19th International Computer Science and Engineering Conference (ICSEC)
Chiang Mai, Thailand, 23-26 November, 2015

Unsupervised Construction of a Word List on Tourism From Wikipedia

Dittaya Wanvarie
Department of Mathematics and Computer Science
Faculty of Science
Chulalongkorn University
Email: Dittaya.W@chula.ac.th

Sansanee Ek-atchariya
Chinese Section
Department of Eastern Languages
Faculty of Arts
Chulalongkorn University
Email: Sansanee.E@chula.ac.th

Thanakon Kaewwipat
German Section
Department of Western Languages
Faculty of Arts
Chulalongkorn University
Email: th.kaewwipat@yahoo.com

***Abstract*—The demand for word lists in a specialized domain is increasing in language learning. We propose an unsupervised framework to extract a word list from Wikipedia data for a language learning class specialized on tourism. We extract topics in Wikipedia articles using non-negative matrix factorization. Each topic is classified as tourism related or not using articles in WikiVoyage. We choose paragraphs in Wikipedia that are classified as in-domain and rank words in such paragraphs by their frequencies. The proposed framework retrieves more than 90% of words in the gold list, but the extracted list still includes a large number of general terms.**

***Index Terms*—Natural language processing; Information filters; Data mining;**

I. Introduction

Tourism is one of the major service industries in South East ASEAN countries. The number of tourists visiting ASEAN countries continues increasing in recent years. More than a half of arrival tourists are coming from outside the region, especially China and Europe. One demand that also increases due to the tourism growth is language teaching for the tourism industry.

One of the fundamental resources in language learning and teaching is word lists. On one hand, teachers can utilize word lists to construct learning materials. On the other hand, the coverage of a word list is also an important factor in the successful communication of a learner since the successful learning depends on the vocabulary size a learner acquires.

A general word list contains vocabularies from broad areas. However, a general list may contain few words in a specialized domain. For example, tourism-related vocabularies include geographical, cultural and religion topics. These topics are normally beyond the scope of general language learning and may not be included in a normal textbook. Tourism-related vocabularies are also unique to the geographical area. For instance, tropical island and temple are common in South East Asian area but occur less frequent in the European context.

To construct a word list for a particular domain, one should collect materials in the target domain and rank words with their frequencies. If the collected data size is large enough, the list of frequent words is reliable as a word list for the domain. However, collecting domain-specific data manually is not trivial since it requires experts time and consideration. At present, computer-readable data is widely used and easily accessed on the Internet. One of the clean and large corpora available on the Internet is Wikipedia, which is an open encyclopedia written by users. Wikipedia also has a strong convention of formal writing styles that conform to language learning. Moreover, Wikipedia has millions of articles in several languages. Therefore, Wikipedia may be a good and cheap source to start gathering the data. One of the problems in using Wikipedia data is that its articles are not categorized. Selecting articles in the target topic domain may not be trivial.

In this paper, we propose an automatic and unsupervised framework to extract a Chinese and a German word list from Wikipedia articles. Our domain is the sightseeing information in ASEAN countries. We have WikiVoyage articles that mostly contains tourism-related documents but the collection size is limited. Instead of directly extracting a word list from the small set of WikiVoyage articles, we propose to classify Wikipedia paragraphs using information extracted from WikiVoyage articles. Finally, the word list is constructed from the tourism-related paragraphs in Wikipedia articles. The number of filtered articles in Wikipedia is still greatly larger than the number of WikiVoyage articles. Hence, the word list will be more reliable.

The organization of this paper is as follows. Section II contains several pieces of the related work. We discuss some current word lists in Chinese and German in Section II-A. The outline of non-negative matrix factorization which is adopted in our work, is described in Section II-B. Our word list construction approach is explained in Section III. We evaluate the proposed framework in Section IV and discuss the result in Section V. Finally, we conclude our contribution and discussing the future work in Section VI.

II. Related Work

A. Word Lists

Word list construction is mostly based on frequency count of words in the given corpus. Hence, selecting articles to build

978-1-4799-4963-2/14/$31.00

a corpus plays a significant role in the construction. Normally, in-domain articles are chosen by linguists or experts in each domain.

Chinese HSK Word list [1] is a list of 5,000 words in Chinese Language Proficiency Test. The test is comparable to the Test Of English as Foreign Language (TOEFL). DeReWo list [2] is a general word list in German consisting of common 320,000 German words. Both HSK and DeReWo lists do not provide frequency information.

SUBTLEX series are word lists extracted from films and television series subtitles. The series contain lists in Dutch, American English, Chinese, Spanish, German, Greek, British English, and Polish. We choose SUBTLEX-CH [3] for Chinese and SUBTLEX-DE [4] for German as a part of the general list in this paper.

Wikipedia also extracted a word list from its articles in some languages. The most frequent 2,000 German words are provided in Wiktionary [5]. Note that these words are not domain specific. Wikipedia does not have its Chinese word list. Instead, it provides the 10,000 most frequent words of Mandarin Chinese from Academia Sinica [6] in both the traditional Chinese and the simplified Chinese variations.

B. Non-Negative Matrix Factorization (NMF)

Given a non-negative matrix $\mathbf{V}_{D\times N}$, non-negative matrix factorization or NMF decomposes $\mathbf{V}$ into two non-negative matrices $\mathbf{W}_{D\times T}$ and $\mathbf{H}_{T\times N}$ such that it minimizes the reconstruction error

$$f(\mathbf{W},\mathbf{H}) = \frac{1}{2}||\mathbf{V}-\mathbf{WH}||_{\mathbf{F}}^2. \quad (1)$$

Note that there may be no exact solution of the factorization. The implementations of the algorithm usually employ numerical methods and find an approximate solution.

Non-negative matrix factorization is a successful unsupervised feature extraction or dimensionality reduction method in document clustering [7] and classification [8], [9]. In text processing tasks, the document-by-word matrix $\mathbf{V}_{D\times N}$ is a corpus consisted of document vector $\mathbf{v}_i$. The dimension D is the number of documents in the corpus and N is the number of unique words in the corpus.

Normally, N is substantially large and a single word contains few useful information. Instead of using the word-level representation, we can model a document using a broader concept called topics. Suppose that a document is focused on tourism, in-domain words such as "beaches" and "temples" are more likely to occur in the document. In contrast, off-topic words such as "computer" and "star" are less frequent.

From the minimization of (1), the factorization will decompose the given $\mathbf{V}_{D\times N}$ into two matrices $\mathbf{W}_{D\times T}$ and $\mathbf{H}_{T\times N}$. $\mathbf{W}$ is the document-by-topic matrix and $\mathbf{H}$ is the topic-by-word matrix, respectively. Each score $v_{i,j}$ in the input matrix $\mathbf{V}_{D\times N}$ can be any non-negative score that represents the relativeness of word j in document i such as a simple or weighted frequency, or a term frequency-inverse document frequency (tf-idf) score.

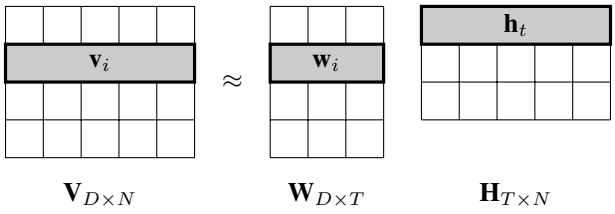


Fig. 1. Illustration of NMF and topic modeling

T in this task is the number of topics in the corpus. Each $h_{t,j}$ in $\mathbf{H}_{T\times N}$ is a score of word j in topic t. If $h_{t,j}$ is high, it indicates that word j is closely related to topic t. Each $w_{i,t}$ in $\mathbf{W}_{D\times T}$ is a score of document i being in topic t. Similarly, if $w_{i,t}$ is high, document i should be highly related to topic t. Both $\mathbf{V}$ and $\mathbf{W}$ represent the same document in the corpus but with different aspects. $\mathbf{V}$ employs the word-level representation while $\mathbf{W}$ employs the topic-level representation.

The illustration of topic extraction using NMF is in Fig. 1. Row $\mathbf{v}_i$ and $\mathbf{w}_i$ are the same document in the collection with different representation. Row $\mathbf{h}_t$ is the word distribution score of topic t. Each $\mathbf{h}_t$ is invariant to the document but is specified to a topic. Hence, given a matrix $\mathbf{v}_i$, we can transform $\mathbf{v}_i$, whose dimension is N, to $\mathbf{w}_i$, whose dimension is T. Since T is usually much smaller than N, the dimension is then reduced. The representation of a document is also transformed from the combination of words to the combination of topics embedded in the document.

III. Word List Construction

Our goal is to construct a list of words on tourism domain for the language learning purpose. Since a word list is usually constructed from the most frequent list of words, we need to gather only documents related to tourism domain and extract words from the collection. Apart from the document selection, words that are necessary for our task are only content words. We also need to choose only words that fit our criteria.

We divide our task into two steps, the document selection in Section III-A and the word list construction in Section III-B.

A. Document Selection

Wikipedia does not have hierarchical structures of its articles. Instead, each article is mostly independent of others. Although we can find topics from all Wikipedia articles, the topics will be very broad and cannot suitably represent our target domain. Moreover, the computational cost on extracting topics from the entire Wikipedia would be very expensive since there are millions of articles.

Firstly, we propose to filter articles with simple keywords. Since our target domain is tourism in ASEAN countries, a list of ASEAN country names is served as a simple primary filtering criterion. All articles that contain any of the ASEAN country names will be included in the primary collection.

The country name alone is not sufficient to extract tourism data. For example, an article on a television program obviously contains the country name of its origin but the article itself

may not relate to the tourism domain. Thus, We need more fine grain filtering criteria.

An article is usually long and consists of several sections and paragraphs. Normally, a paragraph contains one main idea thus one topic. Instead of further filtering documents at an article level, we propose to choose paragraphs that are related to tourism topics using non-negative matrix factorization (NMF) described in Section II-B. From now on, we treat a paragraph as a piece of document. The corpus becomes a collection of paragraphs instead of a set of articles.

Several preprocessing steps are necessary before constructing the paragraph-by-word matrix. We need to perform word segmentation for languages without explicit boundaries such as Chinese. We also remove stop words form paragraphs since these words are too general and are clueless to any topic. Topics also depend mostly on root words. However, variations of words due to morpheme insertion in word construction will decrease the frequency count of words. In this paper, we employ the Snowball stemming algorithm [10], if necessary, and use stemmed words to find topics. After segmentation, removing stop words, and stemming, a paragraph vector $\mathbf{v}$ is constructed using tf-idf metrics. Hence, our paragraph-by-word matrix $\mathbf{V}$ is a tf-idf matrix whose element is the term frequency-inverse document frequency of word j in paragraph i.

From Section II-B, we can factorize $\mathbf{V}$ into two non-negative matrix $\mathbf{W}$ and $\mathbf{H}$. $\mathbf{V}$ is a paragraph-by-word matrix. $\mathbf{W}$ is a paragraph-by-topic matrix, and $\mathbf{H}$ is a topic-by-word matrix, respectively. Given $\mathbf{V}$, we can approximate the factorization of $\mathbf{V}$ with $\mathbf{W}$ and $\mathbf{H}$ through the minimization of (1) as follows;

$$\mathbf{V} \approx \mathbf{WH}. \tag{2}$$

From (2), each row of $\mathbf{H}$ is a topic that either relates or does not relate to the tourism domain. We can transform any paragraph $\mathbf{v}$ in $\mathbf{V}$ into $\mathbf{w}$ in $\mathbf{W}$ such that each $w_{i,t}$ is the score of paragraph i in topic t. If $w_{i,t}$ is high and t is a tourism-related topic, paragraph i should also be related to the tourism domain.

Firstly, we extract topics from selected Wikipedia paragraphs. Secondly, we will classify each topics using information from WikiVoyage paragraphs which mostly contains tourism related data. All WikiVoyage paragraphs are converted into tf-idf paragraph-by-word matrix $\mathbf{V}_{voyage}$ using the vectorizer constructed from Wikipedia. $\mathbf{W}_{voyage}$ is then approximated using $\mathbf{H}$ from the factorization of $\mathbf{V}$ in (2) and $\mathbf{V}_{voyage}$, i.e.

$$\mathbf{W}_{voyage} \approx \mathbf{V}_{voyage}\mathbf{H}^{-1} \tag{3}$$

Note that $\mathbf{H}^{-1}$ in (3) is the inverse matrix of $\mathbf{H}$ in (2).

Each $w_{i,t}$ in $\mathbf{W}_{voyage}$ indicates the relativeness of WikiVoyage paragraph i, to topic t extracted from Wikipedia. We can find the most related topic, $tmax_i$, for WikiVoyage paragraph i with the following equation;

$$tmax_i = \arg\max_t w_{i,t} \quad ; \text{for } i \in D_{voyage}.$$

i represents the index of each paragraph in WikiVoyage collection, D_{voyage}. We count the frequency of each $tmax_i$ for every paragraph in WikiVoyage collection. Top 25% of the most related topics are listed as our target topics, T_{target}. Finally, we re-evaluate each Wikipedia paragraph by finding its most related topic. Any paragraph i in Wikipedia collection whose most related topic is in T_{target} are selected. The filtered documents are in D_{target} where

$$D_{target} = \{\text{paragraph } i \text{ in } D \text{ if } \arg\max_t w_{i,t} \in T_{target}\}. \tag{4}$$

Note that paragraph i is from the filtered Wikipedia collection whose representation is either $\mathbf{v}_i$ or $\mathbf{w}_i$.

TABLE I
STATISTICS OF DATASETS

Chinese			
Statistics	GoldCH	WikiCH	WikiVoyageCH
Document counts	925	544,220	593
Word counts	288,235	16,296,905	79,102
Unique word counts	28,785	905,444	16,807
German			
Statistics	GoldDE	WikiDE	WikiVoyageDE
Document counts	2,089	1,337,948	30,615
Word counts	284,902	25,767,286	423,825
Unique word counts	35,640	1,628,800	48,930

B. Word Selection

After we obtain D_{target} from (4) in Section III-A, we rank each word by its frequency in D_{target}. Since we need only content words, we will count only nouns, verbs, and adjectives. We tag each paragraph in the selected set using Stanford part-of-speech (POS) tagger [11], [12]. Note that each paragraph i in the dataset is from the original text. The paragraph still includes stop words, and their words are not stemmed.

We count the same word with different part-of-speech together and ranked them by their frequency counts in descending order. We finally prune all words with frequencies less than 30.

IV. EXPERIMENTS AND RESULT

A. Datasets and Experiment Settings

We evaluate the proposed framework on two languages, Chinese and German. The target domain is geographic, architectural and cultural information for tourism in ASEAN countries. We have a gold data set for each language, extracted from guidebooks. An expert in each language reviews the books and chooses paragraphs that are in-domain manually. Note that not every content in the book is in-domain. Paragraphs on history, politics and shopping information are excluded from the gold datasets. The gold word list is extracted with the method described in Section III-B.

Chinese gold collection set contains 925 documents with approximately 280k words. The Gold word list, GoldCH, extracted from this set contains 894 words. German gold collection contains 2,089 documents with approximately 280k words. The Gold word list, GoldDE, extracted from this set

TABLE II
WORD LIST SIZE

List name	Chinese	German
Gold	894	748
HSK	5,000	-
DeReWo	-	326,655
SUBTLEX-30	22,625	23,790
WikiList	10,000	2,000
NMF-20	13,026	19,749
NMF-50	10,363	14,679
NMF-100	9,022	18,577

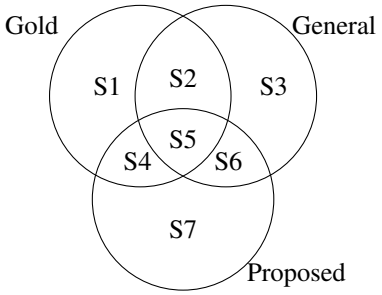


Fig. 2. Word Distribution Reference

contains 748 words. Some documents in the gold set contain several paragraphs instead of a single paragraph. Since we do not need to extract topics from the gold set, we do not split the document into paragraphs. We also measure the size of each gold set by its word counts instead of document counts.

We will extract word lists from Wikipedia dump in Chinese [13] and German [14]. The topics are extracted from WikiVoyage dump in Chinese [15] and German [16], respectively.

The characteristics of each dataset are summarized in Table I. Chinese articles have both traditional Chinese and simplified Chinese character variations. Wikipedia has a character map for conversion between the traditional Chinese words and the simplified Chinese words [17]. We use the same character map to normalize all extracted Wikipedia and WikiVoyage paragraphs to use simplified Chinese words.

We also have several general word lists for each language. We have HSK, SUBTLEX-CH, and the list from Wiktionary for Chinese. Similarly, we have DeReWo, SUBTLEX-DE, and Wiktionary list for German. We merge all general lists into a single general set for each language.

For SUBTLEX lists, we choose only words with the frequency more than 30. The result pruned list is SUBTLEX-30. SUBTLEX-CH also provides part-of-speech and corresponding frequency counts of a word in each part-of-speech. We choose only words that are nouns, verbs, adjectives, and adverbs. However, SUBTLEX-DE does not contain any part-of-speech information. Therefore, we include all words in the general set.

The number of words in each list is shown in the middle rows of Table II. We also compare the extracted word lists with the general set to analyze the coverage of the extracted lists in a broad domain.

B. Experiment Settings

We evaluate the proposed framework in Section III with Wikipedia and WikiVoyage data. The NMF implementation adopted in this experiment is from Scikit-learn package [18]. The number of topics for topic extraction is set to 20, 50 and 100, respectively. All other parameters are default values.

Top 25% of related topics in T_{target} are selected. Finally, the size of the extracted word list in each setting is shown in the bottom rows of Table II. NMF-20, NMF-50, and NMF-100 represent the proposed word list from NMF with 20-, 50-, and 100-topic settings, respectively.

We compare the result word list with the gold set and the general set to evaluate our construction framework. We divide words in all lists into seven categories according to the diagram in Fig. 2. We also have two standard metrics, precision and recall, to evaluate our extracted word list.

Let U be the union operator of two set and $|S|$ be the number of words in set S. The precision metric is the percentage of words in the gold list found in our extracted list;

$$Precision = \frac{|S4 \cup S5|}{|S4 \cup S5 \cup S6 \cup S7|}. \quad (5)$$

The recall metric or the retrieval rate is the percentage of extracted gold words found in the gold list;

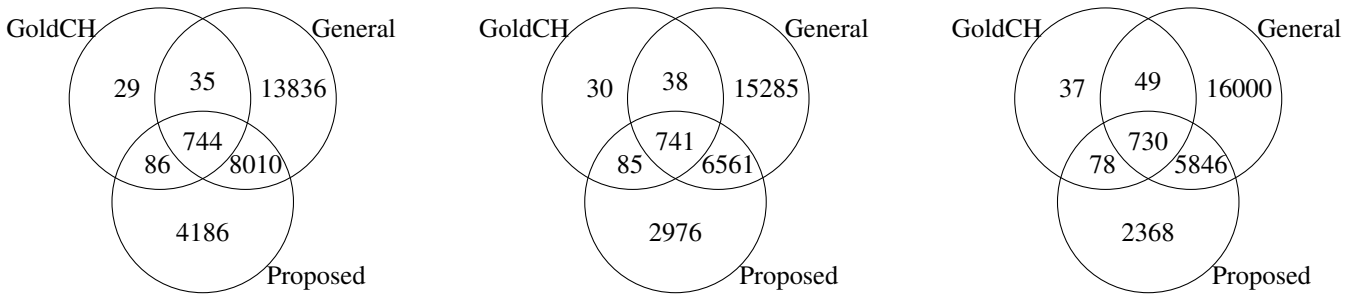
$$Recall = \frac{|S4 \cup S5|}{|S1 \cup S2 \cup S4 \cup S5|}. \quad (6)$$

V. RESULT AND DICUSSION

We will discuss the result from experiments in Chinese and German separately in the following subsections.

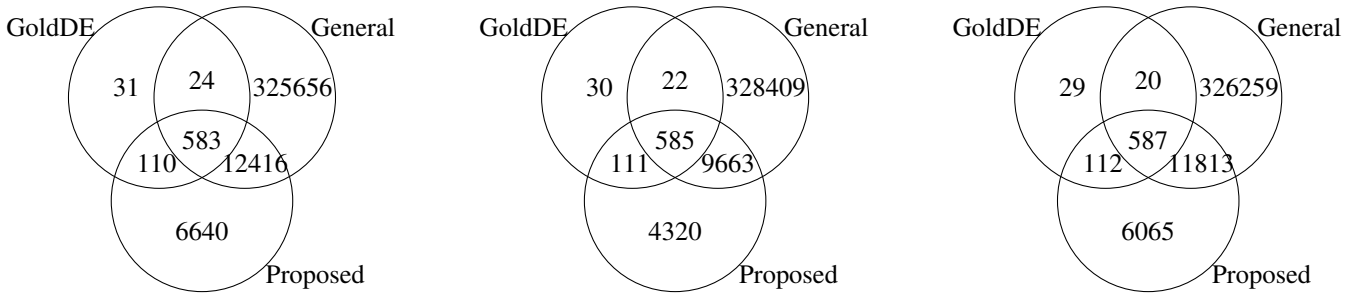
TABLE III
EXAMPLES OF FREQUENT WORDS IN CHINESE NMF-20 LIST

	Chinese
S1	曼德勒 (Mandalay), 佛殿 (Vihara), 石灰岩 (limestone), 巴厘岛 (Bali island), 矗立 (stands)
S2	美景 (beautiful scenery), 迷人 (charming), 山洞 (cave), 徒步 (on foot), 壮观 (spectacular), 预订 (booking), 清澈 (clean/clear), 艺品 (works of art)
S3	谢谢 (thanks), 是的 (yes), 所有 (all), 抱歉 (sorry), 看看 (take a look),
S4	佛塔 (pagoda), 佛像 (Buddha statue), 寺 (temple), 佛教 (Buddhism), 而 (and), 其中 (among), 因此 (therefore), 但是 (still) 之一 (one)
S5	是 (be), 有 (have/possess), 上 (above/up), 最 (most), 中 (center/middle), 不 (not), 好 (good), 会 (can/ability/possibility), 中国 (China), 为 (to)
S6	知道 (know), 没 (not), 眞 (true), 事 (thing), 别 (do not), 香港 (Hong Kong), 航空 (aviation), 亦 (also), 系统 (system), 表示 (representation)
S7	猪笼草 (Nepenthes), 气旋 (cyclone), 升格 (upgraded), 气象厅 (Meteorological Agency)



(a) NMF-50 (b) NMF-50 (c) NMF-100

Fig. 3. Word Distribution in Chinese List. Figures inside a circle are the number of unique words in the category.



(a) NMF-20 (b) NMF-50 (c) NMF-100

Fig. 4. Word Distribution in German List. Figures inside a circle are the number of unique words in the catebory.

TABLE IV
EXAMPLES OF FREQUENT WORDS IN GERMAN NMF-20 LIST

	German
S1	Ao (Bay), Jh. (century), Vieng (town), Cave (cave), Teluk
S2	Std. (hours), Sandstrand (sandy beach), Touren (tour), lohnt (worth), Mt., er (he), großes (great), schöner (beautiful), langer (long), Sonnenuntergang (sunset)
S3	Ich (I), nicht (not), Sie (they), Das (the), die (the)
S4	Ko (island), Hier (here), Hat (beach), Auch (also), Gunung, Philippinen (Philippines), So (so), Dort (there), Da (there)
S5	ist (is), sich (-self), sind (are), auch (too), werden (will), zu (to), hat (has), so (so), war (was)
S6	du (you), mit (with), mich (me), habe (have), bin (am), New, Juni (June), August, September, Januar (January)
S7	%,Amphoe (district), Tambon (subdistrict), B., US-Dollar

A. Chinese Experiments

From the distribution of words in Fig 3, the number of retrieved words from the gold set are quite similar in all settings. Therefore the recall rate is rather high, ranging from 90.38% to 92.84%. However, the precision is low, varying from 6.37% to 8.96%, due to thousands of extracted words that do not belong to the gold list. The setting with 20 topics achieves the highest precision, but the lowest recall among the three settings since its number of extracted words is the largest.

Among the extracted words in the list, about one-fourth to one-third of them are not included in both the gold list and the general list. These words are in category S7. Examples of words in this class are shown in row S7 of Table III. 猪笼草 (Nepenthes) is a local plant in South East Asia. 气旋 (cyclone), 升格 (upgraded) and 气象厅 (Meteorological Agency) are words related to the great storm in the past.

779 of 894 words in the gold set are already included in the general list as suggested in category S2 and S5. Words that are included in all lists are in category S5. Most of them are simple and common words with high frequency in any domain, e.g. 是 (be) and 有 (have, possess). Our framework extracts more than 90% of general words in the gold list. However, our methods cannot retrieve words in cluster S2. Some of words in this category such as 美景 (beautiful scenery) and 迷人 (charming) are general words that are modifiers for scenes and places. These words are scarcely used in encyclopedia articles.

Another interesting category is S4, whose words are found in both the gold list and our extracted list but are not general. There are about 80 words in this category depending on the experiment settings. Example words are shown in row S4 of Table III. This category mostly contains proper nouns for places and culture-related words such as 佛塔 (pagoda) and 佛像 (Buddha statue). However, a word ambiguously determined either as an adverb or a conjunction such as 而 (and), 其中 (among) and 因此 (therefore), are also in this category. Although these adverb/conjunction are very general, their part-of-speech tag in the general set are conjunction; hence they are not included in the general set.

B. German Experiments

The distribution of words in German lists is shown in Fig. 4. Among 748 words in the gold set, our proposed framework retrieves words back more than 90% in all settings. However, the precision rate is low similar to the result in Chinese, varying from only 3.51% to 4.71%. The reason is the number of general words extracted is rather high. Specifically, more than 65% of words in our list are also found in the general list. The three settings achieves comparable performance in terms of precision and recall.

We drew the most frequent words in each category of Fig. 2 and showed in Table IV. Some words in category S1 are proper names and names in foreign languages, e.g. *Ao* (Bay in Thai) and *Teluk* (Bay in Indonesian). However, several words of the same types are successfully extracted in category S4 and S7, e.g. *Hat* (beach in Thai), *Gunung* (mountain in Indonesian and Malay), *Amphoe* (district in Thai) and *US-Dollar*.

Category S2 contains some general words describing the scenery, e.g., *Sandstrand* (sandy beach), *Sonnenuntergang* (sunset). Similar to the result of the Chinese experiments, these words are not common in encyclopedia articles. Most frequent words in category S3 are pronouns such as *Ich* (I). Note that we did not include pronoun in the target word list. However, we cannot prune these words from the general list since there is no part-of-speech information provided in any of the general set. Frequent words in category S5 and S6 are mostly common verbs such as *ist* (is), *habe* (have).

C. Discussion

We found that words in category S2, especially adjectives and adverbs describing the scenery, are not retrieved in both the Chinese and German experiments. From our inspection, Wikipedia articles that have formal encyclopedia writing style will scarcely have such words. However, we successfully retrieve several domain specific words in category S4.

The previous work in word list construction counts the word without any normalization. We then decided to follow the previous convention by counting its original form. However, variations of words due to the grammatical rules in German need to be normalized in the topic extraction step. In this paper, we use a stemming algorithm to normalize these variations but it leads to over-normalizing side-effect. For example, *Hat* (beach in Thai) and *hat* (has) are different words but are normalized to the same word. One possible solution to this problem is to change the stemming algorithm to a more robust approach such as using a lemmatizer. We leave the task to the future work.

VI. Conclusion

We propose an unsupervised framework to extract a word list in tourism-related domain. The settings with the higher number of topics are likely to eliminate unrelated paragraphs with the trade-off of the recall rate. However, some of the extracted topics are still out-of-domain. We may utilize seeding topics to achieve more accurate topics of the interest.

There are also rooms for improvement in the topic extraction. As the document is a linear combination of topics, we can classify Wikipedia articles based on its combination instead of its single best-fit topic.

Acknowledgment

This research has been supported by the Ratchadaphiseksomphot Endowment Fund of Chulalongkorn University (CU-56-327-HS).

References

[1] "新汉语水平考试（HSK）词汇（2012年修订版） (New Chinese proficiency test (HSK) glossary (2012 Revision)." [Online]. Available: http://www.chinesetest.cn/userfiles/file/HSK/HSK-2012.xls

[2] *Korpusbasierte Wortgrundformenliste DEREWO, v-ww-bll-320000g-2012-12-31-1.0, mit Benutzerdokumentation*, Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, 2012. [Online]. Available: http://www.ids-mannheim.de/derewo

[3] Q. Cai and M. Brysbaert, "SUBTLEX-CH: Chinese Word and Character Frequencies Based on Film Subtitles," *PLoS ONE*, vol. 5, no. 6, p. e10729, 06 2010.

[4] M. Brysbaert, M. Buchmeier, M. Conrad, A. M. Jacobs, J. Bölte, and A. Böhl, "The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German," *Experimental Psychology*, vol. 58, pp. 412–424, 2011.

[5] "Top 2000 Most Commonly Used Words in German Wikipedia," February 2015. [Online]. Available: https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/top_2000_German_Wikipedia_words

[6] "Appendix:Mandarin Frequency lists," May 2013. [Online]. Available: https://en.wiktionary.org/wiki/Appendix:Mandarin_Frequency_lists

[7] W. Xu, X. Liu, and Y. Gong, "Document Clustering Based on Non-negative Matrix Factorization," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. New York, NY, USA: ACM, 2003, pp. 267–273.

[8] P. C. Barman and S.-Y. Lee, "Document Classification with Unsupervised Nonnegative Matrix Factorization and Supervised Percetron Learning," in *International Conference on Information Acquisition*, July 2007, pp. 182–186.

[9] M. W. Berry, N. Gillis, and F. Glineur, "Document classification using nonnegative matrix factorization and underapproximation," in *IEEE International Symposium on Circuits and Systems, 2009.*, May 2009, pp. 2782–2785.

[10] M. Porter, "German Stemming Algorithm," 2014, retrieved on 2014-10-10. [Online]. Available: http://snowball.tartarus.org/algorithms/german/stemmer.html

[11] R. Levy and C. Manning, "Is it harder to parse chinese, or the chinese treebank?" in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ser. ACL '03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 439–446.

[12] A. N. Rafferty and C. D. Manning, "Parsing three german treebanks: Lexicalized and unlexicalized baselines," in *Proceedings of the Workshop on Parsing German*. Stroudsburg, PA, USA: ACL, 2008, pp. 40–46. [Online]. Available: http://dl.acm.org/citation.cfm?id=1621401.1621407

[13] January 2015. [Online]. Available: http://dumps.wikimedia.org/zhwiki/20150109/

[14] "German Wikipedia dumps," January 2015. [Online]. Available: http://dumps.wikimedia.org/dewiki/20150106/

[15] "Chinese WikiVoyage dumps," January 2015. [Online]. Available: http://dumps.wikimedia.org/zhwikivoyage/20150123/

[16] "German WikiVoyage dumps," January 2015. [Online]. Available: http://dumps.wikimedia.org/dewikivoyage/20150123/

[17] "ZhConversion.php File Reference," May 2015. [Online]. Available: https://doc.wikimedia.org/mediawiki-core/master/php/ZhConversion_8php.html

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.