Thai Monitor Corpus:
Challenges and Contribution to Thai NLP

Wirote Aroonmanakun[1]

Nattawut Nupairoj[2]

Veera Muangsing[2]

Songphan Choemprayong[3]

Abstract

Building a corpus has been a necessary task for NLP and other research fields like linguistics, language teaching, and translation. Only a few Thai corpora have been created and released. Most of them are static and small in size. They are not designed to be a monitor corpus, which can grow over time. The concept of a monitor corpus bears similarity to the new research area named Big Data, which has gained more interests in the past few years because of the extensive growth of data available online. In this paper, the differences between monitor corpus and Big Data will be first discussed. Then, the design and the framework for developing a Thai monitor corpus will be outlined. To carry out this task, techniques and methods used in Big Data research that are suitable for storing texts will be selected and summarized. The progress of this work will be reported in section 3, and the plan for further development and the use of TMC will be sketched. The paper is concluded by pointing out the relationship between the two research fields, NLP and Big Data. Contributions to each other will be reviewed.

Keywords : Thai corpus, monitor corpus, NLP, Big Data

[1] Department of Linguistics, Faculty of Arts, Chulalongkorn University

[2] Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University

[3] Department of Library Science, Faculty of Arts, Chulalongkorn University

บทคัดย่อ

การสร้างคลังข้อมูลภาษาเป็นงานที่จำเป็นสำหรับงานด้านการประมวลผลภาษาและงานวิจัยด้านอื่น ๆ ได้แก่ ภาษาศาสตร์ การสอนภาษา และการแปล ในปัจจุบัน คลังข้อมูลภาษาไทยที่สร้างและเผยแพร่ให้ใช้มีจำนวนไม่มาก และเกือบทั้งหมดเป็นคลังข้อมูลที่เล็กและมีขนาดจำกัด ไม่ได้ถูกออกแบบมาให้เป็นคลังข้อมูลแบบสั่งสมที่มีขนาดใหญ่ขึ้นเรื่อย ๆ ได้ แนวคิดเรื่องการสร้างคลังข้อมูลแบบสั่งสมนี้สอดคล้องกับงานวิจัยเกิดใหม่ด้านข้อมูลใหญ่หรือบิ๊กดาต้า ซึ่งเป็นงานที่ได้รับความสนใจอย่างมากในปัจจุบันเนื่องด้วยจำนวนข้อมูลที่มีเพิ่มมากมายมหาศาลในโลกออนไลน์ ในบทความนี้ จะเริ่มด้วยการกล่าวถึงความแตกต่างของคลังข้อมูลสั่งสมกับข้อมูลใหญ่ จากนั้นจะกล่าวถึงการออกแบบและสร้างคลังข้อมูลสั่งสมภาษาไทย ซึ่งต้องอาศัยเทคนิคและวิธีการที่ใช้กันในงานข้อมูลใหญ่ ตอนที่สามจะรายงานความก้าวหน้าของการสร้างคลังข้อมูลสั่งสมภาษาไทย และแผนการพัฒนาและใช้ประโยชน์จากคลังข้อมูลดังกล่าว ในตอนท้าย จะสรุปถึงความสัมพันธ์และเกื้อกูลกันระหว่างงานวิจัยสองด้าน คือ งานการประมวลผลภาษาและงานข้อมูลใหญ่

คำสำคัญ : คลังข้อมูลภาษาไทย, คลังข้อมูลแบบสั่งสม, การประมวลผลภาษา, ข้อมูลใหญ่

1. Corpus and Big Data

The rapid growth of data available online recently has created a major interest in a new research field, namely Big Data. The term “Big Data” has been defined differently, but there are four characteristics that have been commonly accepted. These are volume, velocity, variety, and veracity. Volume indicates the large volume of data that has been increasing recently. Velocity is concerned with the rate of data generated and processed. Variety is about the complexity of data that could be in various forms, i.e. numerical, texts, audios, images, videos. The last one is about the trustworthiness of data. (IBM Big Data & Analytics Hub, 2014)

For NLP researchers, the term that is used for referring to language data is “corpus”. Corpora are generally used for training and testing an NLP system. The term is also widely recognized in other fields, such as linguistics, translation and language teaching. Various kinds of corpora have been created and used in many areas of studies. Most of the corpora are static and limited in size. A few corpora are dynamic and increasing in size. They are called “monitor

corpus". It is generally designed to monitor changes of a language, which is very important for lexicography work. COCA[4] is an example of a monitor corpus. It continues to grow over time.

In general, the corpus used in an NLP work is not considered large compared to the size of data in Big Data research. Even when the data can be increased over time in the case of a monitor corpus, data in a corpus is still fundamentally different from Big Data. The data in Big Data can have various forms, e.g. text, numerical data, images, sounds, videos, etc., while data in a corpus is mainly text or language data. However, as many NLP systems began to process language data generated from users in the internet, e.g. twitters, web boards, blogs, reviews, etc., the data starts to be enormously increasing in size and the result from language processing is expected to be generated in real time. In doing this, it is necessary to use technologies related to Big Data to process the data. For example, Hadoop is a framework in which distributed storage is implemented as Hadoop Distributed File System, and parallel processing is done by MapReduce (Hedlund 2011). NoSQL is a non-relational database that is more suitable to store data with flexible structure in each record. MongoDB is an example of NoSQL database that is widely used.

As for Thai language data, the largest Thai corpus, to the best of our knowledge, is HSE Thai corpus. It is about 50 million words collected mostly from news websites. The data can be searched online or downloaded from the project website[5]. Another large corpus is Thai National Corpus (Aroonmanakun 2007). The size of TNC currently is 33 million words. Since It is designed to be a general corpus like the British National Corpus, various text types and genres are included. Other Thai corpora are small and a specialized corpus, For example, Orchid corpus (Sornlertlamvanich et al. 1997) is a collection of academic papers, which are word segmented and part-of-speech tagged. Its size is about 350,000 words. Thai-NEST corpus (Theeramunkong et al. 2010) is a Thai named entity tagged corpus. It created by collecting texts from 10,000 news articles. Another well known corpus is BEST corpus (Boriboon et al. 2009). It is manually word segmented and crated as a data set for word segmentation contest. Its size is about 5 million words. All of these corpora are all static corpora. In this paper, we proposed to do an experiment in creating a Thai monitor corpus, which will continue to grow in

[4] http://corpus.byu.edu/coca/

[5] http://web-corpora.net/ThaiCorpus/search/

size. Thus, the basic technology for storing and processing Big Data will be used and implemented as a fundamental setting of the Thai monitor corpus.

2. Design of Thai Monitor Corpus

Thai Monitor Corpus (TMC) is designed with the primary goal to be constructed quickly with a large amount of Thai texts. By doing this, only a few text sources will be selected because extracting texts from each source would require a tailor-made method of extraction. As a result, a variety of text types or genres would be limited at this stage. However, a few criteria for selecting text sources are used to ensure that TMC has texts in various domains and widely used. The criteria for selecting text sources are as follows. First, the source must be publicly available on the internet and its popularity can be measured from the ranking reported in truehits.net. Each source should represent different text types. Data in the source should be mainly texts, in which Thai texts can be systematically crawled and extracted. Lastly, each source should have texts categorized into various subject domains.

From these criteria, we came up with the following sources: pantip.com, bloggang.com, twitter.com, www.khaosod.co.th.[6] The first is the most popular web board in Thai. It contains various domains of discussions. The language is somewhat in a spoken and informal style, in which a topic is posted and the responses and comments can be added by the public. All of the texts and meta-data marking text status was extracted and preserved. The second source is the most popular site for bloggers. It is arranged into sub-groups for different domains. Each blog is like an article in which comments can be added at the end. The language is a kind of informal written style. Again, all of the texts and meta-data were extracted and kept in the corpus. The third source is the twitter. Though Facebook is more popular in Thailand than twitter, collecting Facebook data is not straight forward as collecting twitter data. Thus, twitter is chosen to be a representative of social media data. The language is limited in size due to the constraint of twitter. Each tweet has no more than 140 characters. But some of them have information in terms of geo-location. Though twitter data is seen publicly, some of them are written to specific persons. The language in twitter then is varied in styles and domains. It can be a personal message or a public message, a written language or a spoken language, a

[6] Selection is based on statistical data reported in truehits.net on January 1, 2015

formal or informal style. The last data source is the news reports from Khaosod, which is the most visited news website. The language from this source is a formal and written language. The texts can be categorized into different section representing different domains of texts. However, to avoid any political bias that might occur in collecting only one news source, we select another news website which shares the same group of audiences but bear different political views, Khomchadluek. The data collected and its characteristics is summarized in Table 1.

Table 1: Types of data selected for TMC

Sources	Genres	Styles	Recipients	Domains
pantip.com	web board	informal, spoken	participants, public	a wide variety, categorized into chat rooms
bloggang.com	blog	informal*, written*	Public	a wide variety, categorized into groups
twitter.com	micro blog	informal*, spoken*	specific, public	a wide variety
www.khaosod.co.th	news report	formal, written	Public	a wide variety, categorized into sections
www.komchadluek.net	news report	formal, written	Public	a wide variety, categorized into sections
* data are mostly like that				

3. Framework of the system

In this study, TMC is created as a pilot project of Thai monitor corpus. The system is consisted of four parts: collecting data, storing and processing data, analyzing data, and hosting the corpus, which is shown in Figure 1

.

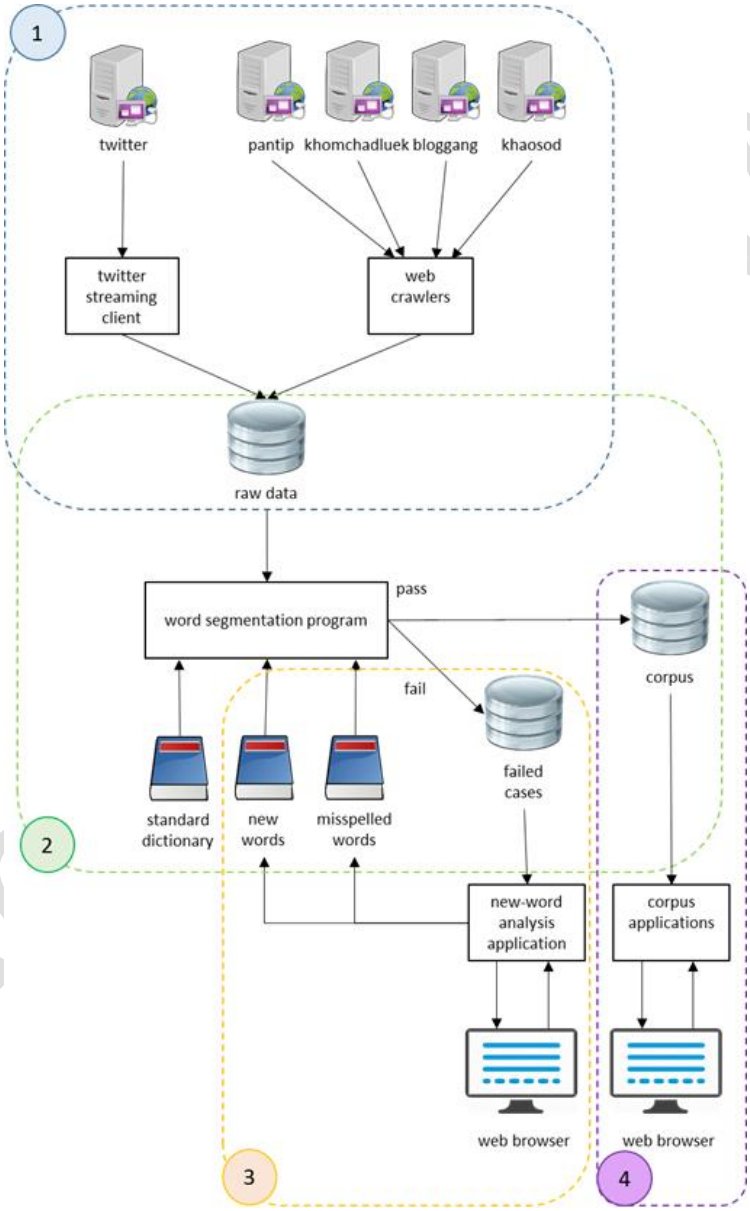

Figure 1: Framework of TMC

In the first part, tweets are collected by the use of twitter API, and the other data sources are collected by specific web crawlers written to match the structure found in each source. Only tweets written in Thai texts are collected by using two methods. The first one is to collect

tweets originated from geo-location set for Thailand. The second method is to collect tweets containing the top four hundred of Thai words found in twitter data collected from previous research. All information of the tweet is stored as it is in a json format. About 200,000 tweets are saved daily.

For other web data sources, we have to manually analyze the web structure to understand how Thai texts are stored on the web page, as well as other meta data that is relevant to the page. Then, a web crawler program is written for each web site to extract Thai texts and meta data. The extracted data is also stored in a json format like the twitter data. All data is then stored in an NoSQL database by using MongoDB. We chose MongoDB because it is easy to convert data from json to bson and it can be used with MapReduce in Hadoop system.

After the data is extracted and stored in bson format, the field that contains Thai texts will be word segmented. Thai word segmentation program written in Perl (Aroonmanakun 2002, Aroonmanakun and Rivepiboon 2004) is used to segment Thai texts because it is designed to recognized well-formed of written Thai. Only words found in the dictionary and written correctly according to Thai orthography rules will be successfully segmented. A sentence containing any unknown words, which can be either a proper name, a transliterated word, a reduplication, a spoonerism, an abbreviation, a shortening, a derivation, a slang, or a coinage, will be marked as a failed segment. It is interesting that a lot of data especially when it is collected from a source containing informal language like twitter cannot be word-segmented successfully. This lead to the next part, the analysis of the data.

Data analysis at this stage focuses on segments that cannot be word-segmented by the program. As stated earlier, failure in word segmentation can be caused from many factors. This is a crucial problem for any Thai NLP systems designed to be used on real data especially data generated from social medias. For any Thai NLP system to work with the real data, all these problems have to be resolved. This issue will be further discussed in Section 5.

The final part is about the use of TMC. Some example of uses will be demonstrated and implemented as a service API, which should show the potential use of TMC. The use of TMC and the future plan will be outlined in detail in section 6.

4. Progress report

The data collected for the prototype system of TMC is about 1,173 million words. The largest portion of data is collected from pantip.com, while the smallest portion is the data from the news websites. The number of words from each source and the time period of collected data are reported in Table 2

Sources	No. of words (millions)	period
pantip.com	1,000	Jan 2014 – Sep 2016
bloggang.com	16	NA
twitter.com	150	Apr 2016 - Dec 2016
www.khaosod.co.th & www.komchadluek.net	7	Feb 2015 – July 2016

Table 2 : Amount and period of collected data

5. Analysis of errors

This section focuses on data that cannot be word-segmented by the program. As stated earlier, these failed segments are caused by either spelling errors or unknown words. Unknown words are those that are not yet included in the standard dictionary used by the segmentation program. They could be abbreviations (ผช is ผู้ชาย /$p^hu{:}3.c^ha{:}j1$/), shortened words (โอ is โอเค /$ʔo{:}1.k^he{:}1$/), proper names, dialect words, interjections, foreign words transliterated into Thai (ท็อป is “top”), reduplication of words (ดี๊ดี /$di{:}4.di{:}1$/), lengthened sound (ค่าาาาาาา /$k^ha{:}3$/), spoonerisms (อะหรี่ดอย /$ʔa2.ri{:}2.dɔ{:}j1$/ is อร่อยดี /$ʔa2.rɔ{:}j2\ di{:}1$/), new words or slangs (จุงเบย /$cuŋ1\ bɤ{:}j1$/ is จังเลย /$caŋ1\ lɤ{:}j1$/), or words written in different variations (แย้ว is แล้ว).

In order to see the proportions of segmentation errors caused from these various sources, we implement a web-based tool for marking types of errors found in the failed segment, as shown in Figure 2. The program will randomly show a chunk of text that could not be word-

segmented. A failed segment is shown in red color. Thai informants were asked to mark the error types and specify the correction. Error types and the intended words will be recorded in a database.

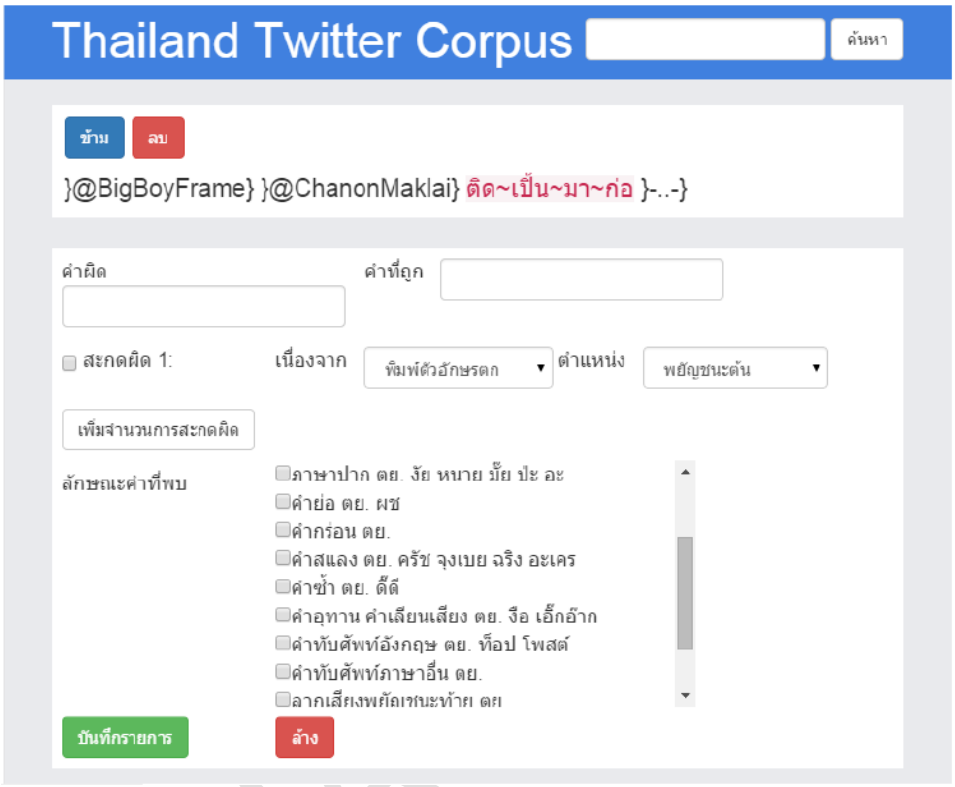

Figure 2 : Tools for marking segmentation errors

To solve problems of segmentation errors, various modules need to be implemented e.g. spelling correction, abbreviation detection, named entity recognition, backward transliteration, shortened word detection, and reduplication detection. These would be basic tools for processing Thai language in general. However, it is interesting that when processing informal texts from internet e.g. twitters and web boards, a lot of slangs and word variants are found in these informal texts. We compare formal texts and informal texts found in TMC by sampling data from twitters. Tweets collected from news agency sources represent formal text while tweets from other sources represent informal texts. It is found that the number of both spelling errors and word variants are found more often in the informal texts. Errors on formal tweets are found only 2% of the data, while errors on informal tweets are up to 17%.

In this section, we will focus only on word variants, which are found more often in informal texts. This is a major problem when processing texts from social media. Since variants

of the same word can be created in many forms, a generation module should be able to generate more than one variant using different methods of variation. It is found that a variation is usually made by sound variant, in which an initial consonant, a vowel, a final consonant sound or a tone is changed to related sound. In a sound variant, orthography will be adjusted to reflect the new sound. The following are some examples of sound variants.

แล้ว /lɛːw4/ > แย้ว /jɛːw4/ change initial consonant sound from /l/ to /j/

ถูกต้อง /tʰuːk2 tɔːŋ1/ > ถ้วกต้อง /tʰuːak3 tɔːŋ3/ change vowel sound from /uː/ to /uːa/

เลย /lɤːj1/ > เล้ย / lɤːj4/ change tone from level /1/ to high tone /4/

A variant could also be created by spelling variant, in which different characters are used but the pronunciation remains the same, e.g. สัตว์ /sat2/ > สัส /sat2/, โทรศัพท์ /tʰoː1.ra4.sap2/ > โทรสับ /tʰoː1.ra4.sap2/. Usually a variant bears similarity to the original word so that listeners can recall the original word. But in some cases, more than one variation can be found in a variant. For example, อะไร /ʔa2.raj1/ > อัลไล /ʔan1laj1/ (0>n, 2>1; r>l), ตลก /ta2.lok2/ > ตัลล๊าก /tan1.laːk4/ (0>n; o>aː, 2>4)

6. The use of TMC and the future plan

TMC is a prototype of a monitor corpus, which will be increasing in size. At this prototype stage, it is a corpus containing specific languages in both formal and informal styles. Since it is not designed as a general corpus, it does not directly reflect the use of Thai language in general. Anyway, it can reflect the use of Thai language in specific. It also can reflect changes of language. To demonstrate the potential use of TMC, we have created API services for TMC. These basic APIs include the ability to search for a word, frequency of the word, and the context in which it occurs, as shown in Figure 3 and Figure 4.

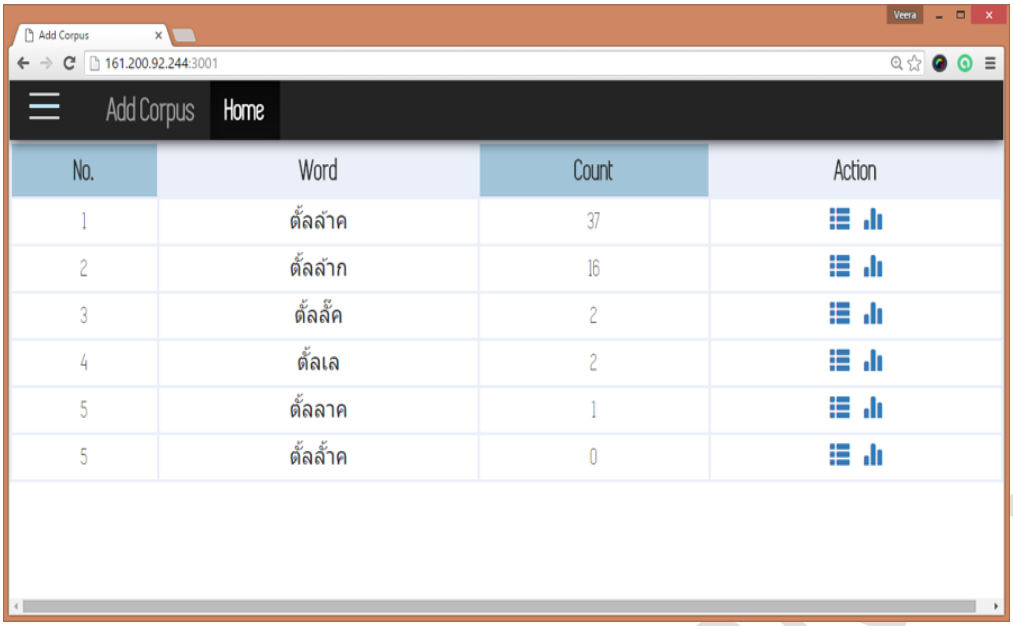

Figure 3 : Word list and frequency result from Corpus API

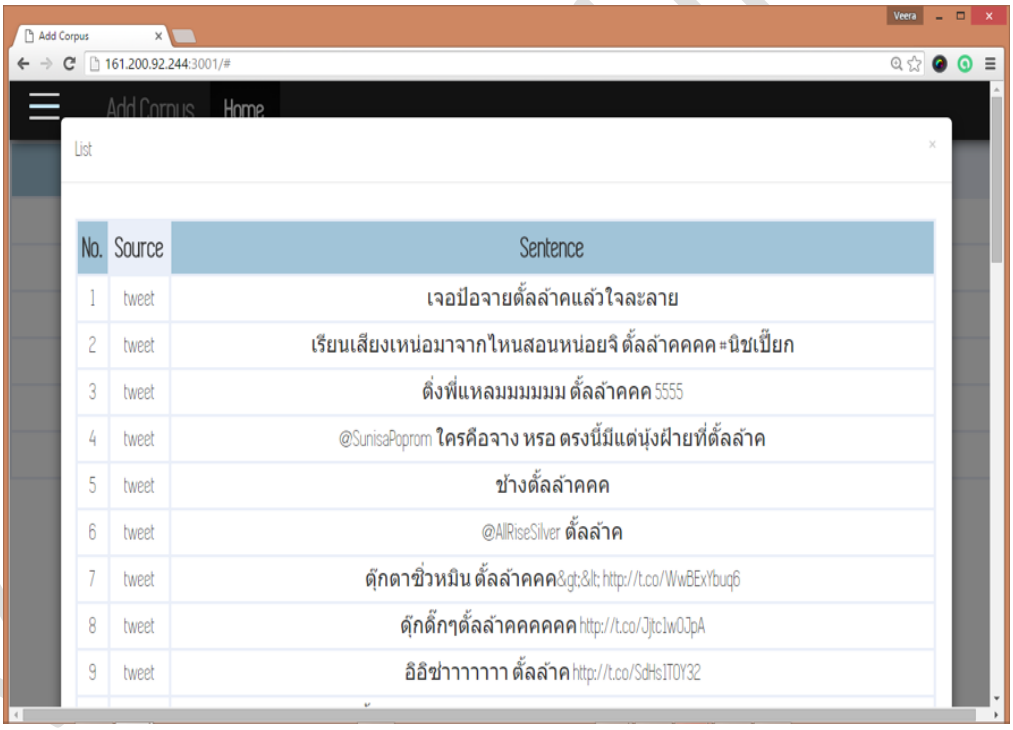

Figure 4 : Words in context result from Corpus API

In addition to these basic information, it is not difficult to create a visualization showing a network of a word and its variants, changes of word frequency over a given period, distribution of words plotted on a map, comparisons of words in different text types. However, the main problem is the implementation of these tools to work on big data and the ability to get the results in real time. For example, to do word segmentation on the data about 7 million words, we spent

59 hours of service on Google Cloud platform with Compute Engine Intel N1 1 VCPUs 3.75 GBs. With the increasing number of data and the more complexity of preprocessing tasks, i.e. part-of-speech tagging, named entity recognition, forward and backward transliteration, dependency parsing, etc., the ability to process a large amount of text would be a crucial problem. Needless to say about the indexing process, if any, to enhance to ability to search words in the corpus. This will be the next focus of our experiment.

7. Conclusion and the future of Thai NLP

In this study, we have created a prototype of Thai monitor corpus by collecting data from four major sources, i.e. twitter, web board, blog, and news. To create a prototype corpus with a large amount of data, a large number of texts are collected from each source as much as possible. This is to demonstrate how varieties of text data can be collected and included in a monitor corpus. To store and process a lot of data, technology related to Big Data research is used in this study. The corpus is preprocessed by segmenting words and stored in bson format in a MongoDB. The database is stored and processed on a Hadoop framework.

Working on real data generated from social media like twitters suggests that ell-formed data is commonly found. To be able to process real informal data, a lot of modules would be needed, e.g. named entity recognition, forward and backward transliteration, abbreviation detection, shorten word detection, reduplication and character repetition, etc. Detection of word variants is also a major concern for informal texts. Since variations are found both for sounds and letters, a word can have more than one variant. All of these preprocessing modules as well as standard modules like POS tagging and syntactic parsing would impose a lot of processing time for a large amount of data in TMC.

In general, when training corpus is increasing in size, the language model would be better. Big Data contributes to NLP not only in terms of the amount of available data but also the challenges to NLP tasks. The main concern is to employ an NLP system that works on very large data volume in real time. Recent research starts to explore more on this direction. For example, Wang et al. (2012) proposed a system to do sentiment analysis from twitter in a real time using IBM's InfoSphere Streams platform. Nesi et al. (2015) used Hadoop as a distributed system for natural language processing when dealing with very large data like web pages. Agerri et al. (2015), on the other hand, proposed a new distributed architecture for NLP, in

which Storm is used instead of MapReduce. By using streaming architecture, existing NLP modules can be used in a distributed system without re-implementing them. Plale (2013) discussed the relations between Big data and other related fields like information retrieval, text mining, and NLP. Han et al. (2015) talked about the data mining methods for NLP and the new research direction that integrates both NLP and data mining. In return, the fruitful of NLP results would yields a better data to be discovered for Big Data research. In sum, we are going to see more or more crossing research between NLP and Bid Data, especially when data is concerned with language data.

References

Agerri, R., Artola, X., Beloki, Z., Rigau G., & Soroa, A. (2015). Big data for Natural Language Processing: A streaming approach. *Knowledge-Based Systems*. 79.36-42.

Aroonmanakun, W. (2002). *Collocation and Thai Word Segmentation*, In Sornlertlamvanich, T.T.a.V. (Ed.), Proceedings of the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Worksho*p*. Sirindhorn International Institute of Technology, Pathumthani, pp. 68-75.

Aroonmanakun, W. (2007). Creating the Thai National Corpus. *Manusaya* Special Issue No.13, 4-17.

Aroonmanakun, W. & Rivepiboon, W. (2004). *A Unified Model of Thai Word Segmentation and Romanization,* In Proceedings of The 18th Pacific Asia Conference on Language, Information and Computation, Tokyo, Japan, pp. 205 - 214.

Boriboon, M., Kriengket, K., Chootrakool, P., Phaholphinyo, S., Purodakananda, S., Thanakulwarapas, T., & Kosawat, K. (2009). *BEST Corpus Development and Analysis*. In Proceedings of International Conference on Asian Language Processing, IALP '09.

Han, J., Ji, H., & Sun, Y. (2015). *Successful Data Mining Methods for NLP*. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 1-4. Beijing, China: Association for Computational Linguistics.

Hedlund, Brad. (2011). *Understanding Hadoop Clusters and the Network*. Available from http://bradhedlund.com/2011/09/10/understanding-hadoop-clusters-and-the-network/ [Accessed August 21, 2016]

IBM Big Data & Analytics Hub. (2014). *The Four V's of Big Data.* Available from http://www.ibmbigdatahub.com/infographic/four-vs-big-data [Accessed 1 Febuary 2016]

Nesi, P., Pantaleo G., & Sanesi, G. (2015). A hadoop based platform for natural language processing of web pages and documents. *Journal of Visual Languages & Computing* 31, Part B.130-38.

Plale, B., (2013). *Big data opportunities and challenges for IR, text mining and NLP*, In Proceedings of the 2013 international workshop on Mining unstructured big data using natural language processing. ACM, San Francisco, California, USA, pp. 1-2.

Sornlertlamvanich, V, Charoenporn, T., & Isahara, H. (1997). *ORCHID: Thai Part-Of-Speech Tagged Corpus. National Electronics and Computer Technology Center Technical Report*, 5-19

Theeramunkong, T., Boriboon, M., Haruechaiyasak, C., Kittiphattanabawon, N., Kosawat, K., Onsuwan, C., Siriwat, I., Suwanapong, T., & Tongtep,N. (2010). *THAI-NEST: A framework for Thai named entity tagging specification and tools*. In Proceedings of the 2nd International Conference on Corpus Linguistics (CILC10), May 13-15, 2010, University of A Coruña, Spain.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). *A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle*. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 115-120. Jeju Island, Korea: Association for Computational Linguistics.

Acknowledgement

This article is part of a research project titled, "Creating a Thai Monitor Corpus and the Search for Orthographic Errors and Coinages" funded by the National Research Council of Thailand (NRCT) and the Thailand Research Fund (TRF), grant no. RDG58H0001. The content and views are those of the authors and do not necessarily reflect the views of NRCT and TRF.