Thai Lexical Semantic Annotation by UW

Virach Sornlertlamvanich, Tanapong Potipiti and Thatsanee Charoenporn
National Electronics and Computer Technology Center,
National Science and Technology Development Agency,
Ministry of Science Technology and Environment,
22nd Floor Gypsum Metropolitan Tower 539/2 Sriayudhya Rd. Rajthevi Bangkok 10400 Thailand.
Email: {virach, tanapong, thatsanee}@nectec.or.th

Abstract

Universal Words (UW) – a set of interlingual acceptations – are the fundamental of the UNL (Universal Networking Language). UNL researchers in different countries annotate words in their languages with UWs. This paper presents a system for annotating Thai lexicons with UWs. The whole process consists of word extraction, word sense disambiguation, and UW annotation. This paper also proposes a computable method to find the appropriate UW to annotate a Thai word.

Introduction

The UNL project ([8]) has been proposed under the aegis of the United Nations University, Japan since 1996. The UNL project is a collaborative work of research institutions from 16 countries. UNL aims to be an international semantic annotation standard for network oriented multilingual communication. UNL represents natural language meaning through semantic graphs in which nodes are Universal Words (UW) – interlingual concepts –. One of the crucial tasks in introducing UNL is the procedure in defining the UWs and how partners in different countries annotate their words with UWs consistently. This paper presents the solution employed by the Thai UNL development team. Our approach will also be applicable to other languages and other interlingua-based machine translations as well.

1 UNL specification

The existing interlingua-based machine translation systems translate source languages to an interlingua and then translate the interlingua to the target language. The errors in creating the interlingua propagate to the target language generation. This drawback in the interlingual approach has impeded the progress in practical use. To improve the translation accuracy, UNL proposes a new paradigm in which the users directly prepare the interlingual documents called UNL as the source documents. So that the source language for the target language generation is the flawless interlingua. Supporting the UNL framework, the UNL documents are designed to contain no semantic ambiguities.

UNL is a project for multilingual networking communication initiated by the United Nations University, Japan. UNL bases on an interlingual approach represented by a hypergraph. A UNL graph consists of nodes and links. A node is formed by a UW attaching with a list of *attributes* (such as *@entry* indicating the entry node of the UNL graph; *@pl* indicating the plurality of the concept; *@def* indicating the definiteness of the concept). A link is a directed arc labeled by a semantic relation between the corresponding two nodes. A UNL document is a text encoding a set of UNL graphs. More details on UNL can be found in [1], [5] and [8]. Figure 1 and 2 show an example of a UNL graph and UNL text.

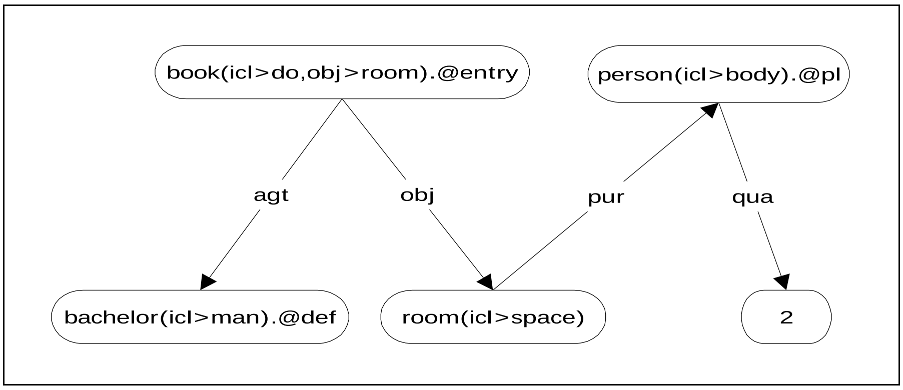

Figure 1: An example of UNL graph for "The bachelor books a room for two persons."

obj(book(icl>do, obj>room).@entry, room(icl>space))
agt(book(icl>do, obj>room).@entry, bachelor(icl>man).@def)
pur(room(icl>space), person(icl>volitional thing).@pl)
qua(person(icl>volitional thing).@pl, 2)

Figure 2: The UNL text encoding the UNL graph in Figure 1.

2 Universal words and the computational concept alignment

The primitive approach in defining a concept is based on dictionary approach such as EDR ([2]). A concept is described by a natural language as in a dictionary. Though this approach is obvious, it does not support computability unless the natural language understanding is supported. Wordnet ([4]) proposes a concept defining scheme in which a concept is defined by linking it to a set of synonyms (Synset) and 8 simple semantic relations to other concepts. However, with only the Synset and semantic relations the concepts cannot be identified accurately and yet contain some ambiguities. Thereafter, UNL ([8]) proposes UW; a new approach to solve the semantic ambiguity and support the computational concept alignment. A UW employs a concept headword and logical semantic constraints to disambiguate the word senses. Table 1 compares the *tired* concept representation in each scheme. Further discussion on concept alignment can be found in [6].

EDR	Wordnet 1.5	UW
- having or displaying a need for rest or an exhaustion - having lost interest - lack of imagination	- A1: tired (vs. rested) - A2: bromidic, commonplace, hackneyed, … - V1: tire, pall, grow weary, fatigue - V2: tire, wear upon, fag out, … - V3: run down, exhaust, sap, … - V4: bore, tire, …	- tired - tired(*agt>use*) - tired(*icl>do*) - tired(*aoj>volitional thing*) - tired(*gol>activity*) - tired(*icl>occur*) :

Table 1: Concept representation in EDR, Wordnet and UW

A UW denotes an interlingual acceptation used for concept representation in UNL. Theoretically, a UW has only one meaning. In other words, UWs do not allow semantic ambiguity. The reasons why English words are employed in UW construction are that (i) English is known by all UNL developers, and (ii) there are a lot of good bi-lingual dictionaries between a local language and English available.

The expression of UW is: "*<headword>(<list of restrictions>)*" e.g. *book(icl>do,obj> room)*. Restrictions are the composition of the following constraints:

1) *Icl* (stands for *inclusion*) is the restriction defining the semantic class where the UW is included. A part of UNL class hierarchy is shown in Figure 3. For example, “*car(icl> movable thing)*” indicates that this UW is in the class of *movable thing*.
2) Any semantic relations, available for the UNL arcs, with a UNL class name can be used in restricting the meaning of the English headword. For example, *eat*(*agt>volitional thing*, *obj>food*) indicates that the agent of this UW is restricted to be the UWs in the class of *volitional thing* and the object of this UW is restricted to be the UWs in the class of *food*.

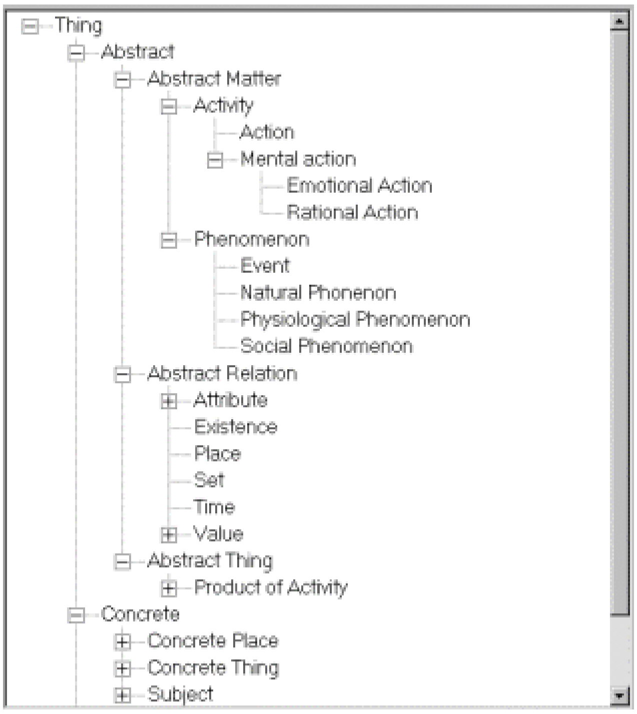

Figure 3: A part of UNL class hierarchy

3 UW annotation for Thai

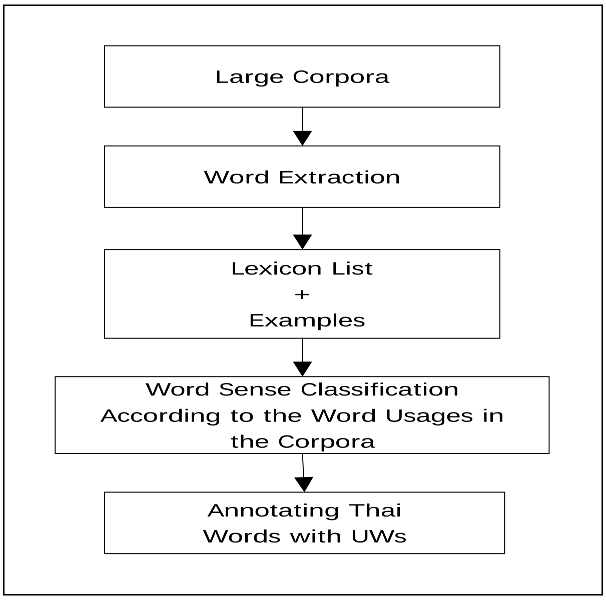
Large Corpora
Word Extraction
Lexicon List + Examples
Word Sense Classification According to the Word Usages in the Corpora
Annotating Thai Words with UWs

Figure 4: UW annotation for Thai: an overview

Mainly, there are five steps in UW annotation for Thai. The first step is corpus collection. The corpus is collected automatically from the Internet. After the corpus is collected, corpus-based word extraction ([7]) comes in the process. Based on the statistical features such as the mutual information and the entropy of character strings, a candidate word list together with their corresponding examples are listed. Each word and its examples in the corpus are then classified according to their senses. When all possible senses are classified, each sense will lead to the final UW annotation step.

3.1 Corpus-based word extraction

Unlike the western languages, Thai has no explicit word boundary. Thai word extraction is therefore an essential task for natural language processing in general. [7] has proposed a statistical Thai corpus-based word extraction with a high accuracy result. Their algorithm is applied to our task in preparing the candidate word list. Because this algorithm still produces some erroneous strings, the word list extracted by the algorithm is manually checked by lexicographers.

3.2 Word sense classification

After the word candidates are extracted from the corpus, each entry is semantically classified according to its context in the corpus. As a result, the sentences containing the considering word with a unique sense are piled up into a cluster. In this process, the unsupervised word sense disambiguation proposed by [9] has been applied to facilitate the lexicographer. For example, *เกาะ* in Thai exhibits two meanings: *to attach* and *an island*. This word is then classified in two clusters as follows.

<table>
<tr><td colspan="2">เกาะ (sense1: <i>to attach</i>)</td></tr>
<tr><td>… มัน <i>เกาะ</i> ตัวเองกับกิ่งไม้ …</td><td>(It <i>clings</i> itself on a tree)</td></tr>
<tr><td>… ผู้โดยสารไม่จำเป็นต้องยืน <i>เกาะ</i> ห่วงอีกต่อไปแล้ว …</td><td>(Passengers don’t have to <i>hold</i> peddles anymore.)</td></tr>
<tr><td colspan="2">เกาะ (sense2: <i>an island</i>)</td></tr>
<tr><td>...บ้านผมอยู่ที่ <i>เกาะ</i> สมุย…</td><td>(I live at the Samui <i>island</i>)</td></tr>
<tr><td>...ประเทศญี่ปุ่นประกอบด้วย <i>เกาะ</i> ใหญ่ 4 เกาะ…</td><td>(There are four big <i>islands</i> in Japan.)</td></tr>
</table>

Table 1: Word sense classification

3.3 Annotating Words with UWs

3.3.1 Headword and Thai-English Dictionary

The first and most obvious step to match the Thai words with UWs is to employ a Thai-English dictionary and UW headwords. By looking up the dictionary, we can find the English words that have similar meaning to the considering Thai word. The UWs employing these English words as headwords will be listed. The annotator selects an appropriate UW from the candidates by considering the Thai word context comparing with the restrictions of the UWs.

1) From the Thai-English dictionary:
เกาะ = island, isle, hold, attach, …

2) The UWs that occupy the headwords above are listed:
island(icl>concrete thing)
island(icl>place)
isle(fld>poem,icl>place)

attach(agt>volitional thing, icl>do, obj>thing)
hold(gol>organization, icl>do, obj>matter)
3) The result of UW annotation:
เกาะ (sense1) is annotated with UW *island(icl>place).*
เกาะ (sense2) is annotated with UW *attach(agt>volitional thing, icl>do, obj>thing).*

Figure 5: Using Bilingual Dictionary for UW annotation

3.3.2 Restriction Similarity

In case that there is no appropriate UW in the UW candidates listed by the headwords, the annotator may consider other UWs guided by the restriction similarity. However, the annotator may create a new UW by forming a new set of restrictions attaching to a headword when there are no any appropriate choices.

เกาะ (sense1: *to attach*)	
...มัน *เกาะ* ตัวเองกับกิ่งไม้ ...	(It *clings* itself on a tree)
... ผู้โดยสารไม่จำเป็นต้องยืน *เกาะ* ห่วงอีกต่อไปแล้ว ...	(Passengers don’t have to *hold* peddles anymore.)

Figure 6: Word sense classification

According to the examples shown in Figure 6, a lexicographer may restrict the finding concept with *(icl>do, agt>volitional thing, obj>concrete thing).* As a result, the UWs with similar structure of restrictions will be selected as candidates based on the score of restriction-similarity. The score is calculated according to the following scheme.

1. The initial score is set to be 0.
2. The score is unchanged for the exact matched restriction pair.
3. In case of a pair of restrictions under the same UNL semantic relation but attaching to different classes, the score is decreased by the distance between those 2 classes. The distance is measured by the number of branches between two classes in the UW class hierarchy.
4. For any unmatched restrictions, the similarity score is decreased by 10 points per each.

Following is an example of determining the restriction-similarity score of two UWs:

w1(agt>volitional thing, icl>thing)
w2(agt>volitional thing, icl> concrete thing, fld>science)

The initial restriction-similarity score is 0. The first restrictions of both UWs are exactly the same, therefore the similarity score is unchanged; 0. For the second restrictions: *icl>thing* and *icl>concrete thing*, the score depends on the distance between *thing* and *concrete thing* in the UW class hierarchy. From the UW class hierarchy in Figure 3, there are 2 branches between *thing* and *concrete thing*, the score is therefore decreased by 2. The last restriction of the second UW which cannot be matched, the score is further decreased by 10. In concluding, the score is resulted in 0-2-10 = -12.

Conclusion and further research

In this paper, we have presented a system for UW annotating for Thai. The system consists of extracting words from corpus, defining word senses and annotating words with UW. This system is also useful for multilingual or thesaurus construction. Our future works is to develop an automatic annotation. In many cases, the existing UWs cannot be matched with the considered word. The annotator has to create a new UW. The most difficulty of UW creation is to scheme its *inclusion* (the UW class it belongs to). We are considering an automatic *inclusion* suggestion scheme by using the vector-based similarity [3] between the considered word and UW classes. The proposed system is a workbench for supporting lexicographers in annotating the word sense with a UW. It is a result of making use of the computability of UW expression.

References

[1] Bouguslavsky, I., Frid, N. and Iomdin, L. (2000). Creating a Universal Networking Module within an Advanced NLP System. *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 83-89.

[2] EDR. *EDR Electronic Dictionary Specification Guide*, TR-041, Electronic Dictionary Research Laboratory

[3] Lochbaum, K. E. and Streeter, L. A. (1989). Comparing and Combining the Effectiveness of Latent Semantic Indexing and the Ordinary Vector Space Model for Information Retrieval. *Information Processing and Management*, 25(6), pp. 665-76.

[4] Miller, G. A., Bechwith, R., Fellbaum, C., Gross, D. and Miller, K. (1993). *Five Papers on Wordnet*, Princeton University, CSL, Report 43.

[5] Serrasset, G. and Boitet, C. (2000). On UNL as the Future “html of the linguistic content” & the Reuse of Existing NLP Components in UNL-related Applications with the Example of a UNL-French Deconverter. *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 768-771.

[6] Sornlertlamvanich, V. (1999). Alignment of Concepts and the Hierarchies. *Proceedings of the Third Meeting of Special Interest Group on AI Challenges*, Japanese Society for Artificial Intelligence, pp. 28-31.

[7] Sornlertlamvanich, V., Potipiti, T., and Charoenporn, T. (2000). Automatic Corpus Based Thai Word Extraction with the C4.5 Machine Learning Algorithm. *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 802-807.

[8] Uchida, H., Zhu, M. and Della Senta, T. (2000). *UNL: A Gift for a Millennium*. The United Nations University.

[9] Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189-196.