Linguistic Research 32(1), 1-20
DOI: 10.17250/khisli.32.1.201504.001

The use of context vectors in determining Thai compounds*

Wirote Aroonmanakun
(Chulalongkorn University)

Aroonmanakun, Wirote. 2015. The use of context vectors in determining Thai compounds. *Linguistic Research* 32(1), 1-20. In this paper we first discuss the problem of identifying compound words in Thai. It will be shown that the structures of compound words are often identical to the structure of phrases or sentences. Determining whether a sequence of words is a compound or a phrase or a sentence has to be determined within the context in which it occurs. Therefore, there is no clear cut way of determining the boundary of a compound. The longer the word sequence, the less likely it is that it will be considered a compound. In this study, we focus on extracting compound words consisting of two words from a large corpus using a vector space model. The basic assumption is that the context in which a compound occurs should be different from the context in which its parts occur. Two experiments were conducted on known compounds and on general bigram words. The test on known compounds was to verify that the cosine similarity of the context vector clearly indicates the differences of context vectors between the compound and its parts. The test on general bigram words was to further verify that the cosine similarity of context vectors between a non-compound bigram and its parts is different from that found in known compounds. When applying the cosine similarity of context vectors to compound candidates which have been extracted from a large corpus and ranked by statistics of collocation, we can determine a compound correctly with the F-measure at 0.81. The results indicate that the cosine similarity of context vectors is useful for determining a compound in Thai. **(Chulalongkorn University)**

Keywords Thai compound, vector space model, compound extraction, context vector

* I would like to thank two anonymous reviewers for their valuable comments and suggestions. This research is supported by the Project on Language and Human Security in Thailand, within the Integrated Academic Innovation Initiative, Chulalongkorn University's Academic Development Plan (CU Centenary) and partly supported by the Ratchadaphiseksomphot Endowment Fund of Chulalongkorn University (RES560530083-HS).

1. Introduction

Thai is an isolating language which does not have any markers for word boundaries or sentence boundaries. It is an alphabetical language in which consonant and vowel characters are written continuously without a space or any delimiter. Without explicit markers for word boundaries, ambiguity of segmentation is often found in Thai texts. For example, โคลงเรือ could be segmented as three words โค-ลง-เรือ kho:1-loŋ1-ri:a1 [cow | get down | boat], or two words โคลง-เรือ khlo:ŋ1-ri:a1 [shake | boat]. Therefore, when processing the Thai language, segmentation is a fundamental task that has to be resolved. One of the difficulties in Thai word segmentation comes from unknown words, which could be proper names or compounds not yet included in the dictionary (Aroonmanakun 2002). Given that compounding is a productive process for creating a new word in many languages, including Thai, extracting new compound words from a corpus is a challenging and important task for natural language processing. Even in a language that has word boundary markers, a compound which is composed of many words still needs to be identified. Much work related to extracting compound words has been done on many languages, such as English (Church and Hank 1990, Smadja et al. 1996), Chinese (Jian et al. 2000, Liu et al. 2003), Japanese (Nakagawa and Mohi 2002), Thai (Sornlertlamvanich and Tanaka 1996, Sornlertlamvanich et al. 2000, Suwanno et al. 2005, Aroonmanakun 2009), Hindi (Kunchukuttan and Damani 2008). etc. In fact, determining whether a given string is a compound is a basic problem in linguistics so, in this paper, we discuss the nature of compounding in Thai and focus on the problem of identifying Thai compounds. The problem of distinguishing between a compound and a phrase or a sentence will be discussed and then the idea of using context vectors to analyze 2-word compounds will be explored.

2. Thai Compounds

A compound is a complex word consisting of two or more lexemes (Aronoff and Fudeman 2011:47, Booij 2007:137). It is a kind of multiword expressions, which are a key problem in natural language processing (Sag et al. 2002). In general, the meaning of a compound is not the same as the sum of the meanings of its parts. The

meaning can be entirely different from its parts, such as หาง-เสือ ha:ŋ5-si:a5 [tail | tiger] = 'helm', ลูก-น้ำ lu:k3-na:m4 [child | water] = 'mosquito larva' and น้ำ-หนัก na:m4-nak2 [water | heavy] = 'weight'. For some compounds, the meaning can be partially related to its parts, for example, หมอ-ดู mɔ:5-du:1 [doctor | watch] = 'fortune teller' is a kind of specialist who can foresee the future. The word ของ-กิน khɔ:ŋ5-kin1 [thing | eat] = 'food stuff' is stuff that is edible. The word ชาว-นา cha:w1-na:1 [people | field] = 'farmer' refers to someone who grows rice in a field. Compounds in Thai include those involving reduplication of form or meaning, for example, โกรธ-เคือง kro:t2-khi:aŋ1 [angry | angry] means 'be angry' and the word หวาด-กลัว wa:t2-klu:a1 [frightened | fear] means 'be afraid of'.

One of the important questions in compound study is how to determine the boundary of a compound. This question is directly related to another question; how to distinguish between a compound and a phrase? It is often found in Thai that a compound can have an identical form to a phrase or a sentence, for example, the compound คนขับรถ khon1-khap2-rot4 [man | drive | car] 'driver', in some contexts, can also be in a sentence referring to an event in which a man is driving a car. In addition, some research on compounding focuses on the semantic relation of components within a compound and this research may prove useful in identifying such a compound.

Previous pieces of research on the identification of compounds have included compound extraction, identifying multiword expressions and term extraction in their titles and both linguistic rules and statistical methods were used in those studies (Su et al. 1994, Jian et al. 2000, Liu et al. 2003, Nakagawa and Mohi 2002, Attia et al. 2010). In relation to research on identifying semantic relations in a compound, there has been a debate over the number of semantic relations (Lauer 1995, Girju et al. 2005, Spencer 2011). Some have proposed a limited set of relations e.g. Lees (1960), Rosario and Hearst (2001) while some have argued for an open set of semantic relations, e.g. Downing (1977), Spencer (2011). The latter group has argued that relations are not always fixed and they sometimes depend on the context. In addition to this research, in this study, we will use a vector space model in determining compounds in Thai.

The vector space model (VSM) was proposed in Salton et al. (1975) for informational retrieval tasks. The idea is to convert each document into a point or a

vector in a vector space. Points or vectors that are closer in the vector space are more semantically similar. The value of each element in the vector can be derived from word frequency found in a corpus. VSMs have been used in measuring meaning similarity not only in documents but also words and phrases in NLP research (Turney and Pantel 2010).

Given that the meaning of the compound should be entirely different or partially different from its parts, this means that the context in which the compound occurs should be different from the context in which one of its components occurs. The vector space model can be used to measure these differences in terms of vector distance (Turney and Pantel 2010) and has already been successfully used in many works, e.g. those on word sense disambiguation, word sense discrimination, word similarity, text similarity and identifying multiword expressions (Pedersen 2010). In compound analysis, the contexts can be represented as a vector of words found in a specified context span and when they are represented as a vector, their similarity can be measured by calculating the cosine value of the two vectors. Below is an example of context vectors for the compound and its part. Given that xy is a compound consisting of x and y, and wi is the word found in the specified context, the context vector is a series of frequency of wi found near xy, x, or y.

Table 1.

	w1	w2	w3	w4	w5	w6	w7	w8	…
xy	f(xy-w1)	f(xy-w2)	f(xy-w3)	f(xy-w4)	f(xy-w5)	f(xy-w6)	f(xy-w7)	f(xy-w8)	…
x	f(x-w1)	f(x-w2)	f(x-w3)	f(x-w4)	f(x-w5)	f(x-w6)	f(x-w7)	f(x-w8)	…
y	f(y-w1)	f(y-w2)	f(y-w3)	f(y-w4)	f(y-w5)	f(y-w6)	f(y-w7)	f(y-w8)	…

3. Framework of Analysis

Before starting work on compound identification, a list of known compounds should first be analyzed since this will ensure that any proposed method is valid for compounds in Thai. Following this, the same method should be tested on n-gram words to see whether this yields significantly different results from the previous one. This will indicate that the method affects mostly compounds and not n-gram words. In addition, to better understand how to identify a compound, an analysis of known

compounds should be carried out. Possible structures and relations within the compound should be done first and this information can then be used to identify a word sequence that could be a compound.

A list of known compounds can be extracted from a dictionary. A word segmentation program was applied to all words in a Thai dictionary. Each word in the dictionary then was segmented into smaller parts and the result is a list of words that can be viewed as a sequence of smaller words. However, not all of these segmented word sequences in the list are compounds. They have to be manually examined. For example, even though it can be segmented as หน้า-ที่ na:3-thi:3 [face | that], the word หน้าที่ na:3-thi:3 is not a compound because it cannot be inferred how the compound has been created from these words. The meaning of the compound should be inferred from or describe how it is related to the meanings of its parts, for example, กุญแจปากตาย kun1cɛ:1pa:k2ta:j1 is a combination of three words กุญแจ-ปาก-ตาย kun1cɛ:1-pa:k2-ta:j1 [key | mouth | fixed], which means 'wrench'. It can be seen that the meaning of this compound can be inferred from its parts as an instrument which has a part that looks like a mouth and is fixed. From 32,653 words in the Thai dictionary, 18,738 patterns of combinations were found. Within these patterns, only 4,687 patterns could be analyzed as compounds. Most of them (94.54%) were two-word compounds and a few of them were three-word and four-word compounds. Most of the four-word compounds were reduplications, e.g. กลุ้ม-อก-กลุ้ม-ใจ klum3-?ok2-klum3-caj1 [worry | chest | worry | heart], 'be worried'.

To understand how the compounds are constructed, we analyzed their structure and dependency relation between POSs. The results show that many relations are possible but those related to N-N and V-N were the most frequently used pattern (28% and 26%). Relations between V-V, N-V, and N-A were also found for 13%, 12% and 10% respectively. Probabilities of possible POS relations could be used later in determining a compound candidate.

Table 2. Distribution of POS relations in Thai compounds

POS reation	Tokens	Percentage
N-N	1,395	27.58%
V-N	1,323	26.16%
V-V	646	12.77%
N-V	617	12.20%
N-A	494	9.77%
A-N	224	4.43%
V-A	170	3.36%
A-A	132	2.61%
A-V	39	0.77%
N-P	8	0.16%
V-P	8	0.16%
P-A	1	0.02%
P-N	1	0.02%
	5,058	100.00%

In addition, different types of compound were analyzed. Two major types, endocentric compounds and exocentric compounds, are distinguished on the basis of whether or not the core meaning of the compound is similar to the meaning of one part (Bloomfield 1933:235, Aronoff and Fudeman 2011:114, Booij 2007:139). In other words, endocentric compounds have one word functioning as a semantic head of the compound. Examples of Thai endocentric compounds are as follows: เรือ-รบ ri:a1-rop4 [ship | war] 'war ship', ภาพ-นิ่ง pha:p3-niŋ3 [image | still] 'still image', น้ำ-แข็ง na:m4-khɛŋ5 [water | hard] 'ice' etc. Conversely, exocentric compounds do not have one semantic head but both words are equally important, such as ล้ม-ละลาย lom4-la4la:j1 [fall | melt] 'bankrupt', ลูก-เสือ lu:k3-si:a5 [child | tiger] 'scout', ตก-งาน tok2-ŋa:n1 [fall | job] 'out of job' etc. Within exocentric compounds, special types are identified in this study. Reduplication is marked when a part of the word is found to be reduplicated. Coordinate compounds in this study are used in a loose sense including those analyzed as appositive compounds (see Dressler 2006), e.g. พ่อแม่ phɔ:3-mɛ:3 [father | mother], 'parent', and semantic reduplications, e.g. ท่องจำ thɔŋ3-cam1 [memorize | remember], 'memorize'. These compounds consist of words from the same class and they can have the same or different or opposite meanings.

Basic analyses of these known compounds should provide us with some information to determine a possible compound in terms of structure and relations but the question whether a sequence of words is a compound or a phrase or sentence still remains untouched. Singnoi (2000) and Prasithrathsint (2010) have given some examples to indicate that whether or not a sequence of words is a compound cannot be determined solely without the context. For example, the words ไข่เป็ด khai2pet2 could be either a compound or a noun phrase. In the example (1a), ไข่เป็ด khai2pet2 is a compound. This can be seen from the use of the classifier ฟอง fɔ:ŋ1, which is the classifier for 'egg'. But in (1b), ไข่เป็ด khai2pet2 is a noun phrase. This can be seen from the use of the classifier ตัว tu:a1, which is the classifier for 'duck'. In example (2a), ไฟไหม้ fai1mai3 is a compound because the use of the classifier แห่ง modifies the compound 'fire' ไฟไหม้ fai1mai3 but in example (2b), the same string is a reduced noun phrase, which uses the fire that is burning to contrast with the other reduced noun phrase, which is the fire that is being put out.

(1) a. ไข่เป็ด | ฟอง | นี้
khai2pet2 | fɔ:ŋ1 | ni:4
[egg-duck | clss. | this]
'This duck egg'

b. ไข่ | เป็ด | ตัว | นี้
khai2 | pet2 | tu:a1 | ni:4
[egg | duck | clss. | this]
'The egg of this duck' Singnoi (2000)

(2) a. มี | ไฟไหม้ | ห้า | แห่ง | ใน | เมือง.
mi:1 | fai1mai3 | ha:3 | hɛŋ2 | nai1 | mɨ:aŋ1
[there be | fire | five | clss. | in | town]
'There were fires in five places in the town'

b. ไฟ | ไหม้ | น่ากลัว // ไฟ | มอด | ไม่ | น่ากลัว.
fai1 | mai3 | na:3klu:a1 // fai1 | mɔ:t3 | mai1 | na:3klu:a1
[fire | burn | be frighten // fire | off | not | be frighten]
'The fire that is burning is frightening. The fire that is being put out is not frightening.'

Prasithrathsint (2010)

Another example below shows that the same word sequence can be analyzed as a compound or a sentence. In (3a), คนขับรถ khon1-khap2-rot4 [man | drive | car] can be a compound word, while in (3b), it cannot be a compound. Whether it is a compound or not depends on the context.

(3) a. มีคนขับรถมานั่งหน้าบ้าน
mi:1-khon1-khap2-rot4-ma:1-naŋ2-na:3-ba:n3
[there be | man | drive | car | come | sit | in front of | house]
'There is a driver sitting in front of the house'
b. มีคนขับรถมาชนหน้าบ้าน
mi:1-khon1-khap2-rot4-ma:1-chon1-na:3-ba:n3
[there be | man | drive | car | come | hit | in front of | house]
'There is a man driving a car and crashing into the front part of the house'

In addition to the above examples, previous research on compound nouns (Girju et al. 2005) has also indicated that the context plays a significant role in determining semantic relations in a compound. A compound consisting of A-B can be analyzed as a word having the meaning of A(x) and B(y) and some kind of relationship between the two words, R(x,y) where R represents some kind of relation between A and B. It is not difficult to find an example in which the relationship between the two words varies according to the context. For example, the word เก้าอี้-โรงงาน kaw3?i:3-ro:ŋ1ŋa:n1 [chair | factory] can be analyzed as a chair with some relationship to a factory as follows:

(4) a. a chair produced from a factory
b. a chair to be used in a factory
c. a chair that is already in the factory

This kind of example confirms that determining the meaning of the compound depends on the context. Moreover, it can be seen that the difference between a phrase and a compound is a matter of degree rather than an absolute. The more complex the compound is, the less likely it is to be a lexical unit. For example, it

is easy to accept (5a) and (5b) as a compound, while one is less likely to accept (5c) and (5d) as a compound word. The last two examples should be analyzed as a complex noun phrase referring to a specific person.

(5) a. คน-ขับ-รถ
khon1-khap2-rot4
[man | drive | car]
'a driver'

b. คน-ขับ-รถ-บรรทุก
khon1-khap2-rot4-ban1thuk4
[man | drive | car | load]
'a truck driver'

c. ?คน-ขับ-รถ-โดยสาร-ประจำ-ทาง-ปรับ-อากาศ
? khon1-khap2-rot4-do:j1sa:n5-pra2cam1-tha:ŋ1-prap2-?a:1ka:t2
[man | drive | car | take(ride) | regular | route | adjust | air]
'air-conditioned bus driver'

d. ??คน-ขับ-รถ-โดยสาร-ประจำ-ทาง-ปรับ-อากาศ-ใช้-ก๊าซ-ธรรมชาติ
??khon1-khap2-rot4-do:j1sa:n5-pra2cam1-tha:ŋ1-prap2-?a:1ka:t2-chai4-ga:s4-tham1ma4cha:t3
[man | drive | car | take(ride) | regular | route | adjust | air | use | gas | natural]
'air-conditioned CNG-fueled bus driver'

From the above examples, we can say that determining whether a given string is a compound or a phrase should become evident during parsing. Determining semantic relations within a compound also depends on the context. Therefore, this problem cannot be resolved without considering the whole sentence. The problem has to be resolved during the parsing process but this does not mean that extracting new compounds is not a necessary task. We still need to find what could be a possible compound in Thai and put those words in the dictionary. Otherwise, the parser will not know that a given sequence of words could possibly be viewed as a compound. Since a compound is less likely to be a lexical unit consisting as it does of many words, we can focus on finding new compounds that consist of two to four

words. Word sequences longer than four can be presumed and analyzed as a phrase.

In sum, we can conclude that there are two main tasks involved in processing compounds. The first task, which is the focus of this paper, is to extract true compounds from a corpus and add those new words to the dictionary. The second task is a parsing process, in which many possible analyses have to be carried out and the decision as to whether a sequence of words is a compound or a phrase can be reached within that sentence.

4. Compound Extraction

There are two modules for extracting new compounds. The first module is to rank candidate compounds and this module can use information, such as (a)-(d), to rank possible candidates. A probabilistic model can be used to search a corpus and rank compound candidates based on the product of these probabilities.

a. The probability of POS between head and dependent in a compound : Prob(Ci - Cj)
b. The probability that a word could be the head in a compound : Prob(Head(W)).
c. The probability that a compound will have a certain syntactic structure, such as W1->W2, W1->W3 : Prob(Hn->Dm)
d. Statistical collocation between any word forms : Colloc(W1,W2).

The second module, which is the focus of this paper, is to determine which candidate could be a compound and the semantic similarity between the compound (W1…Wn) and its components, W1, W2, … Wn is used in this module. The basic idea is that the meaning of the compound should be different from the meaning of the words inside and the meaning of the word is indirectly reflected by the context in which it occurs. Therefore, a vector space model is used to determine semantic similarity by measuring the cosine similarity between context vectors, i.e. cosine(vector(W1..Wn), vector(Wi) i=1,n).

In this study, we limit the study to known 2-word compounds found in the dictionary, which has 4,431 compounds. Given that the compound CP consists of

two words W1 and W2, the semantic difference between CP and W1 or W2 is measured by cosine(vector(context(CP)), vector(context(W1))) and cosine(vector(context(CP)), vector(context(W2))). If the cosine value is less than 0.5 or 0.7071, which reflects the distance of vectors of at least 60' and 45' respectively, we may assume that the meaning of CP is different from the meaning of W1 or W2. Three experiments were conducted using the 32 million words in the Thai National Corpus (Aroonmanakun 2007) to create context vectors. They were tested with different context spans, i.e. 5 words and 10 words, and with the exclusion of the top 100 words. The difference of span is tested to see whether nearby context (within a 5-word span) is sufficient for determining the difference in meaning. The selection of the top 100 words is for eliminating frequently used words since they may not have much predictive power for measuring meaning differences.

5. Findings

It can be seen in Table 2 that when setting the context span at 5, the result is better than the setting of the span at 10. This suggests that a nearby context is sufficient in determining difference of meaning. In addition, when excluding the top 100 words from the context, the result is better than with the other two settings. Thus, it can be inferred that this method is sensitive to low frequency words. The number “n” is used for measuring only when that bigram is found at least n times in the corpus. As seen in Table 2 and Figure 1, when the number of occurrences is higher, the number of items analyzed and the accuracy rate falls. Although the results seem to suggest that similarity of context vectors can be used to determine whether the two words could be a compound, we need to verify that this method does not give the same result for any two word bigrams. In the first experiment it was found that the local context at a 10 word span and with the exclusion of top 100 words yielded the best results. The same setting could then be used to measure the context vectors of bigram words taken from a random text. A comparison of these results is shown in Table 3. As can be seen in Table 3, the number of bigram words in its context is different from the contexts of its parts and far fewer than those found in known compounds. From the experiments, we can see that measuring semantic difference at a degree greater than 60 is more suitable than measuring at a

degree greater than 45 because, at this setting, the results from known compounds are clearly different from the results from general bigram words.

Table 2. Results when testing known compounds.

	distant > 60' or cosine < 0.5			distant > 45' or cosine < 0.7071		
	n>=50	n>=100	n>=200	n>=50	n>=100	n>=200
span=5	855/1571 54.42%	558/1165 47.90%	317/772 41.06%	1480/1571 94.21%	1075/1165 92.27%	690/772 89.38%
span=10	526/1571 33.48%	315/1165 27.04%	169/772 21.89%	1332/1571 84.79%	932/1165 80.0%	571/772 73.96%
span=10 + excl. Top 100 words	1311/1571 83.45%	926/1165 79.40%	564/772 73.06%	1553/1571 98.85%	1147/1165 98.45%	754/772 97.67%

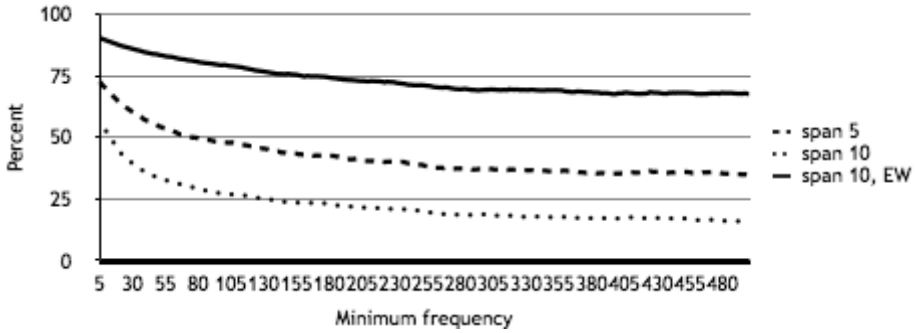

Figure 1. Accuracy rate at a different minimal frequency of bigram testing on a known compound with a degree difference greater than 60

Table 4. Results when testing bigram words from a random text

	distant > 60' or cosine < 0.5			distant > 45' or cosine < 0.7071		
span=10 + excl. Top 100	n>=50	n>=100	n>=200	n>=50	n>=100	n>=200
known compounds	1311/1571 83.45%	926/1165 79.40%	564/772 73.06%	1553/1571 98.85%	1147/1165 98.45%	754/772 97.67%
bigrams	381/1128 33.78%	267/998 26.75%	165/843 19.57%	823/1128 72.96%	693/998 69.44%	542/843 64.29%

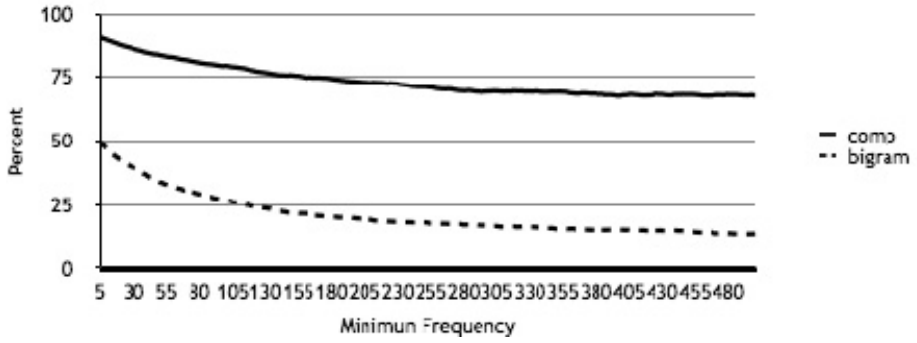

Figure 2. Accuracy rate at a different minimal frequency of bigram testing on a random text with degree difference greater than 60

The result indicates that context vectors can be used to determine a compound. It displays a great difference of contexts when bigram words are taken from known compounds, while it shows less difference when bigram words are taken from a random text. However, how accurate this method is in predicting a compound in a random text is yet to be answered. To answer this question, we tested this method on the list of candidate compounds created in Aroonmanakun (2009), in which collocational strength were mainly used in ranking the candidates. The result is shown in Table 4. The second column shows the accuracy of bigram candidates that are a compound in each cutoff level. It can be seen that when a list of bigram candidates is extracted from a large corpus, context vectors can be used to determine the compound at a rate of 0.73 - 0.81. When looking at the top 50 ranked candidates, this method can identify a compound with a precision of 0.727 and recall of 0.914. When the number of candidates is increased, the precision drops to 0.6 while the recall remains the same. Therefore, it can be concluded that when using context vectors, the precision rates are a bit higher than the accuracy rate found in the candidate sets. Only a few candidates were eliminated by the use of context vectors. A large number of candidates consist of a high-frequency function word like ที่ thi:3 ‘that’, ของ khɔ:ŋ5 ‘of’, ใน nai1 ‘in’, and a number of candidates are a bigram that is a part of larger compound. These candidates cannot be eliminated by the use of context vectors. A further study is needed to eliminate these candidates before applying context vectors. In addition, some of the compounds, such as ทำ-งาน tham1-ŋa:n1 [do | work], ‘work’, phɔ:3-mɛ:3 พ่อ-แม่ [father | mother], ‘parent’ could

not be detected because their contexts are not clearly distinct from the context in which the parts occur.

Table 5. Precision and recall when applying the context vector in determining a compound

Cutoff	Accuracy	Precision	Recall	F-measure
50	0.700 (35/50)	0.727 (32/44)	0.914 (32/35)	0.810
100	0.630 (63/100)	0.644 (56/87)	0.889 (56/63)	0.747
200	0.615 (123/200)	0.628 (108/172)	0.878 (108/123)	0.732
300	0.607 (182/300)	0.624 (164/263)	0.901 (164/182)	0.737
400	0.605 (242/400)	0.620 (219/353)	0.905 (219/242)	0.736
500	0.610 (305/500)	0.621 (277/446)	0.908 (277/305)	0.738
600	0.618 (371/600)	0.630 (339/538)	0.914 (339/371)	0.746
700	0.617 (432/700)	0.629 (397/631)	0.919 (397/432)	0.747
800	0.610 (488/800)	0.623 (451/724)	0.924 (451/488)	0.744
900	0.602 (542/900)	0.615 (504/819)	0.930 (504/542)	0.741
1000	0.592 (592/1000)	0.603 (553/917)	0.934 (553/592)	0.733

6. Analysis of Context Vector on Compound Types

The previous section shows that it is possible to use context vectors in determining a compound in Thai. It would be interesting to see whether this method can be used in determining different types of compound. Since endocentric compounds have one word as a head, we would expect to find that the cosine similarity of context vectors between W1 and CP, or between W2 and CP should be less than 0.5. For exocentric compounds, since there is no clear head, we would expect that the cosine similarity between both W1 and CP and W2 and CP would be less than 0.5. For coordinate compounds, in which W1 and W2 are from the same semantic class, it is expected that the cosine similarity of the context vectors of W1 and W2 would be greater than 0.5. However, the results when applying this method to different types of compound do not conform to those predictions.

Endocentric compounds:

Expected : cosine(vector(either W1 or W2),vector(CP)) < 0.5

Exocentric compounds

Expected : cosine(vector(both W1 and W2),vector(CP)) < 0.5

Coordinate compounds

Expected : cosine(vector(W1,W2)) > 0.5

Table 5. Contextual differences on different types of compound

	Dist(W1, Wc)	Dist(W2, Wc)	~Dist(W1, Wc)	~Dist(W2, Wc)	Dist(W1,Wc) and Dist(W2,Wc)	~Dist(W1,Wc) and ~Dist(W2,Wc)	Dist(W1,Wc) or Dist(W2,Wc)
Endo	610	605	249	254	497	141	718
859	71.01%	70.43%	28.99%	29.57%	57.86%	16.41%	83.59%
Exo	405	393	131	143	351	89	447
536	75.56%	73.32%	24.44%	26.68%	65.49%	16.60%	83.40%
CC	121	116	52	57	91	30	146
173	69.94%	67.63%	30.06%	32.37%	52.60%	17.34%	84.39%

Both endocentric and exocentric compounds including coordinate compounds seem to behave in a similar way. The number of compounds whose context is different from the context of either part is more than 70%. Both types have contextual differences from both W1 and W2 for more than 50%. About 16-17% do not have any contextual difference from either W1 or W2. Therefore, the idea of using context vectors in distinguishing different types of compound is not supported in this study. It is likely that for any type of compound, the compound will normally occur in different contexts from its parts. For example, an endocentric compound กระดาน-ดำ kra1da:n1-dam1 [board | black] 'blackboard', although it can be analyzed as a kind of board, its referent is uniquely different from other kinds of board. It would not occur in the same context as the word กระดาน 'board'. This explains why we cannot distinguish different types of compound using context vectors.

7. Discussion

In this study, we have shown that vector space model can be used for determining new compounds in Thai. Determining whether a sequence of words is a compound or a phrase depends on the context in which it occurs. This has to be done during the parsing process. But new compounds have to be identified from a corpus and added to a dictionary first. Otherwise, the parser would not know that those sequences of words could be analyzed as a compound in Thai. We suggest that only 2-4 word compounds should be extracted from a large corpus since most of Thai compounds consist of 2-4 words. In this paper, we conducted two experiments on extracting 2 word compounds using a vector space model: one on known compounds extracted from a Thai dictionary and the other one on general bigram words. From the above results, we can conclude that context vectors can be used in determining a compound but not its types. However, the cosine similarity of context vectors can reflect the semantic similarity or difference between the compound and its parts. For a known compound, context vectors can be used to decide which word in a compound is more semantically similar to the compound. Table 5 shows some examples of compounds and the cosine similarity of context vectors.

Table 6. Cosine similarity values and semantic relatedness to the compound

CP	W1	W2	cos(vect(CP,W1))	cos(vect(CP,W2))	more related to
น้ำ-เกลือ na:m4-klɨ:a1 ‘saline solution’	น้ำ na:m4 ‘water’	เกลือ klɨ:a1 ‘salt’	0.459	0.488	neither
คน-ใช้ khon1-cha:i4 ‘servant’	คน khon1 ‘man’	ใช้ cha:i4 ‘use’	0.783	0.702	both
ห้อง-น้ำ hɔŋ3-na:m4 ‘restroom’	ห้อง hɔŋ3 ‘room’	น้ำ na:m4 ‘water’	0.83	0.512	‘room’
เล็ก-น้อย lek4-nɔ:j4 ‘little’	เล็ก lek4 ‘small’	น้อย nɔ:j4 ‘little’	0.583	0.8	little’

As can be seen in Table 6, the word น้ำ-เกลือ na:m4-klɨ:a1 [water | salt], ‘saline

solution' has cosine values between the compound and its parts of 0.459 and 0.488 respectively. It means that the contexts of both words are quite different from the context of the compound. In other words, the compound is not really related to either 'water' or 'salt'. In the word คน-ใช้ khon1-cha:i4 [man | use], 'servant' has cosine values of 0.783 and 0.702 and this means that this compound is semantically related to both 'man' and 'use' but veers a little bit more to the 'man' side. In the word ห้อง-น้ำ hɔŋ3-na:m4 [room | water], 'restroom' has cosine values of 0.83 and 0.512 and this indicates that this compound is more related to 'room' than 'water'. In the word เล็ก-น้อย lek4-nɔ:j4 [small | little], 'little' has cosine values of 0.583 and 0.8 and this compound is a semantic reduplication. The value suggests that it is more related to 'little' than 'small'. The use of context vectors can be very useful in determining which part of semantic duplicated compounds is the semantic head of the compound. Whether the semantic head of semantic duplication in Thai compounds should be in the left or the right part can be roughly determined by this method. When testing 143 semantic duplicated two-word compounds, we found that 74 compounds had the right part as the main head while 69 compounds had semantic head on the left part. For example, เจ็บ-ป่วย cep2-pu:aj2 [hurt | sick] 'sick', วิตก-กังวล wi4tok2-kaŋ1won1 [anxious | worry] 'worry', เพ้อฝัน pɔ:4fan5 [drivel | dream], 'dream', etc., all have a meaning related more to the right part but แน่-แท้ nɛ:3-thɛ:4 [sure | real], 'be sure', เจ็บ-ปวด cep2-pu:at2 [hurt | ache] 'hurt', ถก-เถียง thok2-thi:aŋ5 [discuss | argue] 'discuss' etc., are more related to the left part of the compound. Therefore, semantically reduplicated compounds in Thai do not have a preference as to the head initial or head final.

In sum, context vectors can be used not only to indicate the compound but also to determine the semantic similarity of its parts. This information may be useful for teaching Thai compounds to foreign students. For further research, as mentioned in section 5, context vectors cannot be used to determine a compound when its part is a high-frequency function word. How to exclude these word sequences from compound candidates is the next step to be carried out. This could be done by applying POS tagging and eliminating candidates whose POS sequences are unlikely a possible compound. Moreover, the use of context vectors in this study is based mainly on word forms. How can context vectors be used in compounds consisting of a highly polysemous word would be an interesting research topic.

References

Aronoff, Mark and Kirsten Anne Fudeman. 2011. *What is morphology?* Chichester, West Sussex, U.K. ; Malden, MA: Wiley-Blackwell.

Aroonmanakun, Wirote. 2002. Collocation and Thai word segmentation. In Thanaruk Theeramunkong and Virach Sornlertlamvanich, eds. *Proceedings of the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop*. Pathumthani: Sirindhorn International Institute of Technology. 68-75.

Aroonmanakun, Wirote. 2007. Creating the Thai national corpus. *Manusya*. Special Issue 13: 4-17.

Aroonmanakun, Wirote. 2009. Extracting Thai compounds using collocations and POS Bigram probabilities without a POS tagger. Paper presented to the *Asian Language Processing, 2009*. IALP '09. International Conference on, 2009.

Attia, Mohammed, Antonio Toral, Lamia Tounsi, Pavel Pecina and Josef van Genabith. 2010. Automatic extraction of Arabic multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*. Beijing, China: Association for Computational Linguistics.

Bloomfield, Leonard. 1933. *Language.* New York,: H. Holt and Company.

Booij, G. E. 2007. *The Grammar of Words : An Introduction to Linguistic Morphology* New York: Oxford University Press.

Church, Kenneth Ward and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics.* 16(1): 22-29.

Downing, Pamela. 1977. On the creation and use of English compound nouns. Language 53: 810-842.

Dressler, W.U. 2006. Compound Types. *The Representation and Processing of Compound Words. In* G. Libben and G. Jarema (eds.). 15: 242. Oxford ; New York: Oxford University Press.

Girju, Roxana, Dan Moldovan, Marta Tatu and Daniel Antohe. 2005. On the semantics of noun compounds. *Computer Speech and Language* 19: 479-496.

Jian, Zhang, Gao Jianfeng and Zhou Ming. 2000. Extraction of Chinese compound words: an experimental study on a very large corpus. In *Proceedings of the second workshop on Chinese language processing: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 12. Hong Kong*: Association for Computational Linguistics.

Kunchukuttan, Anoop and Om P. Damani. 2008. A system for compound noun multiword expression extraction for Hindi. Paper presented to the *6th International Conference on Natural Language Processing*, 2008.

Lauer, Mark. 1995. Corpus statistics meet the noun compound: Some empirical results.

Paper presented at the *The 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, Massachusetts, USA.

Lees, Robert B. 1960. *The Grammar of English Nominalizations*. Bloomington, IN.

Liu, Jianzhou, Tingting He and Xiaohua Liu. 2003. Extracting Chinese multi-word units from large-scale balanced corpus. Paper presented to *The 17th Pacific Asia Conference on Language, Information and Computation, 2003*.

Nakagawa, Hiroshi and Tatsunori Mori. 2002. A simple but powerful automatic term extraction method. Paper presented at the *COLING-02 on COMPUTERM 2002: Second International Workshop on Computational Terminology* - Volume 14.

Pedersen, Ted. 2010. *Computational Approaches to Measuring the Similarity of Short Contexts : A Review of Applications and Methods*. University of Minnesota Super Computuing Institute UMSI 2010/165.

Prasithrathsint, Amara. 2010. Lexicalization of syntactic constructions in Thai. Paper presented to *The 20th Anniversary Meeting of the Southeast Asian Linguistics Society, Zurich, Switzerland 2010*.

Rosario, Barbara and Marti Hearst. 2001. Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. Paper presented at the *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. *Computational Linguistics and Intelligent Text Processing.* In A. Gelbukh (ed.). 1-15. Springer Berlin Heidelberg.

Salton, Gerard, A. Wong and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM* 18: 613-620.

Singnoi, Unchalee. 2000. *Nominal Constructions in Thai*: University of Oregon Ph.D. Dissertation.

Smadja, Frank, Kathleen R. McKeown and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*. 22(1): 1-38.

Sornlertlamvanich, Virach, Tanapong Potipiti and Thatsanee Charoenporn. 2000. Automatic corpus-based Thai eord extraction with the c4.5 learning algorithm. Paper presented at the *Proceedings of the 18th conference on Computational linguistics* - Volume 2, Saarbrücken, Germany.

Sornlertlamvanich, Virach & Hozumi Tanaka. 1996. The automatic extraction of open compounds from text corpora, Paper presented to the *COLING '96.*

Spencer, Andrew. 2011. What's in a compound? *Journal of Linguistics* 47: 481-507.

Su, Keh-Yih, Ming-Wen Wu and Jing-Shin Chang. 1994. A corpus-based approach to automatic compound extraction. Paper presented at the *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, Las Cruces, New Mexico.

Suwanno, N., Y. Suzuki and H. Yamazaki. 2005. Extracting Thai compound nouns for para-

graph extraction in Thai text. Paper presented to the *Natural Language Processing and Knowledge Engineering*, 2005. IEEE NLP-KE '05. *Proceedings of 2005 IEEE International Conference on*, 2005.

Turney, Peter D. and Patrick Pantel. 2010. *From Frequency to Meaning: Vector Space Models of Semantics.* CoRR abs/1003.1141.

Wirote Aroonmanakun
Department of Linguistics
Faculty of Arts, Chulalongkorn University
254 Phyathai Rd., Prathumwan,
Bangkok 10330, Thailand
E-mail: awirote@chula.ac.th

Received: 2014. 10. 26.
Revised: 2014. 12. 02.
Accepted: 2014. 12. 02.