Common Base Concepts of the 1st – the 3rd Order Entity WordNet of English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino- Burmese Languages for Internet Usage

Panornuag Sudasnna Na Ayudhya
Bansomdejchaopraya Rajabhat University

Received: 15 January 2019 Accepted: 10 Feb 2019

Abstract—Common base concepts of the 1st – the 3rd order entity WordNet of English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino- Burmese Languages for Internet Usage was aimed 1. To examine the translation equivalent pairs of common base concept in English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino languages and 2. To develop the translation equivalence of English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino- Burmese Languages of common base concepts for internet usage. The research methodology consists of the procedure of bi-directional translation to examine translation equivalent pairs of common base concepts of the 1st – the 3rd order entity words in English-Thai-Vietnamese-Lao-Cambodian- Bahasa-Filipino-Burmese languages, the procedure of expanded approach and focus groups conducted in related expertise to develop selected translation equivalent pairs of common base concepts of the 1st – the 3rd order entity WordNet for internet usage, and the procedure of users’ satisfaction towards the developed WordNet for internet usage. The results revealed that there are 1556 selected translation equivalent pairs of common base concepts of the 1st – the 3rd order entity of English-Thai-Vietnamese-Lao-Cambodian- Bahasa-Filipino- Burmese languages included in WordNet. The developed WordNet is the implementation of linguistic methodology as bi-directional translation to select the translation equivalent pairs in the development of multiple languages translation databases of major languages used in Southeast Asia Region. This method can be used to further development of multiple language translation databases in order to prove that the selected translation pairs are translation equivalent.

Index Terms—common base concepts, multiple languages, translation equivalent, WordNet

1. INTRODUCTION

Based on the Twelfth National Economic and Social Development Plan (2017-2021), in order to develop international standards and ability to integrate cultures into daily life of human capital, languages are the important instrument. Thus, the present investigation is aimed to develop a language innovation called “WordNet”. WordNet is initially developed by a group of researchers at Princeton University, the United States of America [1], [2].
by the Global WordNet Association as the core component of WordNet.

English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino-Burmese Languages are included in the development of translation equivalence WordNet. This is because in order to develop international standard and cultural integration in Thai human capital, understanding the languages of neighboring countries is an important tool and medium. In addition, the present work is the extension and further WordNet development based on the previous works of [7], [8], which were supported by National Research Council of Thailand. The present investigation implies the bidirectional translation method to examine English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino-Burmese translation equivalent pairs. Bidirectional translation method is a technique of translating source language to target language, and vice versa. Thus, bidirectional translation technique will be used to select the translation equivalent pairs.

2. THEORETICAL BACKGROUND

There are two issues of theoretical background as translation equivalence and WordNet.

i. Translation Equivalence

Translation equivalence is an important concept of theory of translation. The notion of equivalence in translation has been intensively discussed in the linguistic disciplinary. Finding translation equivalents is one of the important problems of the translation process and there is a lot of attempts to distinguish the concept of equivalence. As Catford states, “the central problem of translation-practice is that of finding target language equivalents. A central task of translation theory is that of defining the nature and conditions of translation equivalence. [9]”

Regarding to [10], translation is a form of communication and it is important to establish equivalence between the source text and the target text. According to the theory of translation, the term “equivalence” is used in its general meaning because it is difficult to find absolutely identical words and notions in different languages. This is because different languages have different phonetic, grammar, syntax and vocabulary structures.

The works of [10], [11], and [12] have been extensively discussed on the nature, types, and degrees of equivalence in translation and translation is viewed as a linguistic process where notions from one language are translated into another one. However, the work of [13] considered that it is an impossible to reach translation equivalence. Despite, there are different points of view towards the concept of equivalence, most of the specialists pay much attention to its meaning in the theory of translation.

In conclusion, the above theories reveal that the notion of equivalence is controversial and relative in nature. Most translation theorists and researchers pay attention to the importance of equivalence in translation, eventually their opinions on translation equivalence are different as the agreement that the equivalence is as a synonym of the translation process and the agreement that the equivalence should not cause the loss of main message of the text. Despite, there are different approaches, the equivalence has been concerned as an important notion in the translation process.

For the present research, the translation equivalence across languages is based on communicative translation type of Newmark, which will be examined using bidirectional translation method.

ii. WordNet

WordNet was initially developed in 1986 at Princeton University as a manually constructed electronic lexical database [14]
in order to test the underlying theories on a large scale.

WordNet consists of four separate components as nouns, verbs, adjectives, and adverbs. Each component contains synsets with words. Synsets are unordered sets of synonymous words and phrases [15]. Each member of a synset represents the same meaning; however, all synset members are not interchangeable in all contexts. All synsets include a brief definition, and one or more sentences illustrating the synonyms’ usage. All word members in a synset illustrate the synonymy and a word member in multiple synonyms reflects that word’s polysemy [16].

In conclusion, English WordNet developed at Princeton University is considered as a standard for the lexical-semantic representation of the natural language for English and a reference of new WordNets in other languages are being built.

3. Research Objectives

1. To examine the translation equivalent pairs of common base concept in English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino languages.

4. Research Methodology

i. Examining Translation Equivalence using Bidirectional Translation Method

The research procedure of examining translation equivalent pairs of common base concept in English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino languages is as follow.

1. Selecting English source words or noun phrases of first, second, and third order entities from synsets, which are included in WordNet Version 3.1 of Princeton University [17].

2. Applying bidirectional translation method to examine translation equivalent pairs of the selected first order entities, the selected second order entities, and the selected third order entities in English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino languages as the procedure below.

   a) Translating selected English first, second, and third order entities (Source text) to Thai, Vietnamese, Lao, Cambodian, Bahasa, and Filipino languages, orderly (Target text).

   b) Examining the translation words in the target text to the gloss of each synset as the horizontal evaluation and to the structure of each synset as the vertical evaluation. The translation words in the target text, which are not matched will be cut off.

   c) Examining the selected translation words in target language with bidirectional translation method. The three native speakers of target language will be asked to translate the selected translation words in target language back to the words in source language and the three native speakers of source language will be asked to translate the words in source language back to the translation words in target language. The translation words in target language, which can be translated back to source language and vice versa will be included in the developed WordNet.

ii. Development of the WordNet

The first procedure has been focused on the organization of words included in WordNet and has followed by the software which was applied to the development of WordNet.

The translation equivalent pairs, which are nouns are hierarchically organized starting from several nodes, from which the
rest of nodes are directly or indirectly linked through the semantic relation of hyponymy.

Whereas, the translation equivalent pairs, which are verbs are arranged hierarchically through the relation of troponymy.

The software which was applied to the development of WordNet consists of developing database and interface.

For the development of database, a freeware database, the Mini SQL de Hughes Technologies10 (Mini SQL) was used. The Mini SQL de Hughes Technologies10 offers Application Program Interfaces (APIs). The procedure includes the complete loading of WordNet 3.1, which consists of: a) the meanings of every word with its connection to the synset and b) the gloss related to the synset. For the recent study, the multilingual connection of the meanings of the words is conducted through the synsets.

In order to construct the WordNet, the interface has been developed to work through the Web.

5. RESULTS

The results are presented as the results of translation equivalent pairs of the first order entities, the second order entities, and the third order entities in English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino languages and the results of developed WordNet.

i. The Translation Equivalent Pairs of the First Order Entities, the Second Order Entities, and the Third Order Entities in English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino languages

There are 801 of the selected translation equivalent pairs of the first order entities, 504 of the selected translation equivalent pairs of the second order entities, and 251 of the selected translation equivalent pairs of the third order entities in English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino languages included in the developed WordNet.


The developed translation equivalence WordNet of English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino-Burmese Languages of common base concepts for internet usage includes the following interface.

a) Searching Multilingual WordNet

The interface for searching the multilingual WordNet is presented in Fig. 1.

Fig. 1 Searching Interface

According to Fig. 1, users can select a particular language as the source text using the icon (1) and select a particular language as the target text using the icon (2). The results will be shown in (3).

b) Editing Multilingual WordNet.

The interface for editing the multilingual WordNet is presented in Fig. 2 and Fig. 3.
As in Fig. 2, an editor must put a username and a password in order to access the editing function of WordNet.

WordNet can be edited using the editing interface, which allows to add more languages and more word information. With this function, the developed WordNet can be further included the other languages rather than the English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino-Burmese (as in Fig. 3).

c) Consulting Language Resources.

The interface for consulting language resources is presented in Fig. 4. Users can use this function to download free fonts and to access free dictionaries of each language of the developed WordNet.

d) Editing Statistics.

The interface for editing statistics is presented in Fig. 5. This function allows the administrator of program to collect the number of users and usage.

6. CONCLUSIONS AND DISCUSSIONS

The results revealed that the translation equivalent pairs of the first order entities obtained from bidirectional translation method were at the highest number comparing to the translation equivalent pairs of the second order entities and the third order entities. Whereas, the translation equivalent pairs of the third order entities obtained from bidirectional translation method were at the lowest number.

These results imply that the first order entities, the second order entities, and the
third order entities provide different levels of difficulty for examining translation equivalent words in cross languages. According to [4], the meanings of first order entities refer to physical objects. The meanings of second order entities refer to “events, processes, states-of-affairs, etc, which are located in time and which, in English, are said to occur or take place, rather than to exist”. The meanings of third order entities refer to “such abstract entities as propositions, which are outside space and time”[4]. The distinction among the three types of common base concepts illustrates that the main difference is based on the concreteness or abstractness of meanings. In the other words, the meanings of the first order entities are concrete whereas the meanings of the second and the third order entities are more abstract.

The distinction between “abstract” and “concrete” concepts and words has been agreed upon in various investigations. Regarding [18], the “abstract–concrete dimension” was differently reflected from language users’ understanding and comprehension of word meanings.

In addition, the works of [19] and [20] studied people’s judgement towards the concreteness and abstractness of words; it was shown that this linguistic and cultural influence is particularly strong for abstract, as compared to concrete words [21].

The influence of linguistics and culture on the concrete and abstract words has also been revealed by the works of [22]-[25]. According to [22], the study of translation equivalent pairs of kinship terms in Thai, Chinese, and English illustrated that there are incongruent family systems existing in Thai, Chinese, and English cultures, leading to the difficulty to find the translation equivalent pairs of kinship terms between these languages. The findings are supported by the work of [23]. That study revealed that there are different linguistic and cultural gaps among the translations of Thai, English and Chinese languages. However, it is these linguistic and cultural gaps that provide the path to the possibility of translation equivalence among Thai, English, and Chinese languages. The results suggest that translating concrete words between Thai and Chinese has a greater level of equivalence than translating concrete words between Thai and English, and between Chinese and English. In addition, the influence of linguistic and cultural gaps was more strongly observed in the translation of abstract words as compared to the translation of concrete words. The influence of linguistic and cultural gaps on words with concrete and abstract meanings is also found in the later works of [24] and [25].

The influence of linguistic and cultural gaps on the translation of concrete and abstract words between languages can be summarized by saying that, if linguistic and cultural gaps are less pronounced, the translation of concrete and abstract words between languages is more equivalent rather than if linguistic and cultural gaps are greater. As a result, there are more translation equivalent pairs obtained from two languages with less linguistic and cultural gaps than those obtained from two languages with more linguistic and cultural gaps. In addition, the linguistic and cultural gaps provide stronger influence on words with abstract meanings than on words with concrete meanings.

This explanation is supported by the numbers of translation equivalent pairs of the first order entities, the second order entities, and the third order entities in English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino languages found using bidirectional translation method. That is, the numbers of translation equivalent pairs of the first order entities were higher than the numbers of translation equivalent pairs of the second order entities and the third order entities.

Regarding the development of the translation equivalence WordNet of English-Thai-Vietnamese-Lao-Cambodian-Bahasa-Filipino-Burmese Languages of common base concepts for internet usage in the present investigation, the core structure of the present developed WordNet can be summarized as follows (as in Fig.6 and Fig. 7).

1. The examined English words in the Princeton WordNet can have multiple
translation equivalence words in the target languages.

2. One translation equivalent word in the target language can be translated to multiple English words in the Princeton WordNet.

3. This is possibility that there is more than one link established from English words to synsets.

4. The developed WordNet has been extended from the original WordNet 3.1 Synsets. The translation equivalent words of target languages are extended to the remaining WordNet 3.1 Synsets. The design of this developed WordNet also allows for further extension to other target languages. The design of this developed WordNet focuses on extension of the data sets. This is represented by Fig. 7.

As illustrated in Fig. 8, it is common for links from the examined English to connect to more than one translation equivalence words in the target languages, and for the links from translation equivalent word in the target language to link to multimple English words. In addition, there is more than one link established from English words to synsets.

Based on the recent language investigations [26]-[28], multilingual language data sets and processing provide obvious benefits, and this type of functionality is necessary in various fields of language communication and usage. For instance, the study of [29], revealed that multilingual language data sets and processing are useful because the information content found in different languages has recently been found to be complementary. However, most multilingual analysis tools have been applied to only small sets of languages. This is because the development effort for each language is large.

There is an existing need to develop multilingual language data sets and WordNet for many languages. This section will focus on the perspective of how to achieve the future development of multilingual data sets and WordNet.

Recent works [30]-[32] have focused on developing resources and tools for the development of data sets and WordNet of multiple languages. These investigations propose implementation which relies on reusing the resources from preexisting data sets of a language, and adapting them to new languages. In addition, these works also illustrate the important core techniques of reusing the preexisting language data sets to be simplicity of rules and the lexicon [32].
and [33], and uniform input and output structures [32] and [34].

It is assumed that simple applications can more easily be achieved with simple developmental technique. It is obvious that in order to develop language data sets and processing, the tools and developmental methodology should be simple. This principle was used in the present investigation for linking translation equivalent words collected from the present study to English WordNet synsets. However, the present developed WordNet produced in this way is neither error free nor complete, and a Web interface is implemented to facilitate manual interaction.

The results show that the multilingual WordNet can be quickly constructed using a method of linking translation equivalent words collected from the present study to English WordNet synsets, and that it can be made functional using a process of manual refinement of the acquired data.

The construction of the core of a multilingual Lexical knowledge base from preexisting lexical resources is proposed as the complementary techniques for linking the preexisting lexical resource as English Princeton WordNet and the present developed cross-language WordNet. However, due to the different aspects of semantic and lexical characteristics of each language included in the present developed WordNet, automatic construction should be focused on the semantic and lexical differences among the languages included in WordNet.

REFERENCES


------

Remaining WordNet → Translation Equivalent

Design of Extension of Translation

The present developed WordNet, automatic construction should be focused on the...


