Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education

Assisting Tools for Selecting Proper Semantic Meaning by Disambiguation of the Interference of the First Language

Nattapol KRITSUTHIKUL[a,c,d,*], Shinobu HASEGAWA[b], Cholwich NATTEE[a], & Thepchai SUPNITHI[d]

[a]*Information, Computer, and Communication Technology (ICT), SIIT, TU, Thailand*
[b]*Center for Graduate Education Initiative (CGEI), JAIST, Japan*
[c]*School of Information Science (IS), JAIST, Japan*
[d]*National Electronic and Computer Technology Center (NECTEC), Thailand*
*nattapol.kritsuthikul@gmail.com

Abstract: In this work, we proposed an assisting tool to EFL Thai students to find the incorrect concept word usage in writing. To prevent selecting incorrect term effect by translation from native language, a list of commonly confusing words along with a method to score words in co-occurrence are exploited. The proposed method can indicate the miss-using word in terms of semantic and suggest the possibly correct one with detail and reason.

Keywords: English as Foreign Language (EFL), Writing Skill, Natural Language Processing (NLP), Semantic, Meaning, n-gram, search engine

1. Introduction

Most of EFL students are influenced of translating their native language to English language when they try to communicate in English. However, semantic of words in languages is apparently not equal in terms of scope, sense, usage, etc. (Speaks, 2014), and translation often applies incorrectly due to the lack of a clear understanding of the word meanings. While selecting the word in translation, students often select the word with a board meaning or the frequently seen word confused by its polysemy as they assume the word has a sense of explanation in the same form.

The issues of the translated words from native language to English are greatly noticeable in the work of writing. EFL students use incorrect word to express their content because they do not know the word that represents their concept. The incorrect issues can be categorized into four types: (1) using word with boarder meaning (hypernym), (2) using word with excessive specific meaning (hyponym), (3) using frequently seen word with similar but incorrect meaning or usage (disjoint similar concept) and (4) using direct word-to-word translation in the proverb or grammatical pattern (ignoring correlated concept). These issues are originated with the interference of their native language because of translation and the original language not containing the concepts.

From observation, case (1), using the word with boarder or too general sense, has been found the most. Unfortunately, common students clearly know the concept that they want to mention, but they do not know the equivalent translated word in English thus they select the words with general concept which are in their knowledge. Moreover, the other cases can be happened in their writing from time to time based on students' limitation of English knowledge. For example, the concept "sandal" is wanted to be expressed. For case (1), students who do not know the word may select the word "shoe" because the word is the only concept they have for footwear. Some students may use the word "slipper" as case (3) because they do not know the difference among those concepts, which are the indoor and outdoor usage purpose, since the

detail may be missing from a bilingual dictionary. Furthermore, there is a case that the concept "sandal" is expressed as a compound word in their native language, such as Thai for 'รองเท้า (/rongtao/ - shoe - [noun]) - แตะ'(/tae/ - sound made by flip-flop shoe - [noun but it is polysemy to common word for verb as 'to touch']), hence they may invent new words, for instance "*touch shoe" or "*tae shoe", to represent the concept because they lack the English word for the concept or they believe there is no such concept in English.

These issues become more commonly found in EFL writing works because for them, these concepts are very confusing in terms of conceptual ambiguity from the divergence of their native language. Moreover, the translation method in expressing English for EFL students cannot be prevented since English language is greatly different from their native, and it interrupts and limits the way of their recognition. Hence, reducing the interference of the native language (L1 or the first language, henceforth) is the key to improve their English expressing. There are the words that often found as confusing words in translation provided by published dictionaries and guideline for translation by veteran translators. This information is a good hint to assist on disambiguation for those words.

In this work, we aim to develop an assisting tool to help EFL students in Thailand on selecting a proper English word. Since the case of using the hypernym is the most common and foremost issue, we aim to handle this issue first. By comparing the English written output from students with widely-published data, the likeness of pattern and wording in the written work is matched. This will result as if consulting student individual writing sentence with the large corpus to check the co-occurrence of the given words to evaluate the selection of accompanying words. By employing a list of confusable words in translation from Thai to English and WordNet, the tool is designed to reduce the interference of Thai native language on selecting the English words with commonly found cases, using hypernym and hyponym. We expect that this tool will improve their English writing skill and help them to gain more understanding of semantic concepts of English words.

2. System Architecture and Prototype System

The system is designed to find the inappropriate English words in terms of semantic meaning in the writing work of EFL student in Thailand, and to suggest a list of the better words for manual selection. The system architecture is illustrated in Figure 1.

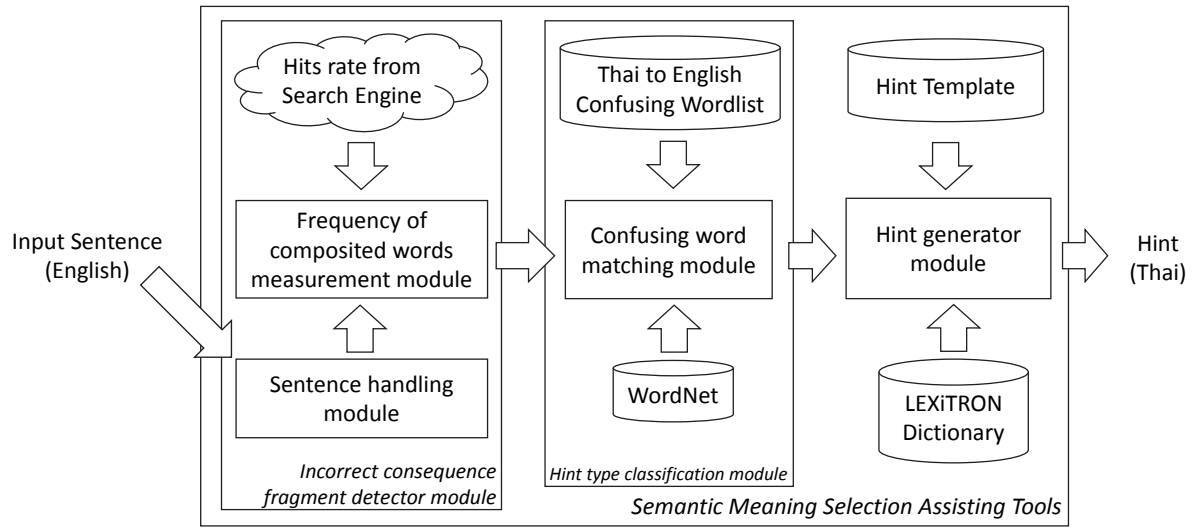

Figure 1. System architecture of the proposed system

2.1 Sentence handling module

This module gets an input as a digital English sentence made by EFL Thai students. To find an incorrectly semantically used English word among sentence, words are chunked with the *n*-gram model. The system divides the sentence into a group of consequent fragments.

For example, an input sentence from a student is "I write a picture". It will be assigned in *n*-gram model as given in Figure 2.

***n*-gram**	**consequent fragments**			
	1st-word	**2nd-word**	**3rd-word**	**4th-word**
4	I	write	a	picture
3 #1	I	write	a	
3 #2		write	a	picture
2 #1	I	write		
2 #2		write	a	
2 #3			a	picture

Figure 2. An *n*-gram model of the sentence "I write a picture"

2.2 Frequency of composited words measurement module

To acknowledge with the words with an inappropriate semantic meaning within the sentence, Hits rate (Number of search result) from search engine (e.g. Bing (Microsoft Bing, 2009), Google (Google, 1998), and so on) is employed as a concordance for measurement. Since we believe that the less frequency the words are used in co-occurrence, the more chance they are incorrectly composed with the wrong semantic meaning together. The rate is assigned to every roll. As a result from Figure 2, the hits rates of their responding consequent fragments are exemplified in Figure 3.

***n*-gram**	**consequent fragments**				**Hits rate from Bing**	**Hits rate from Google**
	1st-word	**2nd-word**	**3rd-word**	**4th-word**		
4	I	write	a	picture	**7**	**43,400**
3 #1	I	write	a		8,420,000	190,000,000
3 #2		write	a	picture	5,850,000	**295,000**
2 #1	I	write			18,400,000	10,900,000
2 #2		write	a		57,700,000	278,000,000
2 #3			a	picture	53,400,000	48,400,000

Figure 3. Hits rates of each gram by using Bing and Google

From Figure 3, according to extremely low hits rate from both Bing and Google for *n*-gram where $n = 4$, it is assumed that *n*-gram where $n = 4$ is incorrect. Moreover, *n*-gram where $n = 3$ starting from 'write' (3 #2 roll) also gets a remarkably low once comparing to other hit rates.

2.3 Confusing word matching module

Once the consequent fragments are found with low Hits rate, each fragment is examined through the Confusing Wordlist provided from published dictionaries and a guideline from veteran translators. Within the Confusing Wordlist, words ambiguous to each other are given together in a list format with their POS. An example of Confusing Wordlist is shown in Figure 4.

1.	PEOPLE@N	CITIZEN@N	POPULATION@N	NATIVE@N	INHABITANT@N
2.	ABOVE@PREP	OVER @PREP	HIGHER@ADJ		
3.	WRITE@V	DRAW@V		PAINT@V	
4.	PREVENT@V	PROTECT@V			
5.	ADVICE@V	INTRODUCE@V	SUGGEST@V	GUIDE@V	
6.	SANDAL@N ESPADRILLE@N	SLIPPERS@N MULE@N	FLIPFLOP@N	FLIP-FLOP@N	THONGS@N
7.	TALK@V	SPEAK@V	SAY@V	TELL@V	CONVERSE@V

Figure 4. An example of Confusing Wordlist

With words given in Confusing Wordlist, an example from Figure 3 is found with the word in given in line#3 from Figure 4 therefore the system attempts to replace the found word with the given alternative words in the list and re-do the Frequency of composited words measurement module With replacing word ‘write’ with word ‘draw’, we gain the result demonstrated in Figure 5.

***n*-gram**	**consequent fragments**				**Hits rate from Bing**	**Hits rate from Google**
	1st-word	**2nd-word**	**3rd-word**	**4th-word**		
4	I	draw	a	picture	**156,000**	**2,080,000**
3 #1	I	draw	a		1,110,000	34,100,000
3 #2		draw	a	picture	5,190,000	**39,900,000**
2 #1	I	draw			4,870,000	3,190,000
2 #2		draw	a		32,200,000	8,900,000
2 #3			a	picture	53,400,000	48,400,000

Figure 5. Hits rates of each gram after replacing the confusing word ‘write’ with word ‘draw’

From comparing Figure 3 and 5, we found that the Hits rate of roll 3 #2 from Google is boosted about 135 times while the whole sentence “I draw a picture” obtains much higher hits rate ratio than the sentence “I write a picture”.

2.4 Hint generation module

To give a suggestion with reason, WordNet (Princeton University, 2010) is exploited to this work to show a relation between written word and correct word. There are three cases.

- If the written word is a **hypernym** by WordNet of the word returning better Hits rate, the template which mentions the word in use is “*too general term*” will be shown.
- If the written word is a **hyponym** by WordNet of the word returning better Hits rate, the template which mentions the word in use is “*too specific term*” will be shown.
- If both words are **not related** within WordNet, only the suggested words are given as a possible better word based on Confusing Wordlist.

Moreover, the suggested word will be given which Thai definition and Thai translation provided by digital Thai-English bilingual dictionary, LEXiTRON (NECTEC LEXiTRON, 1995) to give more details for learners.

3. Experiment

The objective of this experiment was to evaluate the usability of the prototype. Research subject is fifteen volunteered Thai students studying in Grade 8 from a provincial boarding school in Chonburi, Thailand. Each student was assigned to compose ten sentences in both Thai and English. Specific words from Confusing Wordlist shown in Figure 6 must once be used in the each of written English sentences. The total of those 150 sentences were given to the system and returned results given in Figure 7.

able / above / accept / accord / accordance / according to / advice / affect / agree / bring / capable / citizen / come / converse / draw / effect / enable / equal / equivalence / espadrille / exact / except / flip-flop / flipflop / go / going to / gonna / good / guide / higher / in accordance with accurate / inhabitant / introduce / lay / lie / mule / native / over / people / population / prevent / protect / raise / rise / said / same same / sandal / say / slippers / speak / suggest / take / talk / tell / thongs / well / write

Figure 6. Specific words from Confusing Wordlist assigned to students

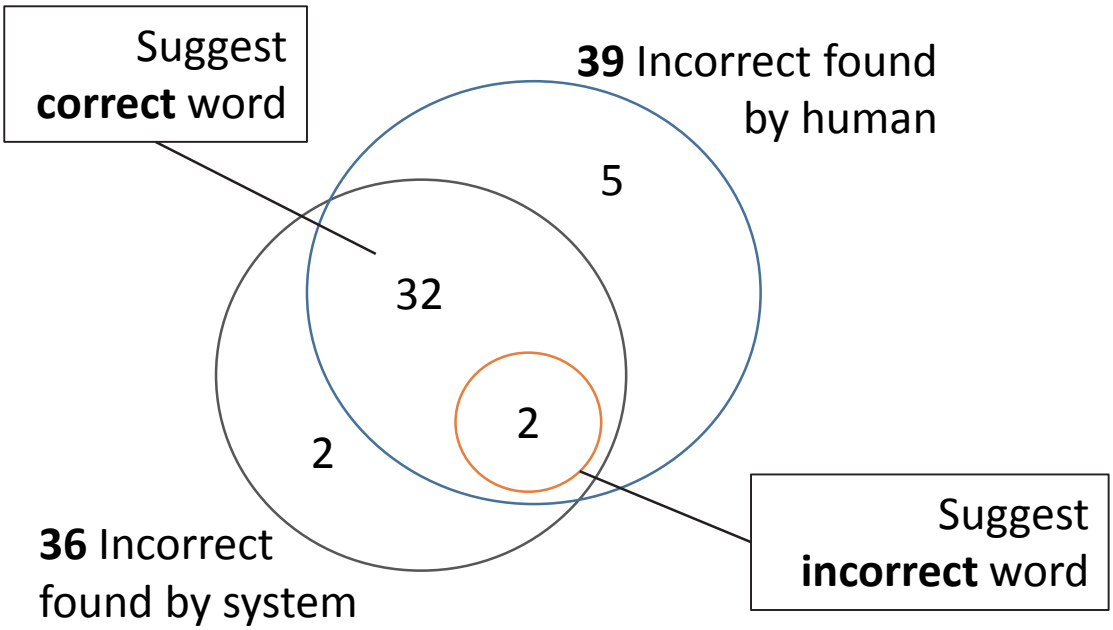

Figure 7. Results of System comparing with human

4. Discussion

From results given in Figure 7, the system returned 36 found incorrect words from 150 sentences. From those 36 found words, there were two words that are apparently correct but the system marked as incorrect. These findings were caused by the lower Hits rate of the correct one. In details, the found one is “I speak English well”. Since the n-gram chunked the sentence into gram based fragments, the immediate word ‘I speak’ was focused while the word ‘speak’

is in the Confusing Wordlist along with ‘say, talk, etc.’ as shown in Figure 4 line 7. Hench, the system decided to try ‘I say’ which gives much higher Hits ratio. This issue was generated from considering low gram number. The less n-gram word is searched for Hits, the more Hits ratio will be returned. Therefore, bi-gram word should be avoided and high number of n-gram should take higher priority.

Moreover, there are five incorrect cases that the system cannot find. This issue comes from the insufficiency of the words in the Confusing wordlist. Since the Confusing wordlist was only gathered with publicly provided data, there are other possible confusing words. It is best to gather more and more words to cover the list. Furthermore, type of confusing should be categorized to increase the scope and detail of the confusion.

Last, there are two of the suggested hints led to incorrect word selection. Since the system provides the list of possible words, students, who lack knowledge of the words though there are additional details given by bilingual dictionary, cannot select the proper word among them. It is possible that the details given by the bilingual dictionary are insufficient for low-proficiency student to understand. Thus, to facilitate more details, another detail related to the word in Thai should be added. Furthermore, an image of the word will be greatly helpful to exemplify the visual instance of the word mentioning an entity while video of motion is an explicit example of acting semantic.

5. Conclusion and Future work

This paper presents an assisting tool to help EFL Thai learners to select proper words by their semantic that they intend to. This work applies the Hits ratio from search engine to consider the words in use as statistical concordance. The low Hits result is used to find an inappropriate words using in co-occurrence. By employing confusing wordlist, the found incorrect words in consequence are given with the reason why the words are incorrect. The result of the system is to find the incorrect word among writing work with the suggest words and reason of the confusion. From testing with 150 written sentences by Grade 8 Thai students, 34 from 39 incorrect words were found while 32 from 34 found results were suggested accurately with words containing appropriate concepts to the context.

To improve the system, we plan on adding more confusing words based on Thai to English translation to increase coverage. Types of confusing will be categorized to scope and give better reason for word incorrectly used in writing. Additional information such as an image of entity and a short video of motion will be attached to suggested hint for learner to clearly understand the implicit concept of word.

Acknowledgements

This research was conducted under a grant in the SIIT-JAIST-NECTEC Dual Doctoral Degree Program.

References

Speaks, Jeff, "Theories of Meaning", The Stanford Encyclopedia of Philosophy (Fall 2014 Edition), Edward N. Zalta (ed.), forthcoming. (April 23, 2014) Retrieved http://plato.stanford.edu/archives/fall2014/entries/meaning/.

Microsoft Bing (June 1, 2009). Retrieved http://www.bing.com/

Google (September 4, 1998). Retrieved http://www.google.com/

Princeton University “About WordNet.” WordNet. (2010). Retrieved http://wordnet.princeton.edu/

NECTEC LEXiTRON “LEXiTRON: an online Thai-English Electronic Dictionary.” (1995). Retrieved http:://lexitron.nectec.or.th/