Automatic Extraction of Thai-English Term Translations and Synonyms from Medical Web Using Iterative Candidate Generation with Association Measures

Kobkrit Viriyayudhakorn[1], Thanaruk Theeramunkong[1],
Cholwich Nattee[1], Thepchai Supnithi[2], and Manabu Okumura[3]

[1] Sirindhorn International Institute of Technology, Thammasat University
131 Moo 5, Tiwanont Rd., Bangkadi, Muang, Pathumthani, 12000, Thailand
Tel: +66 (0) 2501 3505-20, Fax: +66 (0) 2501 3524
[2] National Electronics and Computer Technology Center
112 Paholyothin Road, Klong 1,
Klongluang, Pathumthani, 12120, Thailand
[3] Precision and Intelligence Laboratory
Tokyo Institute of Technology
4259 Nagatsuta Midori Yokohama 226-8503, Japan
{kobkrit,cholwich,thanaruk}@siit.tu.ac.th,
thepchai.supnithi@nectec.or.th,oku@pi.titech.ac.jp

Abstract. Electronic technical documents available on the Internet are a powerful source for automatic extraction of term translations and synonyms. This paper presents an association-based approach to extract possible translations and synonyms by iterative candidate generation using a search engine. The plausible candidate pairs can be chosen by calculating their co-occurring statistics. In our experiment to extract Thai-English medical term pairs, four possible alternative associations; namely confidence, support, lift and conviction, are investigated and their performances are compared by ten-fold cross validation. The experimental results show that lift achieves the best performance with 73.1% f-measure with 67% precision and 84.2% recall on translation pair extraction, 68.7% f-measure with 71.5% precision and 67.7% recall on Thai synonym term extraction and 72.8% f-measure with 72.0% precision and 75.1% recall on English synonym term extraction. The precision of our approach in Thai-English translation, Thai synonym and English synonym extraction are 4 times, 3.5 times and 5.5 times higher than baseline precision respectively.

Keywords: Association rule mining, Thai-English medical term translation, Conviction, Lift, Iterative approach, Synonym extraction.

T. Theeramunkong et al. (Eds.): PAKDD Workshops 2009, LNAI 5669, pp. 141–155, 2010.

1 Introduction

In recent years, there have been several attempts to extend text mining techniques to mine specific health-related knowledge, such as medical, pharmaceutical and biological [1,2,3] practices. Since health-related articles usually include a lot of technical terms, processing such terminologies becomes an important factor towards success of automating analysis of those articles. There are still a lot of challenges in developing Thai health-related terminology due to at least two reasons. First, currently there has been no standardization of health-related terminology in Thai languages. Second, it is a backbreaking task to add new terms or to modify information of terms in a conventional paper-based or online dictionary. Nowadays, since there are a lot of web pages providing information or knowledge related to health science, it is possible to use such pages as resources to construct health-related terminology. Normally, like texts in several non-English languages, Thai medical texts often include Thai technical terms followed by their corresponding English translations since English is widely recognized as a common language for interchanging technical information. Among several patterns of translation pairs, a common one is that the English translation of a Thai term is enclosed in a parenthesis or placed immediately after a term. For example, 'โรคกระเพาะ (peptic ulcer)' or 'โรคกระเพาะ peptic ulcer' denotes that the Thai word 'โรคกระเพาะ' has the term 'peptic ulcer' as its English translation.

This common regularity enables us to extract Thai-English translation pairs. However, there have been a few difficulties in extracting translation pairs from texts. First, a technical term may be translated into several different terms due to lacking of standardization. For instance, a term 'peptic ulcer' can be translated into four Thai translation terms, (1)'โรคกระเพาะอาหาร', (2)'โรคแผลในกระเพาะอาหาร', (3)'แผลเป็บติก' or (4)'แผลเพ็ปติก' where the first two terms are direct translation, and the last two terms are transliteration. While an English term can be translated to more than one Thai terms, a Thai term is also able to be mapped to several English terms. Therefore, a mechanism to select the best translation pair is needed. Second, due to authors' writing styles, an English term after a Thai term may not be its translation. As one example, in Thai medical texts, sometimes an English term is used directly as a word in a context, without specifying its corresponding Thai term, such as 'สามารถทำให้เกิด dengue fever' ('can trigger dengue fever'). This irregularity causes difficulty in detecting translation pairs.

Intuitively it is possible to detect synonyms by linking two translation candidate pairs. To create a candidate for Thai-Thai synonym, we link a Thai-English translation pair with another English-Thai translation pair that has identical English term. In the same manner, linking a English-Thai translation pair and a Thai-English translation pair will enable us to obtain a candidate for English-English synonyms.

Towards the above objectives, this paper presents a method to use Web documents as resources for extracting translation and synonym pairs of English and Thai. This paper is organized as follows. Some previous approaches are described in Section 2. Section 3 presents the framework and techniques for constructing translation and synonym pairs. In Section 4, the experimental results are discussed. Finally, a conclusion and some further works are given in Section 5.

2 Related Works

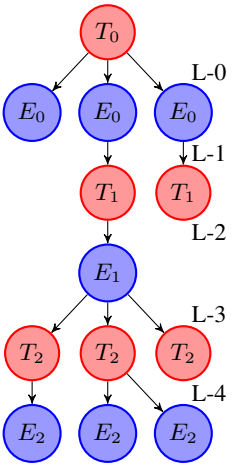
This section gives a survey to research works related to extraction of term translations and synonyms. Some recent works have been conducted to extract term pairs between English and Chinese translations from Chinese texts on the Internet. In [2], Zhang and Vines proposed a method that generated English-Chinese and Chinese-English translation candidates from top-100 search results from a search engine and then find potential translation pairs by exploiting co-occurrence frequency and surface characteristics, such as term length and common substring.

As another work, Wang and his colleagues [4,5,6] showed promising results of exploiting the Web as a source to generate effective translation equivalents for many unknown terms, including proper nouns, technical terms and Web query terms and in assisting bilingual lexicon construction for a real digital library system. Their method applied the Chi-square test and context vector analysis to rank cohesion among terms in web documents to tackle with low-frequency problem.

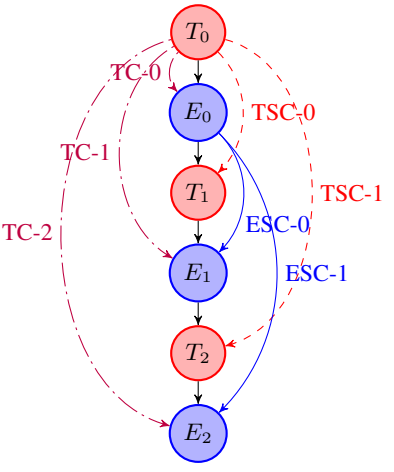
English synonym extraction by using an unsupervised learning algorithm based on statistical data was proposed in [7] and was improved by combining with symbolic knowledge in [8]. For Japanese language, Okamoto[9] extracted a set of near-synonyms by using semantic features from the Thesaurus and then weighting them with their occurrence probabilities under a set of heuristics. Shimohata[10] extracted synonyms from documents whose contents are similar by looking contextual information of surrounding words. In Thai language, the number of works in mining translation pairs is still limited. As our best knowledge, there is still no work on Thai synonym extraction using web documents.

3 Iterative Approach for Candidate Generation and Association-Based Candidate Selection

In this section, an approach to extract translation pairs and synonyms using iterative candidate generation and association-based selection is described. The two main steps in our approach are (1) iterative candidate generation and (2) candidate selection.



(a) Iterative approach for candidate generation



(b) Multi-level candidates of translation pairs and synonym pairs

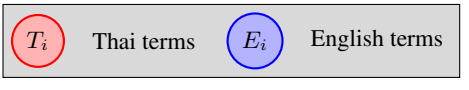


Fig. 1. Mapping candidate generation and multi-level candidates of translation pairs and synonym pairs (L-n: n^{th} iteration, TC-n: The n^{th}-level mapping of Thai-English translation pairs, TSC-n: The n^{th}-level mapping of Thai synonym pairs, ESC-n: The n^{th}-level mapping of English synonym pairs)

In the first step, an initial set of Thai terms are fed to a search engine to get their potential English translations from Thai web documents. Next, these obtained English translations are submitted again to the search engine in order to retrieve their potential Thai translations from Thai web documents. This process is iteratively executed by inputting the obtained Thai translations into the search engine repeatedly to obtain English translations. In general, it is possible to specify the language of search result pages returned from the search engines, such as Google[4] and Yahoo[5]. With this feature, we have chosen Thai as the language of our target pages for all iterations. The search results are used as a corpus to extract term translations and synonyms.

In the second step, after a number of potential translations are obtained, an association-based measure is applied to select the most plausible candidates as translation and synonym pairs. Figure 1 illustrates the whole process where the details are shown in the following subsections.

[4] http://www.google.com
[5] http://www.yahoo.com

3.1 Iterative Candidate Generation

As the first step, our proposed approach starts from a set of initial Thai technical terms and then find their English translations. Each Thai term is submitted as a query to a search engine. Then, a set of the top-k result snippets are obtained from the search engine. The common writing patterns are applied to each snippet to extract English translations of the term. Then, each obtained English translation is re-submitted to the search engine and different Thai terms may be extracted from the results. The process is repeatedly conducted for each newly obtained English or Thai terms until no new term is obtained from the search results or the number of iterations is higher than a pre-defined threshold. The result of the iterative extraction for each term is represented as a directed graph as shown in Figure 1(a). Each node of the graph denotes a term. Each arc links a term with its translated term in another language.

The iterative approach allows us to construct a set of potential translation candidates since a Thai term can be translated into more than one English terms, or there are more than one Thai terms used to refer to an English term. For example, a Thai term 'โรคหูด' is a translation of both 'ringworm' and 'tinea' which are synonym. By linking two related term translation pairs, the proposed method can be also used to generate a pair of synonyms. In this case, a Thai term can be used as an intermediate to obtain a synonym of the given English term. In the same manner, we can link two Thai terms through a English term. This linking can also be done through more than one intermediates as shown in Figure 1(b) to obtain more than one synonyms.

Since, there is no explicit word boundary in the Thai writing system, it is possible to extract an incorrect portion (string) as a word in a running text. For example, we may get a phrase 'มีอาการคล้ายไข้เดงกี' ('has symptoms similar to dengue fever') that includes an useless part 'มีอาการคล้าย' ('has symptoms similar to') in front of a suitable word 'ไข้เดงกี' because, as occurred often in Thai natural writing style, a space may not be inserted between that part and the target term 'ไข้เดงกี'('dengue fever'). To filter out these incorrect pairs, we have proposed to use association-based candidate selection. The basic idea is that the incorrect pairs usually have low occurrence frequency. More details will be described in the next subsection.

3.2 Association-Based Candidate Selection

The association analysis aims to evaluate the relationship between two sets of objects (set A and set B) written as $A \rightarrow B$. The $A \rightarrow B$ indicates that B is likely to occur when A occurs. In our experiments, if A is set to be the source language term and B is set to be the target language term, we can use $A \rightarrow B$ to indicate how likely B will be taken place when A occured. In other words, how likely B is the translation of A when

B is a term that is enclosed by parenthesis or is placed immediately after A in common writing pattern. Normally, association is quantified by a set of well-known measures in association rule mining; namely support, confidence, lift and conviction. They have strong and weak points under different situations. Next, we explain how the association measures are applied to measure the association between terms in a translation pair.

$N(X)$ is the number of pages that include the word X and $N(*)$ is the total amount of existing pages. Unfortunately computing $N(*)$ is impossible since the total number of Web pages indexed by search engine are not precisely estimated. However $N(*)$ may be trivial when only ranking result is concerned.

- **Support** is an undirected measure that specifies the ratio that A and B occur with respect to the total occurrence.

$$Support(A \rightarrow B) = \frac{N(A \wedge B)}{N(*)} \tag{1}$$

- **Confidence** is a directed measure specifying the ratio that A and B occurs when A occurred.

$$Conf(A \rightarrow B) = \frac{N(A \wedge B)}{N(A)} \tag{2}$$

- **Lift** or **Interestingness** is an undirected measure that has an advantage over confidence by exploiting negative association. Lift measures the proportion of A and B occurring together compared to the expected occurrence when they are considered statistically independent.

$$Lift(A \rightarrow B) = \frac{N(*)N(A \wedge B)}{N(A)N(B)} \tag{3}$$

- **Conviction** is a directed measure representing the proportion of A occurrences without B, comparing to the expected occurrence when they are dependent. i.e. $N(A)$ and $N(\neg B)$.

$$Conv(A \rightarrow B) = \frac{N(A)N(\neg B)}{N(*)N(A \wedge \neg B)} \tag{4}$$

To calculate association measures, we submit queries to a search engine and use the number of page hits returned from the search engine as probability estimation. From the above association measures, the search results from the search engine for association measures are obtained by submitting both A and B for $N(A \wedge B)$, either A or B for $N(A)$ and $N(B)$, and both A and $-B$ for $N(A \wedge \neg B)$. B is leaded by the minus sign to specifies that B is an unwanted word.

However, computing conviction requires some assumptions since we cannot submit a query to obtain the web pages not containing B for finding $N(\neg B)$. Anyway, since the number of pages available on the Web is very large and B is a specific technical terms that are rarely found on the Web, we can assume that $\frac{N(\neg B)}{N(*)}$ is very closed to 1. Therefore, the approximated conviction can be computed as

$$Conv^*(A \rightarrow B) = \frac{N(A)}{N(A \wedge \neg B)} \tag{5}$$

As stated above, the association measures are of two types: directed and undirected measure. A directed measure evaluates the relationship of $A \rightarrow B$, (i.e. A causes occurrence of B), differently from the relationship $B \rightarrow A$. In contrast, the undirected association measures do not take into account the direction of occurrence. It measures both relationships in the same manner. Since we suppose that the relation between terms in the translation and synonym pairs are undirected relation, the directed association measure such as conviction and confidence need to be translated into an undirected representation. At this step, three functions; Minimum, Maximum and Mean are proposed to combine two directed measures to be an undirected one.

$$\lambda^{Min}(A \leftrightarrow B) = Min(\lambda(A \rightarrow B), \lambda(B \rightarrow A)) \tag{6}$$

$$\lambda^{Max}(A \leftrightarrow B) = Max(\lambda(A \rightarrow B), \lambda(B \rightarrow A)) \tag{7}$$

$$\lambda^{Mean}(A \leftrightarrow B) = Mean(\lambda(A \rightarrow B), \lambda(B \rightarrow A)) \tag{8}$$

where λ represents an directed association measure which is confidence or conviction, λ^{Min}, λ^{Max} and λ^{Mean} denotes an undirected association measure generated by Minimum function, Maximum function and Mean function respectively.

As an extension, association measures can be applied to find potential synonym pairs. However, a pair of synonyms are rarely occurred together on the same web page because a writer usually selects only one term to express each meaning in a sentence. Therefore, we cannot directly apply the search results to compute the measure for a pair of synonyms. However, with slight application, we can combine the association measures for term translation pairs in order to obtain a synonym pair.

In this paper, how likely two terms are a synonym pair is determined by considering the minimum association measures obtained from their translation intermediates. For example, while the arc labeled as 'TSC-1' in Figure 1(b) presents the example of a Thai-Thai synonym pair, T_0 and T_2, their association value is determined by the minimum value among the association values of all intermediate translation pairs, (T_0, E_0), (E_0, T_1), (T_1, E_1) and (E_1, T_2).

Iterative candidate generation repeatedly extracts the translated term from the product of the previous iteration. In any iteration when the association value is low, resulting

in translated terms weaken the linkage in the next iteration. We infer that the association of synonym pairs is low when their intermediate linkage is weak. For this reason the minimum function is used for the association of the synonym pair $\mu_s(T_1, T_n)$ as defined below.

$$\mu_s(T_1, T_n) = \min_{i \in \{1 \ldots n-1\}} (\mu_t(T_i, T_{i+1})) \tag{9}$$

where T_1 and T_n represent a synonym pair. T_i and $T_i + 1$ denote an intermediate translation pair appearing between T_1 and T_n. μ_s represents an association measure used for evaluating each synonym pair. μ_t represents an association measure used for evaluating each translation pair.

4 Experiments

4.1 Experimental Settings

A number of experiments have been conducted to confirm the performance of the proposed approach. The prototype system is implemented based on the APIs provided by Yahoo Developer Center[6]. The APIs allows us to access and obtain XML results from the Yahoo search engine. An initial set of 510 Thai medical terms manually collected from various web pages are used as the initial set of terms for extracting translation. A set of English translation terms are generated from snippets returned from Yahoo APIs. The process is done repeatedly for five times as shown in Figure 1(a). We perform three experiments to solve three questions as follows.

In the first experiment, the performance of the three functions, namely Min, Max and Mean, used for combining two directed measures to be an undirected measure are compared to select the best one. The training set includes all translation pairs from iterative candidate generation. For each undirected confidence and undirected conviction, its score is ranked in the descending order and compared to another. Here, top-k word pairs are evaluated where k is varied from 1 to the number of all possible word pairs in order to evaluate which function is the best.

In the second experiment, we evaluate translation and synonym pairs obtained by four association measures i.e. support, confidence, lift and conviction. As directed association measures, confidence and conviction require a mechanism to combine two functions that are the result of the first experiment to be the undirected association measures. For each association measure, the top-k translation and synonym word pairs are displayed in the descending order when the result is output. The experiments are conducted to evaluate three levels of generated candidates for translation and two levels of

[6] `http://developer.yahoo.com`

generated candidates for synonym as shown in Figure 1(b). The results are compared with the results labeled by three human evaluators. When the evaluators give different labels on one word pair, majority voting is selected.

In the third experiment, the performance of four association measures is tested with an unseen test set. The ten-fold cross validation is applied. The translation and synonym candidate pairs are equally divided into ten parts. Nine parts stand as a training set and one remaining part is used as a test set with equal distribution among positive and negative examples. The test are conducted repeatedly for ten times. In each time, the test set is changed to another part that never tested before. For each association measure, the value of an association measure that yields the highest f-measure in the top-k training set will be used as the threshold in the test set. For each translation and synonym pair in test set, we have

$$C_{(x)} = \begin{cases} C^+ & \text{if } V_{(x)} \geq \delta, \\ C^- & \text{if } V_{(x)} < \delta. \end{cases}$$

where $C_{(x)}$ stands for the assigned class for a pair x in the test set, C^+ denotes the positive class, C^- denotes the negative class, $V_{(x)}$ represents the association value of the pair x in the test set, δ stands as a threshold in the test set. The performance of our proposed system is evaluated with false positive, precision, recall and f-measure. Here, let S be the set of generated word pairs, and C be the set of correct word pairs. We have

$$FalsePositive = \frac{|S - C|}{|S|} \tag{10}$$

$$Precision = \frac{|S \cap C|}{|S|} \tag{11}$$

$$Recall = \frac{|S \cap C|}{|C|} \tag{12}$$

$$F{-}measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{13}$$

4.2 Experimental Results

Table 1(a) shows the result of candidate generation with the number of input terms, the number of extracted terms, the ratio of extracted terms over input terms. It is found that candidate generation from Thai to English has branching factor of approximately 2 but approximately 0.75 in English to Thai. At the end of all iterations, we got 2,321 Thai words and 5,953 English words in total. Table 1(b) shows the precision of the extracted words is decreased in the later iteration.

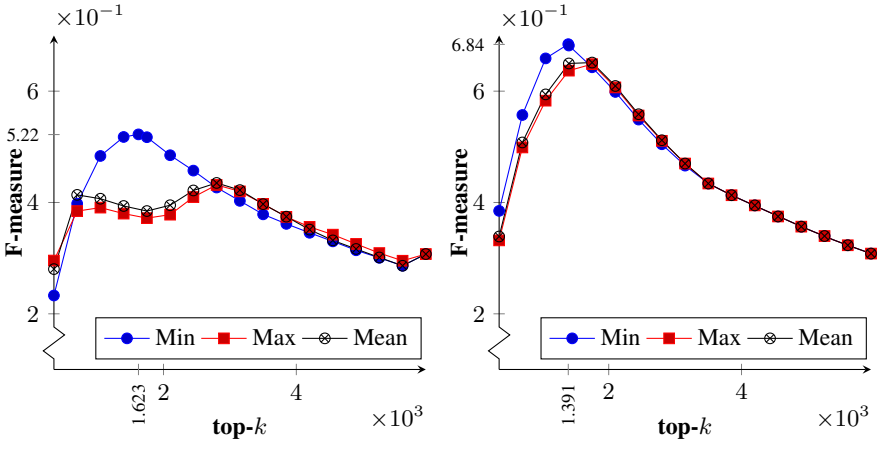
Table 1. Basic characteristic of candidate generation and baseline precision of the translation and synonym pairs

(a) Candidate generation result

Iteration	Input terms	Extracted terms	Ratio(Extracted/Input)
level 0 ($T \rightarrow E$)	510	1129	2.214
level 1 ($E \rightarrow T$)	1129	991	0.878
level 2 ($T \rightarrow E$)	991	2024	2.042
level 3 ($E \rightarrow T$)	2024	1330	0.657
level 4 ($T \rightarrow E$)	1330	2800	2.105

(b) Baseline precision

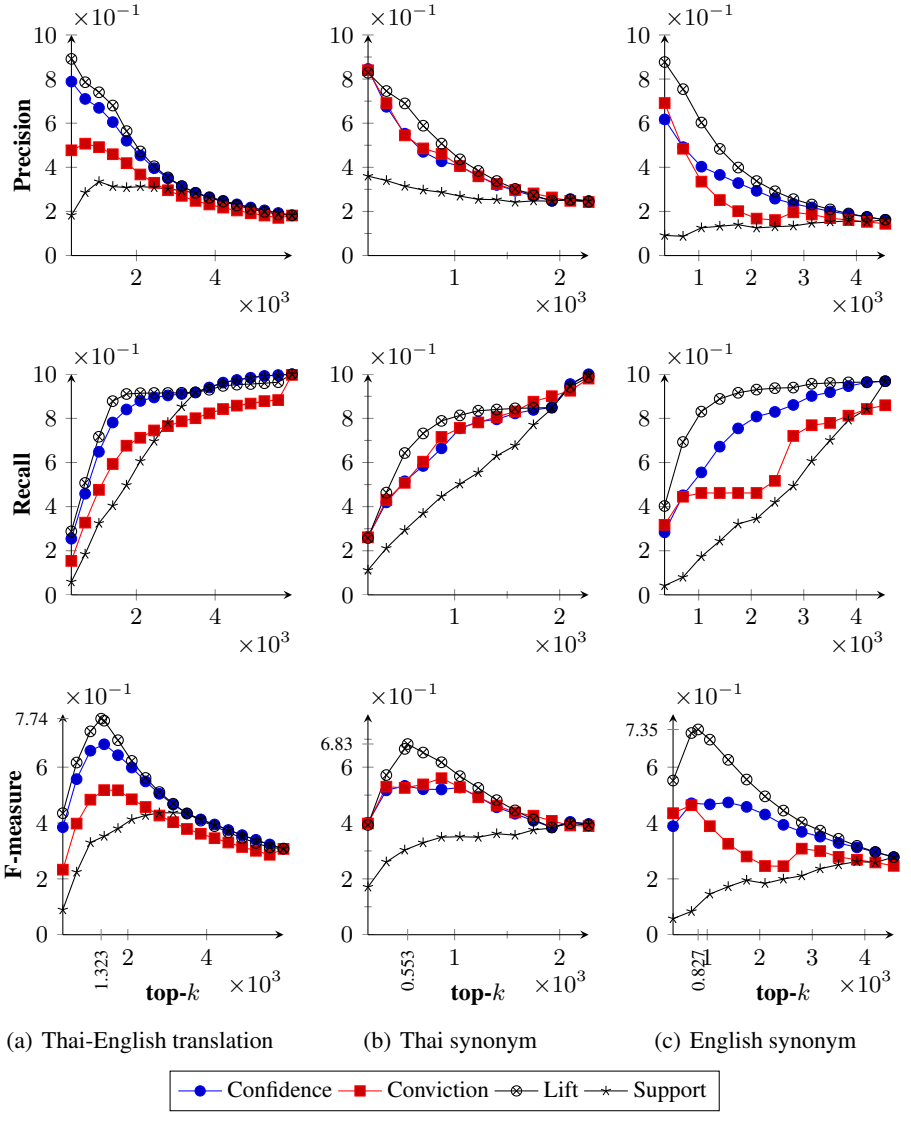
Evaluation level	Baseline Precision(%)		
	Translation	Thai synonym	English synonym
Level 0	63.77	35.62	24.95
Level 1	11.96	6.39	0.42
Level 2	0.57	-	-
Total	16.43	18.87	12.40



(a) Undirected conviction (b) Undirected confidence

Fig. 2. F-measures of the top-k translation pairs in descending order (undirected confidence and undirected conviction). Three conditions considered are minimum, maximum and mean combining function.

In the first experiment, Figure 2 shows the f-measure of the top-k translation pairs of undirected confidence and undirected conviction that are combined by minimum, maximum and mean functions as shown in Section 3.2. The minimum function yielded



(a) Thai-English translation (b) Thai synonym (c) English synonym

Fig. 3. Precision, Recall and F-measure for Thai-English Translation, Thai Synonym and English Synonym. Four association measures are Confidence, Conviction, Lift, Support.

the highest 68.44% f-measure in confidence and the highest 52.23% f-measure in conviction. Hence, we use the minimum function to generate undirected measures in the second and third experiment.

The second experiment investigates the precision, recall and f-measure score of top-k ranked by each association measures using all candidates as the training set and test

set. Figure 3(a), 3(b) and 3(c) show that lift yields the best association measure with the highest 77.4% f-measure (with 70.3% precision and 85.9% recall) for Thai-English translation, the highest 68.3% f-measure (with 68.9% precision and 67.7% recall) for Thai synonym and the highest 73.4% f-measure (with 70.6% precision and 76.6% recall) in English synonym, respectively.

Figure 3(a) shows the precision, recall and f-measure score of top-k Thai-English translation pairs that rank by each association measures using all evaluation levels as training and test set. The lift$\times$confidence achieved the best association measure with the highest 78.7% f-measure (with 70.8% precision and 83.5% recall) at top-1279. The precision of lift$\times$confidence is approximately four times greater than baseline precision as shown in Table 1(b).

Figure 3(b) shows the precision, recall and f-measure score of top-k Thai synonym pairs that rank by each association measures in all evaluation levels. The highest 68.3% f-measure (with 68.9% precision and 67.7% recall) at top-553 obtained by lift. Its precision outperforms baseline precision as shown in Table 1(b) with three and a half times improvement.

For English synonym pairs, figure 3(c) shows the precision, recall and f-measure score of top-k English synonym pairs that rank by each association measures in all evaluation levels. Lift yielded the highest 73.5% f-measure at top-827 (with 70.6% precision and 76.6% recall) which is superior to baseline precision as shown in Table 1(b) with approximately five and a half times improvement.

As the result of the third experiment, Table 2 shows the average false positive rate, average precision, average recall and average f-measure of ten-fold cross validation (subscripted, with '10') when our proposed algorithm is applied. The last column of the table is the f-measure of our proposed algorithm using all candidates as the training test and test set (subscripted with 'all'). The difference between both f-measures is trivial. This means the stability of the proposed system.

Table 2(a) shows lift yields the highest 73.1% average f-measure (with 8.3% average false positive, 67.0% average precision and 84.2% average recall) for Thai-English translation. Its average precision is superior to baseline precision (16.43%) as shown in Table 1(b) with approximately 4 times improvement. Table 2(b) shows that lift yields the highest 68.7% average f-measure (with 9.9% average false positive, 71.5% average precision and 67.7% average recall) for Thai synonym. Its average precision outperforms baseline precision (18.87%) as shown in Table 1(b) with 3.5 times improvement. Table 2(c) shows that lift yields the highest 72.8% f-measure (with 6.0% false positive, the 72.0% precision and 75.1% recall) for English synonym. The average precision of lift is approximately 5.5 times greater than baseline precision (12.40%) as shown in Table 1(b).

Table 2. Experimental Result in Ten-fold Cross Validation

(a) Thai-English Translation

Model	False Positive$_{10}$	Precision$_{10}$	Recall$_{10}$	F-measure$_{10}$	F-measure$_{all}$
Lift	0.083	0.670	0.842	**0.731**	**0.774**
Confidence	0.112	0.599	0.777	0.654	0.684
Conviction	0.230	0.398	0.703	0.493	0.522
Support	0.438	0.296	0.831	0.423	0.442

(b) Thai synonym

Model	False Positive$_{10}$	Precision$_{10}$	Recall$_{10}$	F-measure$_{10}$	F-measure$_{all}$
Lift	0.099	0.715	0.677	**0.687**	**0.683**
Confidence	0.155	0.544	0.537	0.531	0.548
Conviction	0.461	0.396	0.712	0.464	0.561
Support	1.000	0.243	1.000	0.390	0.406

(c) English synonym

Model	False Positive$_{10}$	Precision$_{10}$	Recall$_{10}$	F-measure$_{10}$	F-measure$_{all}$
Lift	0.060	0.720	0.751	**0.728**	**0.735**
Confidence	0.095	0.531	0.493	0.487	0.498
Conviction	0.089	0.515	0.446	0.453	0.477
Support	0.924	0.164	0.966	0.280	0.281

Concludingly, the experimental results evidenced that our system with lift gained the highest f-measure (73.1%) for mining Thai-English translation pairs, compared to the extraction of English synonym pairs (72.8%) and Thai synonym pairs (68.7%). This result is quite intuitive since naturally it is necessary to extract at least two translation pairs to obtain a synonym pair. Moreover, comparing to mining of English synonyms, extracting Thai synonyms is a harder task since Thai language has no explicit word boundary.

5 Conclusion

This paper presented a method to use Web documents as resources for extracting translation and synonym pairs between English and Thai medical terms. Iteratively inputting keywords on a search engine, a set of translation candidate pairs are generated. The potential scores of translation word pairs are calculated using four alternative measures, support, confidence, lift and conviction, commonly used in association rule mining.

By experiments using 510 Thai words, we found out that our approach using lift as association measure achieves the highest average f-measure of ten-fold cross valida-

tion that is 73.1% (with 67% precision and 84.2% recall) for Thai-English translation, 68.7% (with 71.5% precision and 67.7% recall) for Thai synonym and 72.8% (with 72% precision and 75.1% recall) for English synonym. The precision of our approach in Thai-English translation, Thai synonym and English synonym are 4 times, 3.5 times and 5.5 times greater than baseline precision respectively. Lift is the best association measure for extracting both translation and synonym. We also found that the minimum function is the best function for combining two directed measure to be an undirected measure.

As our future work, we plan to improve our approach to using combination of association measures with larger data sets and different specific domains.

Acknowledgment

This research is financially supported by the Thailand Research Fund (TRF), under Grant No. BRG50800013 and NCRT Grant 2009 and Thailand Advanced Institute of Science and Technology - Tokyo Institute of Technology (TAIST-Tokyo Tech), National Science and Technology Development Agency (NSTDA), Tokyo Institute of Technology (Tokyo Tech) and Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU).

References

1. Bodenreider., O.: Lexical, terminological, and ontological resources for biological text mining. In: Ananiadou, S., McNaught, J. (eds.) Text Mining for Biology and Biomedicine, ch. 3, pp. 43–66. Artech House (2006)
2. Zhang, Y., Vines, P.: Using the web for automated translation extraction in cross-language information retrieval. In: Proceedings of the 27th Annual International ACM SIGIR Conference (SIGIR 2004), Sheffield, South Yorkshire, UK, July 2004, pp. 162–169 (2004)
3. Viriyayudhakorn, K., Theeramunkong, T., Nattee, C.: Mining translation pairs for thai-english medical terms. In: Proceedings of the 3rd International Conference on Knowledge, Information and Creativity Support Systems (KICSS 2008), December 2008, pp. 104–111. Hanoi National University of Education (HNUE), Hanoi (2008)
4. Wang, J.-H., Teng, J.-W., Cheng, P.-J., Lu, W.-H., Chien, L.-F.: Translating unknown cross-lingual queries in digital libraries using a web-based approach. In: Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries (JCDL 2004), Tucson, Arizona, USA, June 2004, pp. 108–116 (2004)
5. Lu, W.-H., Lin, S.-J., Chan, Y.-C., Chen, K.-H.: Semi-automatic construction of the chinese-english MeSH using web-based term translation method. In: Proceedings of American Medical Informatics Association 2005 Symposium, pp. 475–479 (2005)
6. Wang, J.-H., Teng, J.-W., Lu, W.-H., Chien, L.-F.: Exploiting the web as the multilingual corpus for unknown query translation. J. Am. Soc. Inf. Sci. Technol. 57(5), 660–670 (2006)

7. Turney, P.D.: Mining the web for synonyms: Pmi-ir versus lsa on toefl. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 491–502. Springer, Heidelberg (2001)
8. Inkpen, D.: A statistical model for near-synonym choice. ACM Trans. Speech Lang. Process. 4(1), 2 (2007)
9. Okamoto, H., Sato, K., Saito, H.: Preferential presentation of japanese near-synonyms using definition statements. In: Proceedings of the second international workshop on Paraphrasing, vol. 16, pp. 17–24 (2003)
10. Shimohata, M., Sumita, E.: Acquiring synonyms from monolingual comparable texts. In: Dale, R., Wong, K.-F., Su, J., Kwong, O.Y. (eds.) IJCNLP 2005. LNCS (LNAI), vol. 3651, p. 233. Springer, Heidelberg (2005)