UNL Document Summarization

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

This paper proposes an approach on UNL document summarization. Our approach employs both the surface and semantic information of UNL annotation to summarize documents. With the merit of semantic annotation of the UNL, the essence of the document is efficiently collected which facilitates the abstraction function for language generation. The multilinguality can also be realized through the language decoverters from the summarized UNL document to the target languages under the UNL framework. The experiment result shows the improvement of the summarization quality in using the UNL annotation comparing with the original plain text.

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. The UNL framework provides a mean for representing the meaning of natural language document with a set semantic graphs. This paper introduces a summarization method to UNL document for a better summarization result. Rather than employing only the superficial information, we directly process the UNL semantic information to extract the essence of the document. Our work shows the improvement of the summarization quality in using the UNL annotation.

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, the UNL project 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 universal word (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], [4], [5] and [8]. Figure 1 and 2 show an example of a UNL graph and UNL text.

2 Universal words

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. ([5])

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*.

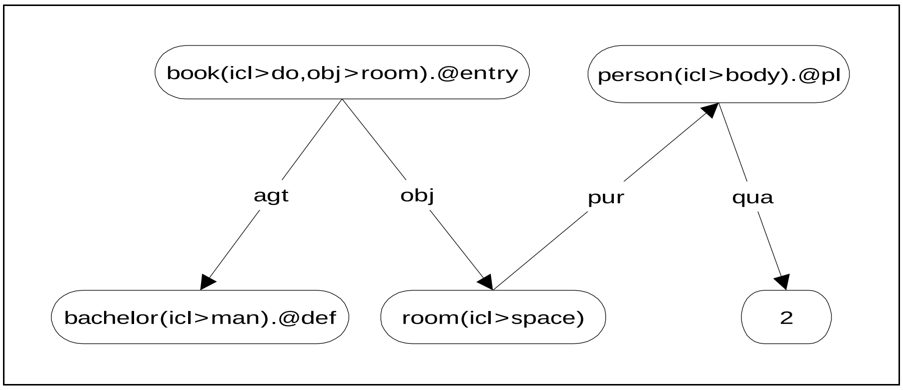
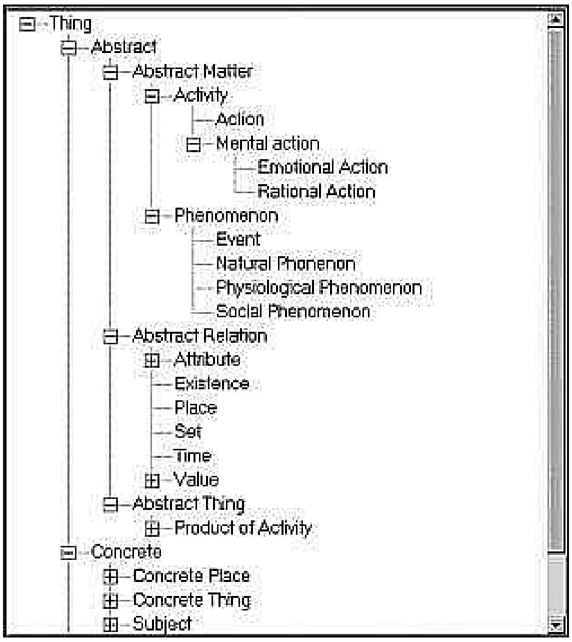


Figure 1: An example of UNL graph for "The bachelor books a room for two persons."

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.



- Thing
  - Abstract
    - Abstract Matter
      - Activity
        - Action
        - Mental action
          - Emotional Action
          - Rational Action
      - Phenomenon
        - Event
        - Natural Phonenon
        - Physiological Phenomenon
        - Social Phenomenon
    - Abstract Relation
      - Attribute
      - Existence
      - Place
      - Set
      - Time
      - Value
    - Abstract Thing
      - Product of Activity
  - Concrete
    - Concrete Place
    - Concrete Thing
    - Subject

Figure 3: A part of UNL class hierarchy

3 UNL annotation and text summarization

Most of existing works on text summarization such as [2], [3] and [7] rely on surface information of documents. Employing the surface information, these approaches select the best sentences and list them together to summarize the whole text. Without employing the semantic information, these approaches have a great drawback. The generated summaries are often not much readable and contain a lot of redundancies. However, for UNL documents, the UNL semantic information is very useful to summarize and generate high quality summaries.

3.1 Advantages of UNL document summarization

3.1.1 Multilinguality

Because UNL provides an interlingua expression framework, the UNL document summarization can be generated in many target languages without any additional work. The decoverter generates the desired target languages from the summarized UNL document.

3.1.2 Unambiguity

The UNL document does not allow semantic ambiguity in the annotation. Summarization of the UNL document ensures a high quality and clarity. For example, to summarize a document on *plants*, an ambiguity on whether *plant* means a factory or a tree may occur. But if the document is annotated by UNL in which different concepts are represented by different UWs, this ambiguity is clarified. The problem in multiple statistical count is consequently avoided.

3.1.3 Semantic information

Rather than employing only the superficial surface information, to summarize UNL documents we also employ the deep semantic information. This semantic information improves the quality of summarization. With this information, we can remove redundancy and combine sentences into a more meaningful and readable one.

No.	English	UNL Expression

1	UNL represents the means to facilitate multilingual communication on the information network.	aoj(represent.@entry.@pred.@present, UNL)
		obj(represent.@entry.@pred.@present, means.@def)
		met(facilitate.@pred, means.@def)
		obj(facilitate.@pred, communication)
		mod(communication, multilingual.@indef)
		mod(communication, network.@def)
		mod(network.@def, information)

2	The language exists only on the information network.	obj(exist.@entry.@pred.@present, language.@def)
		lpl(exist.@entry.@pred.@present, network.@def)
		mod(network.@def, only)
		mod(network.@def, information)

3	UNL is a global-scale common language, being transparent to all languages.	aoj(language.@entry.@pred.@present.@indef, UNL)
		aoj(global-scale, language.@entry.@pred.@present.@indef)
		aoj(common, language.@entry.@pred.@present.@indef)
		aoj(transparent.@pred,language.@entry.@pred.@present.@indef)
		ben(transparent.@pred, language:02.@pl)
		mod(language:02.@pl, all)

4	Information encoded in UNL is converted to an equivalent counterpart written in the target language, through a language generator "deconvertor" prepared for each language.	obj(encode.@pred, information)
		met(encode.@pred, UNL)
		obj(convert.@entry.@pred.@present, information)
		gol(convert.@entry.@pred.@present, counterpart.@indef)
		aoj(equivalent, counterpart.@indef)
		obj(write.@pred, counterpart.@indef)
		met(write.@pred, language:01.@def)
		mod(language:01.@def, target)
		met(convert.@entry.@pred.@present, generator.@indef)
		mod(generator.@indef, language:02)
		cnt(generator.@indef, deconvertor)
		obj(prepare.@pred, generator.@indef)
		ben(prepare.@pred, language:03)
		mod(language:03, each)

5	Complying with the same technical standards, these computer networks comprise the Internet.	aoj(technical, standard.@pl.@def)
		mod(standard.@pl.@def, same)
		gol(comply.@pred, standard.@pl.@def)
		mod(network.@pl, computer)
		mod(network.@pl, these)
		man(comprise.@pred.@present.@entry, comply.@pred)
		aoj(comprise.@pred.@present.@entry, network.@pl)
		obj(comprise.@pred.@present.@entry, Internet.@def)

Table 1: The 5 best sentences selected for summarization.

3.2 UNL document summarization

Mainly, there are 4 steps in UNL document summarization. The first step is to calculate a score for each UNL sentence. According to this score, the n-best sentences for summarization are selected. Employing the UNL semantic information, the redundant words or phases in the selecting sentences are removed. Then some selecting sentences are combined to improve readability and naturalness.

3.2.1 Calculating sentence-score

In order to select the best sentences for summarization, a score is calculated for each sentence. A sentence score is calculated by the weight of each word constituting the sentence. Weight of each word is computed according to its term frequency and inverted document frequency ([6]) as following.

$$S(s) = \sum_{\forall uw_i \in s} W(uw_i) \qquad \text{--- (1)}$$

$$W(uw_i) = Tf(uw_i) * Idf(uw_i) \quad \text{---(2)}$$

$$Idf(uw_i) = \log(N(uw_i) / n(uw_i)) \ \text{-(3)}$$

where

S is the sentence scoring function,
s is the considered sentence,
W is the weighting function,
uw_i is the universal word,
Tf is the term frequency,
Idf is the inverted document frequency,
$N(uw_i)$ is the number of the documents in the corpus,

$n(uw_i)$ is the number of documents in the corpus where *uw* occurs.

3.2.2 Selecting Sentences

Applying the scoring process to a UNL document, the sentences with the highest scores are selected.

A UNL with 100 sentences (including 2,000 words) from *Introduction to UNL* is selected for our experiment. Table 1 shows the 5 best sentences selected for summarization in both English and UNL semantic equivalents.

No.	The sentences generated by the original UNL	The sentences re-generated after removing redundant nodes	Removed words
1	UNL represents the means to facilitate multilingual communication on the information network.	UNL represents the means to facilitate multilingual communication on the network.	information
2	The language exists only on the information network.	The language exists on the network.	only, information
3	UNL is a global-scale common language, being transparent to all languages.	UNL is a global-scale language, being transparent to languages.	common, all
4	Information encoded in UNL is converted to an equivalent counterpart written in the target language, through a language generator "deconvertor" prepared for each language.	Information encoded in UNL is converted to counterpart written in the target language.	through a language generator "deconvertor" prepared for each language.
5	Complying with the same technical standards, these computer networks comprise the Internet.	These networks comprise the Internet, complying with the technical standard.	computer, same
6	(67 words)	(49 words)	(16 words)

Table 2: The sentences before and after removing redundant words.

3.2.3 Removing redundant words

The selected sentences still contain redundant words. Most of the redundant words are the modifiers. These modifiers are easily identified by considering the UNL semantic relations. The semantic relations such as *man*, *mod* and *ben* imply the modifying relationship. If an auxiliary node do not help in clarifying the head node, the auxiliary node can then be removed without distorting the total meaning. The contribution of an auxiliary node to a head node is measured by the following *contribution score*. Contribution score of an auxiliary word is defined as:

$$\mathrm{Con}\,(\mathrm{l}\;(\mathrm{uw}_1, \mathrm{uw}_2)) = \frac{\mathrm{W}\;(\mathrm{uw}_1)}{\mathrm{W}\;(\mathrm{uw}_2)} \quad \text{---(4)}$$

where

Con is the contribution function,

l() is the considered UNL relation,

W is the weighting function, defined in Equation (2).

The auxiliary node will be removed if the contribution score is less than the removing threshold. In our experiment the threshold is set to be 1.5. For example, the contribution score for 4 UNL relations from the first sentence in Table 1 are as follow.

Con(met(facilitate.@pred, means.@def)) = 4.27
Con(mod(communication, network.@def)) = 1.81
Con(mod(communication, multilingual.@indef)) = 1.78
Con(mod(network.@def, information)) = 0.47

Applying the threshold of 1.5, the node *information* is removed from this sentence. Table 2 shows the result when all redundant words are removed.

3.3 Combining sentences

To make the automatic summary more natural and readable, we propose the process of combining UNL sentences into one. Sentences that employ the same UW can be merged to reduce the sentential redundancy. To preserve naturalness and restrict the generation of unexpected very long sentences, only sentences with less than 15 words are merged. Figures 4-6 and Table 3 illustrate how to merge two sentences together.

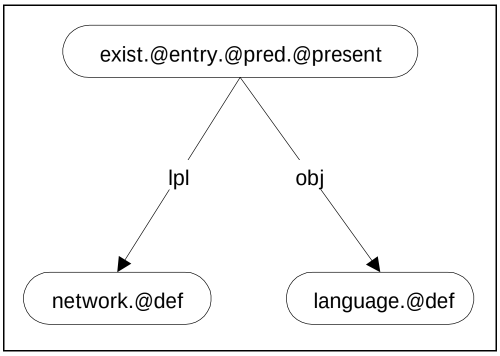


Figure 4: The UNL graph representing the sentence 2 in Table 2.

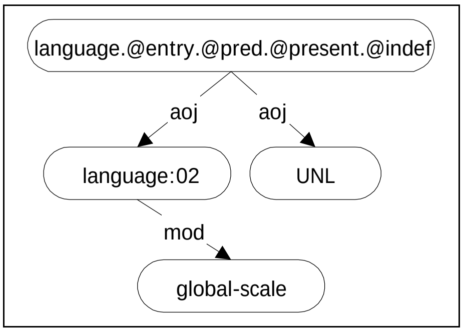


Figure 5: The UNL graph representing the sentence 3 in Table 2.

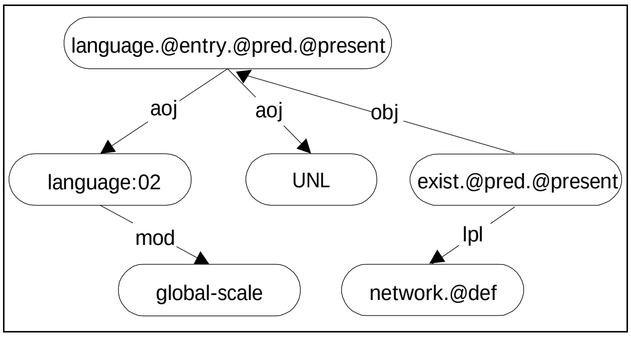


Figure 6: The UNL graph representing the merged sentence.

The first sentence generated in English	The second sentence generated in English	The merged sentence generated in English
The language exits on the network.	The UNL language is a global-scale language.	The UNL language is a global-scale language existing on the network.

Table 3: The sentences generated before and after combining.

Plain text summarization	UNL document 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.	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.
5 sentences, 67 words.	4 sentences, 47 words.

Table 4: The result of UNL document summarization comparing with plain text summarization.

3.3 Summarization Results

Table 4 shows the comparison of the result of the plain text summarization and UNL document summarization. It is significant that summarization of the UNL document can produce a more concise readable and higher quality summary.

Conclusion

In this paper, we propose a methodology for UNL document summarization, which includes (1) representative sentence selecting based on the *sentence scoring function*, (2) redundant words removal based on the *word contribution score* and (3) sentences merging based on the *co-relation words* finding. UNL provides a lot of advantages for summarization. Our experiment significantly shows the improvement of the summarization quality in using the UNL annotation comparing with the original plain text. The UNL semantic information can also be used to improve the naturalness in sentential level of the summarization.

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] Kupiec, J., Pederson, J. and Chen, F. (1995). A Trainable Document Summarizer. *Proceedings of 18th ACM-SIGIR Conference*, pp. 74–82.

[3] Myaeng, S. and Jang D. (1999). Development and Evaluation of Statistically-Based Document Summarization System. *Advances in Automatic Text Summarization*, MIT Press, pp. 61–70.

[4] 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.

[5] Sornlertlamvanich, V., Potipiti, T. and Charoenporn, T. (2000). Thai Lexical Semantic Annotation by UW. *Proceedings of WAINS7.*

[6] Sparck Jones, K. (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, 28(1), pp. 11–20.

[7] Sparck Jones, K. (1999). Automatic Summarizing Factors and Directions. *Advances in Automatic Text Summarization*, MIT Press, pp. 1–12.

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