Thai Lexical Semantic Annotation by UW

Virach Sornlertlamvanich, Tanapong Potipiti
and Thatsanee Charoenporn

Information Research and Development Division

National Electronics and Computer Technology Center
(NECTEC), THAILAND

Overview

- Universal Networking Language (UNL) project
  - UNL specification
  - Universal Word (UW) and the problems in concept alignment
- UW annotation for Thai
  - Corpus-based word extraction
  - Word-sense classification
  - UW annotation
- Conclusion

UNL project

- Initiated by the United Nations University in 1996
- Collaboration of research institution from 16 countries
- International semantic annotation standard for multilingual communication
- Interlingua-based data archive

UNL and existing MT

- Existing interlingual MT

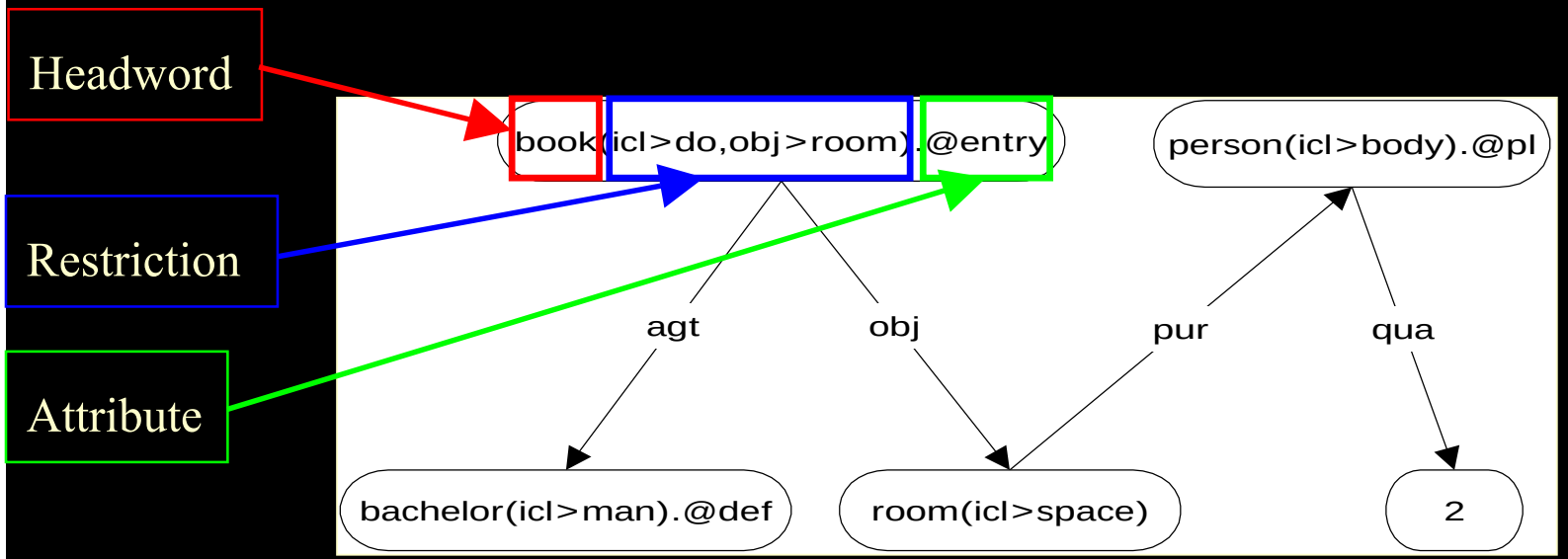
Errors in analysis are propagated into the generation process.

- UNL

No errors in analysis is propagated into the generation process.

UNL specification

- Interlingua in hypergraph representation
  - Node : UW (interlingual acceptation)
  - Link : UNL semantic relation such as agt, obj, pur ...



The UNL graph representing '*The bachelor books a room for 2 persons.*'

UWs and concept alignment (1)

- Concept alignment
  - The fundamental of interlingual approach
  - Define and alignment concepts among languages
    - Concept unification and decomposition
    - How to link a word sense in each language to the interlingual concepts consistently

UWs and concept alignment (2): approaches in concept alignment

- EDR
  - Approach : Word description as employed in dictionaries
  - Problem : Ambiguities and incomputability
- Wordnet
  - Approach : Synonym set and simple semantic relations to other words
  - Problem : Ambiguities
- UW
  - Approach : Headwords and semantic restrictions
  - Advantage : Computability and no ambiguity

UWs and concept alignment (3): approaches in concept alignment

EDR	Wordnet 1.5	UW
-having or displaying a need for rest -having lost of 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(*icl>physical*) -tired(*icl>mental*)

Representation of concept *tired* in different schemes

UW specification(1)

- UW format :

  <headword>(<list of restrictions>)

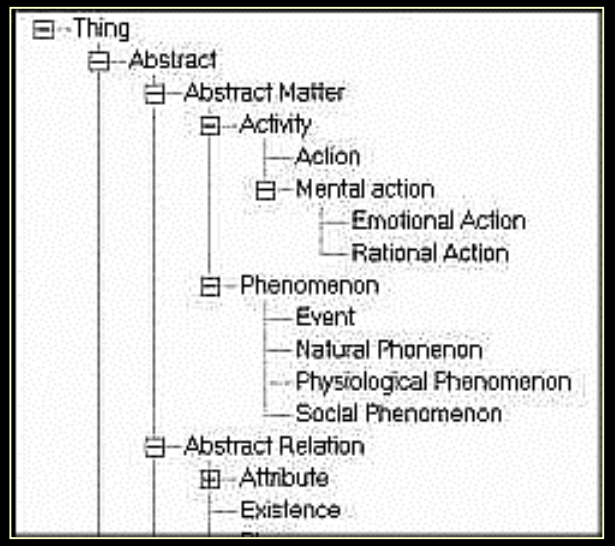
  e.g. *book(icl>do, obj>room)*

- Headword :

  An English word roughly describes the UW sense.

- Restrictions :
  - *Inclusion* (*icl*) to indicate the class of the sense

    e.g. *car(icl>movable thing)*



UW Class Hierarchy

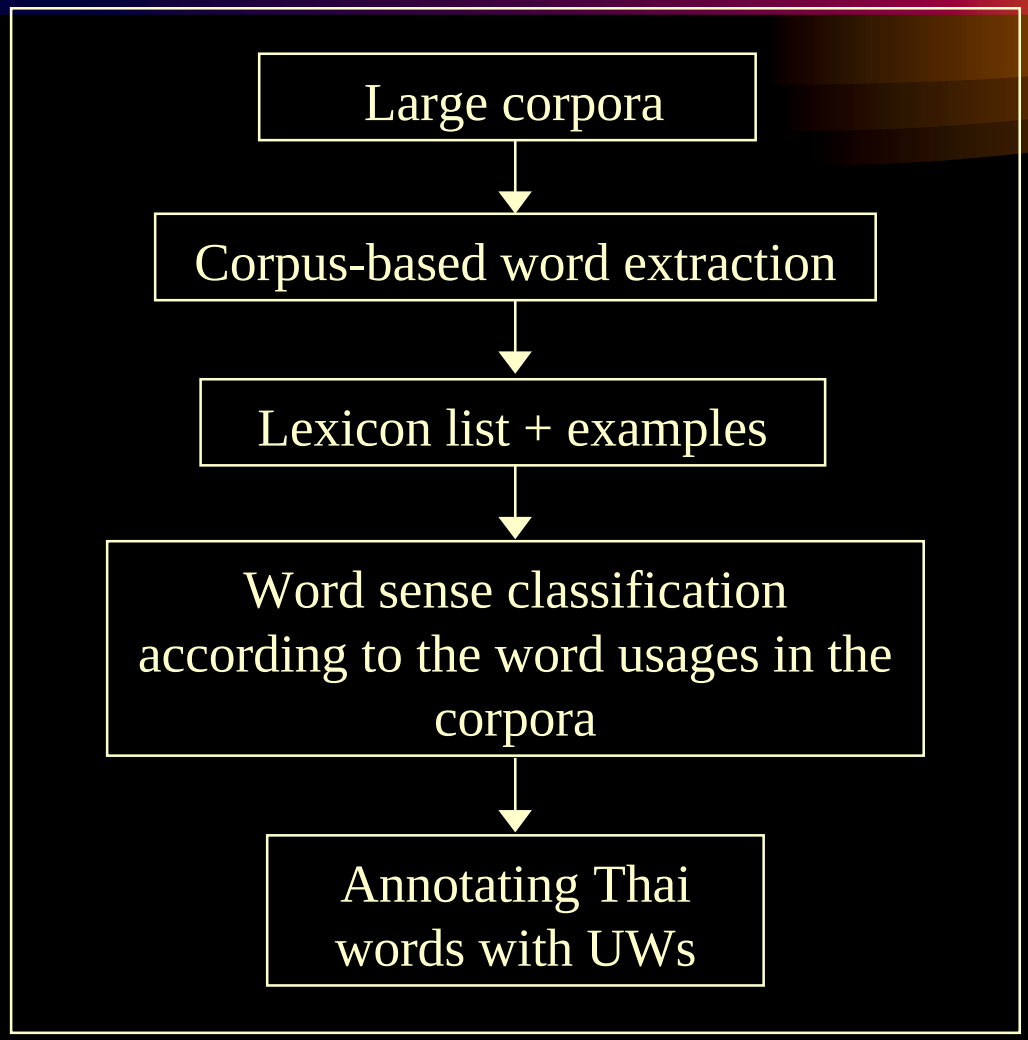
UW specification(2)

- Restrictions (continued)
  - *UNL semantic relations*

e.g. *eat(agt>volitional thing, obj>food)*

The agent of this UW is restricted to be *volitional thing*.
The object of this UW is restricted to be *food*.

UW annotation for Thai: an overview



Corpus-based word extraction(1)

- Corpus-based word extraction
  (Virach et. al. (COLING2000))
  - Machine learning employing statistical features of strings
  - Manual checking

Corpus-based word extraction(2): Mutual Information

$$Lm(xyz) = \frac{p(xyz)}{p(x)\,p(yz)} \qquad Rm(xyz) = \frac{p(xyz)}{p(xy)\,p(z)}$$

where

x is the leftmost character of string *xyz*
y is the middle substring of *xyz*
z is the rightmost character of string *xyz*
p( ) is the probability function.

High mutual information implies that *xyz* co-occurs more than expected by chance. If *xyz* is a word, its *Lm* and *Rm* must be high.
…*Efunction*… and ...*Function*...

Corpus-based word extraction(3): Entropy

$$Le(y) = -\sum_{\text{all } x \in A} p(xy \mid y).\log_2 p(xy \mid y)$$

$$Re(y) = -\sum_{\text{all } z \in A} p(yz \mid y).\log_2 p(yz \mid y)$$

where

x is the leftmost character of string *xyz*
y is the middle substring of *xyz*
z is the rightmost character of string *xyz*
p( ) is the probability function.

Entropy shows the variety of characters before and after a word.
If *y* is a word, its left and right entropy must be high.
Example: ...?*function*... , ...?*unction*...

Corpus-based word extraction(3): Other Features

- Frequency

  Words tend to be used more often than non-word string sequences.

- Length

  Short strings are likely to happen by chance.
  The long and short strings should be treated differently.

- Functional Words

  Functional words are used mostly in phrases. They are useful to disambiguate words and phrases.

Result of subjective test :

Word precision	85%
Word recall	56%

Word-sense classification

- Word and their contexts in the corpora
- Manual word-sense disambiguation according to the contexts.
- Unsupervised word sense disambiguation (Yarowsky 1995)

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

เกาะ (sense2: an island)	
...บ้านผมอยู่ที่ ***เกาะ*** สมุย…	(I live at the Samui ***island***.)
... ญี่ปุ่นประกอบด้วย ***เกาะ*** ใหญ่ 4 เกาะ…	(There are four big ***islands*** in Japan.)

Annotating Thai words with UW: headword and dictionary

- Headword search through the Thai-English dictionary

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)*
attach(agt>volitional thing, icl>do, obj>thing) *hold(gol>organization, icl>do)*

3) The best UWs annotation corresponding to the contexts in the corpora are:
เกาะ (sense1) is annotated with UW *attach(agt>volitional thing, icl>do, obj>thing).*
เกาะ (sense2) is annotated with UW *island(icl>place).*

Annotating Thai words with UW : restriction similarity (1)

- Restriction similarity
  - The annotator can find an appropriate UW by forming a set of restrictions, in case that there is no appropriate UW due to the headword search.

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

From the example above, a lexicographer may restrict the finding concept with *(icl>do, agt>volitional thing, obj>concrete thing).*

Annotating Thai words with UW : restriction similarity (2)

- UWs that have similar restrictions with the created set of restrictions will be listed as candidates.

- Similarity of restrictions will be ranked according to the ***similarity score***.

Annotating Thai Words with UW : restriction similarity score (1)

- ***Similarity score*** is computed as follows:
  The score is calculated according to the following scheme.
  - The initial score is set to be 0.
  - The score is unchanged for an exact matched restriction pair.
  - For a pair of restrictions under the same UNL relation but attaching to different classes, the score is decreased by the distance between those 2 classes.
  - For any unmatched restrictions, the score is decreased by 10 points per each.

Annotating Thai words with UW : restriction similarity Score (2)

- Example: *restriction similarity score* of

(agt>volitional thing, icl>thing)

and

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

	Score	Restrictions applied
	0 - 2 - 10	*agt>volitional thing , agt>volitional thing* *icl>thing, icl> concrete thing* *fld>science*
Total	- 12	

Conclusion and further research

- The process of UW annotation for Thai is presented.
- The computability of UW has been applied.

- Further Research
  - Automatic UW class suggestion applying vector similarity rather than linear similarity score between words in UW classes and the considered Thai word.