Asian WordNet Framework:
Its Web Service and Collaborative Platform

Virach Sornlertlamvanich
virach@gmail.com
SIIT, Thammasat University, Thailand

January 15-16, 2016
Outline

- **Asian WordNet Development**
  - Translation approach
  - Use of the existing bilingual dictionaries
  - Synset assignment
  - KUI for collaborative editing

- **WNMS (WordNet Management System)**
  - Distributed WordNet service
  - Service for cross language WordNet retrieval
Synset Assignment Algorithm (CS=4)

- Accept the Synset that includes more than one English Equivalent with confidence score 4.

Example:
- \( L_0: เป้าหมาย \)
- \( E_0: \text{aim} \)
- \( E_1: \text{target} \)
- \( S_0: \text{purpose, intent, intention, aim, design} \)
- \( S_1: \text{aim, object, objective, target} \)
- \( S_2: \text{aim} \)
Synset Assignment Algorithm (CS=3)

- Accept the Synset that includes more than one English Equivalent from the synonym of the target language with confidence score 3.

Example:

- L0: จ้อง
  L1: เพ่งมอง

- E0: stare
  E1: gaze

- S0: stare
  S1: gaze, stare
Synset Assignment Algorithm (CS=2)

- Accept the only Synset that includes the English Equivalent with confidence score 2.

Example:
- L0: สูติแพทย์
- E0: obstetrician
- S0: obstetrician, accoucheur

January 15-16, 2016
The Second Wordnet Bahasa Workshop, NTU, Singapore
Synset Assignment Algorithm (CS=1)

- Accept more than one Synset that includes each of the English Equivalent with confidence score 1.

Example:

- L0: ช่อง
- E0: hole
- E1: canal
- S0: hole, hollow
- S1: hole, trap, cakehole, maw, yap, gap
- S2: canal, duct, epithelial duct, channel
Asian WordNet Development Process

- Addition
- Discussion
- Lookup
- Correction
- Translation
- Voting

- WN
- X-English
- X-English
- X-English
- Thai-English

- merged-WN
- KUI

- AWN
- GWN

- ML Applications
  - Dictionary
  - Ontology
  - CL-Search
  - MT
  - Summarization
  - IE/IR
  - ....

January 15-16, 2016
WNMS for AWN

- **WNMS (WordNet Management System)**
  - Sense based translation rather word based translation
  - Show the relation between senses
  - System is fully distributed connected through a standard Open API
  - Collaborative editing tools based on KUI concept
WNMS (WordNet Management System)
Participation (Translate)

- Input a word to search
- Input a translated word, and select degree of confidence
- Input comment or memo if have
- Delete

January 15-16, 2016
The Second Wordnet Bahasa Workshop, NTU, Singapore
Participation (Vote)

☐ Read the comment or memo

☐ Vote
  vote up  vote down

January 15-16, 2016
The Second Wordnet Bahasa Workshop, NTU, Singapore
ExpertScore (score for voting)

- Each member is provided an initial value of voting score as 1
- The voting score is changed according to ExpertScore, which is estimated by the value of
  - Expertise
  - Contribution
  - Continuity

of the participation history of each member
ExpertScore

- **Expertise**: a composite score of the accuracy of opinion and vote

\[
\text{Expertise} = \alpha \frac{\text{count}(\text{BestOpinion})}{\text{count}(\text{Opinion})} + \beta \frac{\text{count}(\text{BestVote})}{\text{count}(\text{Vote})}
\] (1)

- **Contribution**: a composite score of the ratio of opinion and vote posting comparing to the total

\[
\text{Contribution} = \gamma \frac{\text{count}(\text{Opinion})}{\text{count}(\text{TotalOpinion})} + \rho \frac{\text{count}(\text{Vote})}{\text{count}(\text{TotalVote})}
\] (2)

- **Continuity**: a regressive function based on the assumption that the absence of participation of a member will gradually decrease its ExpertScore to one after a 365 days of the absence

\[
\text{Continuity} = 1 - \left( \frac{D}{365} \right)^4
\] (3)
ExpertScore

\[ \text{ExpertScore} = f(\text{Expertise}, \text{Contribution}, \text{Continuity}) \]

\[ \text{Expertise} = \alpha \frac{\text{count}(\text{BestOpinion})}{\text{count}(\text{Opinion})} + \beta \frac{\text{count}(\text{BestVote})}{\text{count}(\text{Vote})} \]

\[ \text{Contribution} = \gamma \frac{\text{count}(\text{Opinion})}{\text{count}(\text{TotalOpinion})} + \rho \frac{\text{count}(\text{Vote})}{\text{count}(\text{TotalVote})} \]

\[ \text{Continuity} = 1 - \left( \frac{D}{365} \right)^4 \]

where \[ \alpha + \beta + \gamma + \rho = 1 \]

\( D = \text{number of absent date} \)
ExpertScore

\[ \text{ExpertScore} = \left(1 - \left(\frac{D}{365}\right)^4\right) \times \left\{ \alpha \frac{\text{count(BestOpinion)}}{\text{count(Opinion)}} + \beta \frac{\text{count(BestVote)}}{\text{count(Vote)}} + \gamma \frac{\text{count(Opinion)}}{\text{count(TotalOpinion)}} + \rho \frac{\text{count(Vote)}}{\text{count(TotalVote)}} \right\} \]

- Range of the value of ExpertScore is 1 to 365 according to the accuracy and the rate of contribution of each member
  - Reliable members are rewarded better score for each vote
  - The expertise of the member is decreased by the discontinuity of the participation
Distributed WordNet Service

- Distribute the WordNet service node
- Service node can be locally maintained
- Synset ID (or Synset Offset) is the key to link between nodes
Representation of Synset Translation

<table>
<thead>
<tr>
<th>wn30_data_pos</th>
<th>15952</th>
</tr>
</thead>
<tbody>
<tr>
<td>sense_id</td>
<td>02958343</td>
</tr>
<tr>
<td>synset_offset</td>
<td>06</td>
</tr>
<tr>
<td>lex_filename</td>
<td>n</td>
</tr>
<tr>
<td>ss_type</td>
<td>5</td>
</tr>
<tr>
<td>w_cnt</td>
<td>car, auto, automobile, machine, motorcar</td>
</tr>
<tr>
<td>words</td>
<td>a motor vehicle with four wheels; usually propelled by an internal combustion engine; “he needs a car to get to work”</td>
</tr>
<tr>
<td>gloss</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>contribute</th>
<th>40725</th>
</tr>
</thead>
<tbody>
<tr>
<td>sense_id</td>
<td>40808</td>
</tr>
<tr>
<td>lex_filename</td>
<td>06</td>
</tr>
<tr>
<td>message</td>
<td>15952</td>
</tr>
<tr>
<td>contributed_by</td>
<td>06</td>
</tr>
<tr>
<td>contributed_date</td>
<td>215296</td>
</tr>
<tr>
<td>confidence_score</td>
<td>2007-06-12 09:50:42</td>
</tr>
<tr>
<td>vote_score</td>
<td>3</td>
</tr>
</tbody>
</table>

English WordNet

```
car, auto, automobile, machine, motorcar
```

Sense ID

```
15952
```

Thai WordNet

```
รถ, รถยนต์
```

January 15-16, 2016

The Second Wordnet Bahasa Workshop, NTU, Singapore
Types of Services ‘sense’

- Thai Sense (Get word translation by POS and SYNSET_OFFSET)
  
  **Service URI**: http://th.asianwordnet.org/services/sense/output/[callback]/pos/synset_offset
  
  **Service Name**: sense
  
  **Parameter**:
  
  pos = PartOfSpeech {n,v,r,s}
  synset_offset is an English Princeton WordNet v.3.0 offset, represented in 8 digits

http://th.asianwordnet.org/services/sense/xml/n/02958343

January 15-16, 2016  The Second Wordnet Bahasa Workshop, NTU, Singapore
Types of Services ‘dictionary’

- **E-Dictionary (Get word translation by word entry)**
  
  Service URI:
  http://th.asianwordnet.org/services/dictionary/output/[callback]/type_of_dict/search_word

  Service Name: dictionary
  Parameter: type_of_dict = {en2th, th2en}, search_word is a word you want to search
Types of Services

- **Auto complete** (Get a list of words existing in WordNet by prefix auto completion)
  
  Service URI:
  
  http://th.asianwordnet.org/services/autocomplete/output/[callback]/language/search_word

  Service Name: autocomplete

  Parameter: language = {en,th}, search_word is a word you want to get autocomplete (Result: limit 50 records found)

- **WN-Browser** (Browse WordNet and its semantic relations)
  
  Service URI:
  
  http://th.asianwordnet.org/services/browse/output/[callback]/language/search_word

  Service Name: browse

  Parameter: language = {en,th}, search_word is a word you want to get all relations
Visualization of AWN
(http://www.asianwordnet.org/)

Thai->Indonesian

January 15-16, 2016
The Second Wordnet Bahasa Workshop, NTU, Singapore
Guideline in WordNet Translation

- Word entry must be translated into the appropriate WORD(s) by avoiding phrase and meaning explanation.
- Words in a Synset must be interchangeable in a sentence.
Translational Issues

- There are many cases that a gloss need to be expressed in a phrase or explanation, especially in the case of technical terms and scientific vocabulary.

  Ex. **Chaperon**
  
<table>
<thead>
<tr>
<th>POS</th>
<th>Noun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synset</td>
<td>chaperon, chaperone</td>
</tr>
<tr>
<td>Gloss</td>
<td>one who accompanies and supervises a young woman or gatherings of young people</td>
</tr>
</tbody>
</table>

  Thai: ผู้ตามควบคุมหญิงสาว

- The concept is not common in the Thai language.
A gloss can be expressed by two or more Thai words. These words have the core meaning but occur in different context. Should it be divided into more specific concept?

Ex. **Appear**

- POS: Verb
- Synset: appear, come out
- Gloss: be issued or published; "Did your latest book appear yet?"; "The new Woody Allen film hasn’t come out yet"

**Thai**

- T1 = ตีพิมพ์; T2 = ออกฉาย
  - T1 occurs in the context of printed matter
  - T2 occurs in the context of film or movie
For missing Synset

- => Computational model for the compound Synset (Synset composition/decomposition)
- => Adding new Synset with language/domain specific Synset tag
WNMS 2.0 (Langgrid Plug-in)

Language Grid is an online multilingual service platform which enables easy registration and sharing of language services such as online dictionaries, bilingual corpora, and machine translations. The Language Grid protects Intellectual properties of language resources and can be used for non-profit and research purpose.

- 121 organizations from 18 countries have participated in the Language Grid (2010/08)
- The Language Grid is providing 91 language services (2010/08)
- Federated operation starts between Kyoto University and NECTEC in summer 2010
- Service Grid Open Source activity has started globally
- Language Grid is aiming at helping translate Wikipedia articles
- EU/NSF projects will use the source code of the Language Grid
- Research papers on Language Grid are accepted by ICSOC, SCC, ICWS, etc.

For more information about the Language Grid, please visit NICT Language Grid Project page.
Summary

- Translation approach for Asian language Wordnet development
- Algorithm for synset alignment is proposed
- ExpertScore for weighting the vote
- Langgrid plug-in