Cercone 2000] which supports multilingual translation. In order to build the large-scale lexical database quickly, our fundamental principle is to reuse lexical data that is stored in existing available lexicons when it is possible. Currently, our lexical database combines lexical data that is stored in the GRMT and WordNet [Felbaum 1999] lexicons. It contains more than one hundred thousand words.

In MULEX system, we store words and their associated information into SQL Server 2000 database. We produce a Web-based interface in order to access the database easily. We create and develop programs that extract and convert lexical data from GRMT and WordNet lexicons and that efficiently manage and query lexical data in the database. The extraction and conversion of lexical data are performed in MULEX's adapters, whereas the management and querying of lexical data are performed in MULEX manager.

To build the database, lexical data is extracted and converted from Prolog and lexicographic files in GRMT and WordNet sources, respectively. Then, it is loaded into the MULEX database by MULEX manager. Database loading is made difficult by the existence of conflicts that prevent a straightforward completion of database. One important conflict that happens during the database loading is *missing value* conflict which occurs when an attribute is required in the database but its value is missing. In this case, the loading is considered to be incomplete until the required attributes, which is necessary for the translation process in GRMT, is carried out.

When the database has been set up, GRMT developers can use it without modifying the translation procedures. This leads to the simplicity which is one of MULEX's features. The database can be enlarged either by using a program interface or via Prolog adapters. If the Prolog adapters are used, the loading process is repeated.

6. CONCLUSION

MULEX is a multilingual lexical database system that can efficiently serve machine translation. It provides a simple and extendible environment for storing and managing lexical data from different languages. MULEX provides a web-based interface to enable users to search for word meanings and related words.

We have outlined the conceptual design of a multilingual lexical database that stores words and their associated information. MULEX database consists of a collection of lexicons and a description of database structure. It is designed to be well organized and efficiently accessed. The fundamental principle of building MULEX database is to reuse lexical data that is stored in existing lexicons when it is possible.

REFERENCES

[Ampornnaramveth and Methapisit 2000] Vuthichai Ampornnaramveth and Tasanee Methapisit. 2000. Automatic Word Lookup Service and Client Tool for SAIKAM Online Dictionary. National Institute of Informatics Journal.

[Blanc 1999] Etienne Blanc. 1999. PARAX-UNL: a Large Scale Hypertextual Multilingual Lexical Database. In Proceedings 5th Natural Language Processing Pacific Rim Symposium (NLPRS), pages 507-510, Beijing, China.

[Blaser et al. 1992] Brigitte Blaser, Ulrike Schwall, and Angelika Storrer. In Proceedings of COLING'92, pages 510-516.

[Carpenter and Penn 1999] Bob Carpenter and Gerald. Penn. 1999. ALE: The Attribute Logic Engine User's Guide Version 3.2 Beta, available on line at http://www.sfs.nphil.uni-tuebingen.de/~gpenn/ale.html.

[Chen 1976] P. Chen. 1990. The Entity-Relationship Model: Toward a Unified View of Data. ACM Transactions on Database Systems, pages 9-36.

[Fellbaum 1998] Christiane Fellbaum. 1998. WordNet: An Electronic Lexicon Database. MIT-Press, Cambridge Massachusetts.

[Mangeot 2000] Mathieu Mangeot-Lerebours. 2000. Papillon Lexical Database Project: Monolingual Dictionaries & Interlingual Links. In Proceedings of the 7th Workshop on Advanced Information Network and System (WAINS'7), pages 39-44, Kasetsart University, Bangkok, Thailand.

[Naruedomkul and Cercone 1997] Kanlaya Naruedomkul and Nick Cercone. 1997. Steps Toward Accurate Machine Translation. In Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation, pages 63-75, Santa Fe, New Mexico.

[Naruedomkul and Cercone 1999] Kanlaya Naruedomkul and Nick Cercone. 1999. The Role for Word Association Numbers in Machine Translation. In Proceedings of the Conference Pacific Association for Computational Linguistics (PACLING'99), pages 379-392, Waterloo, Ontario, Canada.

[Naruedomkul and Cercone 2000] Kanlaya Naruedomkul and Nick Cercone. 2000. Generate and Repari Machine Translation. In Proceedings of the Fourth Symposium on Natural Language Processing (SNLP'00), pages 63-79, Chiang Mai, Thailand.

[Pianta et al. 2002] Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing and aligned multilingual database. In Proceedings of the First International Conference on Global WordNet, Mysore, India.

[Serasset and Mangeot 2001] Gilles Sérasset and Mathieu Mangeot. 2001. Papillon Lexical Database Project: Monolingual Dictionaries & Interlingual Links. In Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS), pages 119-125, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan.

[Tiedemann 2002] Jorg Tiedemann. 2002. MatsLex - a Multilingual Lexical Database for Machine Translation. In Proceedings of the Third International Conference on Linguistic Resources and Evaluation (LREC 2002), pages 1902-1912, Las Palmas de Gran Canaria, Spain

[Vossen 1998] P. Vossen. 1998. EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kulwer Academic Publishers, Dordreht.

Appendix D

สถาบันนวัตกรรมและพัฒนากระบวนการเรียนรู้
Institute for Innovation and Development of Learning Process
สถานที่ติดต่อ : คณะวิทยาศาสตร์ มหาวิทยาลัยมหิดล ถนนพระรามหก กรุงเทพฯ 10400
Faculty of Science, Mahidol University, Rama 6 Rd, Bangkok 10400
Tel : 0-2201-5728-9 Fax : 0-2247-7054 E-mail : directil@mahidol.ac.th

23 มิถุนายน 2546

เรื่อง ขอเชิญเป็นวิทยากร

เรียน ดร.กัลยา นฤดมกุล

ตามที่ท่านเป็นผู้ที่มีความเชี่ยวชาญและมีประสบการณ์ทางงานวิจัยด้าน Machine Translation ทาง สถาบันนวัตกรรมและพัฒนากระบวนการเรียนรู้ จึงใคร่ขอเรียนเชิญท่านมาให้สัมมนาแก่นักศึกษาปริญญาเอก สาขาวิทยาศาสตร์และเทคโนโลยีศึกษา แขนงวิชาคอมพิวเตอร์ศึกษา ในเรื่อง Machine Translation ในวันอังคารที่ 24 มิถุนายน 2546 และวันอังคารที่ 1 กรกฎาคม 2546 เวลา 9.00 – 11.00 น. โดยที่ในครั้งที่ 1 นั้น อาจารย์จะเป็นผู้พูดให้นักศึกษาฟัง และในครั้งที่ 2 นักศึกษาจะเป็นผู้นำเนื้อหาที่เกี่ยวข้องมาอภิปรายและถกกัน โดยมีอาจารย์เป็นผู้ช่วยให้คำแนะนำและวิจารณ์

จึงเรียนมาเพื่อโปรดรับเป็นวิทยากร จักขอบคุณยิ่ง

ขอแสดงความนับถือ

ดร.ภิญโญ พานิชพันธ์
ผู้อำนวยการสถาบันนวัตกรรมและพัฒนากระบวนการเรียนรู้

Appendix E

From: virach@tcllab.org

Sent: 15 กันยายน 2547 13:01

To: ak@ku.ac.th; athavisak@yahoo.com; boon@cpe.eng.kmutt.ac.th; boonserm@cp.eng.chula.ac.th; cccpt@ mucc. mahidol.ac.th; kittipat.y@bu.ac. th; kosin.cha@kmutt.ac.th; krit.kosawat@nectec.or.th; Nisachon.T@Chula.ac.th; nuantip.tan@kmutt.ac.th; nuttanart@cpe.eng.kmutt.ac.th; pattarachai@it.kmitl.ac.th; prasarn@ce.kmitl.ac.th; ranat @th.ibm.com; rdk@parthenon.cs.tu.ac.th; scknr@mucc.mahidol.ac.th; surapan@siamguru.com; thepchai@nectec.or.th; wanasanan@eng.cmu.ac.th; wanchai.r@Chula.ac.th; Watit.B@Chula.ac.th; wirote.A@Chula.ac.th

Subject: invitation for MT Summit X local committee (resend)

Dear Colleagues,

I resend the following invitation in case that you did not go over it yet. Please let me know your convenience in accepting to be a local organizing committee.

Regards,

Virach

เรียน ท่านผู้ทรงคุณวุฒิ

ด้วย The Asia-Pacific Association for Machine Translation (AAMT) ร่วมกับ สถาบันเทคโนโลยีสารสนเทศและการสื่อสารแห่งชาติ ประเทศญี่ปุ่น (National Institute of Information and Communications Technology) โดยหน่วยวิจัยภาษาศาสตร์คำนวณ (Thai Computational Linguistics Laboratory), ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ และสถาบันเทคโนโลยีนานาชาติสิรินธร จัดการประชุมเชิงวิชาการนานาชาติ หัวข้อ MT SUMMIT X ในระหว่างวันที่ 12-16 กันยายน 2548 ณ จังหวัดภูเก็ต

การประชุมเชิงวิชาการนานาชาติ MT SUMMIT เป็นการประชุมเชิงวิชาการเกี่ยวกับการแปล ภาษาด้วยคอมพิวเตอร์ ซึ่งครอบคลุมถึงการวิจัยและพัฒนาองค์ประกอบต่างๆ ที่เกี่ยวข้องกับระบบ การแปล การประเมินผล การประยุกต์ใช้ ตลอดจนนโยบายที่เกี่ยวข้อง จัดขึ้นทุกๆ 2 ปี โดย ครั้งนี้เป็นครั้งที่ 10 และ เป็นครั้งแรกที่จะจัดขึ้นที่ประเทศไทย

คณะผู้จัดใคร่ขอเรียนเชิญท่านซึ่งเป็นผู้ที่มีความรู้ ความเชี่ยวชาญ และประสบการณ์เกี่ยวข้องกับ การวิจัยและพัฒนาด้านการแปลภาษาด้วยคอมพิวเตอร์ ร่วมเป็น Program Committee เพื่อร่วมพิจารณาคัดสรรบทความและผลงานที่ทรงคุณค่าเพื่อนำเสนอในการประชุมที่จะจัดขึ้น ตลอดจนสนับสนุนการประชาสัมพันธ์การจัดการประชุมวิชาการดังกล่าวให้แพร่หลายยิ่งขึ้นด้วย

จึงเรียนมาเพื่อพิจารณาตอบรับ และขอขอบคุณอย่างสูงที่ท่านสนับสนุนการจัดการประชุมวิชาการ MT SUMMIT X

ขอแสดงความนับถือ

วิรัช ศรเลิศล้ำวาณิช

หน่วยวิจัยภาษาศาสตร์คำนวณ

Chair of Local Organizing Committee

Appendix F

From: Ishizaki Shun [ishizaki@sfc.keio.ac.jp]
Sent: 23 กันยายน 2547 8:55
To: scknr@mucc.mahidol.ac.th
Subject: PACLING 2005 Conference Committee

Dear Prof. Kanlaya Naruedomkul

I would like to invite you to participate as a member of the conference committee of PACLING'05 which will be held in Tokyo, Japan, in August 24-27 in 2005. I am sure that your contribution from the overseas will be significant for the success of PACLING'05.

Let me say a few words about PACLING.
PACLING (Pacific Association for Computational LINGuistics) has grown out of the very successful Japan-Australia joint symposia on natural language processing held in November 1989 in Melbourne, Australia and in October in Iizuka, Japan in 1991.

The first three meetings of the retitled PACLING, a name designed to express the wider membership, took place in Vancouver, Canada in 1993, in Brisbane, Australia in 1995, in Tokyo, Japan in 1997, in?Waterloo, Canada in 1999, in Kita-kyushu, Japan in 2001 and in Halifax,?Canada in 2003.

PACLING'05 will be a low-profile, high-quality, workshop-oriented meeting whose aim is to promote friendly scientific relations among Pacific Rim countries, with emphasis on interdisciplinary scientific exchange demonstrating openness towards good research falling outside current dominant "schools of thought," and on technological transfer within the Pacific region. The conference represents a unique forum for scientific and technological exchange, being smaller than ACL, COLING, or Applied NLP, and also more regional with extensive representation from the Pacific.

Typically between 36-48 papers are presented at PACLING and there is time for discussion and taking in the sights. We will have a call for papers and a web site up and running soon.

We have about 20 Japanese members for the conference committee. We are expecting of your participation in the conference and waiting for your positive reply to this email.

Best Wishes

Shun ISHIZAKI
President, PACLING (Pacific Association for Computational LINGuistics), Prof. of Keio University, Japan