UNL Document Summarization

Virach Sornlertlamvanich, Tanapong Potipiti and Thatsanee Charoenporn

Information Research and Development Division
National Electronics and Computer Technology Center (NECTEC), THAILAND
{virach, tanapong, thatsanee}@nectec.or.th
Overview

- UNL project
- UNL specification
- UNL document summarization
  - Sentence score
  - N-best sentences
  - Removal of redundant words
  - Merging sentences
- Conclusion
UNL project

- Initiated by the United Nations University in 1996
- Collaboration of research institutions from 16 countries
- International semantic annotation standard for multilingual communication
- Interlingua-based data archiving
UNL and existing MT

- **Existing interlingual MT**

  Source language $\rightarrow$ analysis $\rightarrow$ Interlingua $\rightarrow$ generation $\rightarrow$ Target language

  Errors in analysis are propagated into the generation process.

- **UNL**

  User $\rightarrow$ preparing $\rightarrow$ UNL document $\rightarrow$ generation $\rightarrow$ Target language

  No errors in analysis is propagated into the generation process.
UNL specification

- Interlingual representation
  - Nodes: UWs (interlingual acceptations)
  - Links: UNL semantic relations such as agt, obj, pur ...

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

- **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)**, **field(fld)**
    
    e.g. `car(icl>movable thing)`
  - **Relations**

**UW Class Hierarchy**
Multilinguality
UNL is an interlingua for multilingual application.

Unambiguity
UNL is designed to contain no semantic ambiguity.

Semantic information
Employing UNL, semantic information are employed for high quality summarization.
Four steps in UNL summarization

1: Calculating sentence scores
2: Selecting n-best sentences
3: Removing redundant words
4: Merging sentences

Summary
1: Calculating sentence scores

- A sentence score is calculated as follows:

\[ S(s) = \sum_{\forall u_{w_i} \in S} W(u_{w_i}) \]

where:

- \( S \) = sentence scoring function
- \( s \) = considered sentence
- \( W \) = weighting function
- \( u_{w_i} \) = universal word
- \( Tf \) = term frequency
- \( Idf \) = inverted document frequency
2: Selecting n-best sentences

- Five sentences with the highest scores are selected from the original 100-sentences (2,000 words) text.

UNL represents the means to facilitate multilingual communication on the information network.

The language exists only on the information network.

UNL is a global-scale common language, being transparent to all languages.

Information encoded in UNL is converted to an equivalent counterpart written in the target language, through a language generator "deconvertor" prepared for each language.

Complying with the same technical standard, these computer networks comprise the Internet.
3: Removing redundant words-1

- Insignificant modifier words are removed. Modifier relations are man, mod, ben and such.

\[
Con(I(uw_1, uw_2)) = \frac{W(uw_1)}{W(uw_2)}
\]

Where, \( Con \) contribution function
\( I() \) considered UNL relation
\( W \) uw weighting function
\( uw_1 \) head uw
\( uw_2 \) dependent uw

The links are removed if the contribution score is less than a threshold (1.5 in the experiment).
3: Removing redundant words-2

Threshold of the contribution score is 1.5

“UNL represents the means to facilitate multilingual communication on the information network.”

$\text{Con}(\text{met}(\text{facilitate.@pred, means.@def})) = 4.27$

$\text{Con}(\text{mod}(\text{communication, network.@def})) = 1.81$
$\text{Con}(\text{mod}(\text{communication, multilingual.@indef})) = 1.78$
$\text{Con}(\text{mod}(\text{network.@def, information})) = 0.47$

removed
### Sentences

<table>
<thead>
<tr>
<th>Sentences</th>
<th>Removed words</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNL represents the means to facilitate multilingual communication on the information network.</td>
<td>information</td>
</tr>
<tr>
<td>The language exists only on the information network.</td>
<td>only, information</td>
</tr>
<tr>
<td>UNL is a global-scale common language, being transparent to all languages.</td>
<td>common, all</td>
</tr>
<tr>
<td>Information encoded in UNL is converted to an equivalent counterpart written in the target language, through a language generator &quot;deconvertor&quot; prepared for each language.</td>
<td>Through a language generator deconvertor prepared for each language</td>
</tr>
<tr>
<td>Complying to the same technical standard, these computer networks comprise the Internet.</td>
<td>same, computer</td>
</tr>
</tbody>
</table>
The UNL sentences sharing the same UW are possibly merged to produce a more complex sentence.
<table>
<thead>
<tr>
<th>The first sentence generated in English</th>
<th>The second sentence generated in English</th>
<th>The merged sentence generated in English</th>
</tr>
</thead>
<tbody>
<tr>
<td>The language exists on the network.</td>
<td>The UNL language is a global-scale language.</td>
<td>The UNL language is a global-scale language existing on the network.</td>
</tr>
</tbody>
</table>
### Plain text summarization

UNL represents the means to facilitate multilingual communication on the information network. The language exists only on the information network. UNL is a global-scale common language, being transparent to all languages. Information encoded in UNL is converted to an equivalent counterpart written in the target language, through a language generator "deconvertor" prepared for each language. Complying with the same technical standards, these computer networks comprise the Internet.

5 sentences, 67 words.

### UNL document summarization

UNL represents the means to facilitate multilingual communication on the network. UNL is a global-scale language, being transparent to languages, existing on the network. Information encoded in UNL is converted to counterpart written in the target language. These networks comprise the Internet, complying with the technical standard.

4 sentences, 47 words.
Conclusion

- The process of summarization by UNL has been presented.
- UNL provides many advantages in summarization.
- Our experiment shows that UNL can improve the quality of summarization.
- Applicable to any semantic representation.
- Further research Considering UW class hierarchy as well as the attributes and relations.