Utilizing Wikipedia for Retrieving Synonyms of Trade Security-related Technical Terms

Rafal Rzepka[1], Shinji Muraji[2] and Akihiko Obayashi[3]

[1]Faculty of Information Science and Technology, Hokkaido University, Japan
[2]Graduate School of Information Science and Technology, Hokkaido University, Japan
[3]Center for Innovation and Business Promotion, Hokkaido University, Japan
[1]rzepka@ist.hokudai.ac.jp, [2]shinjimuraji@ist.hokudai.ac.jp, [3]obayashi@mcip.hokudai.ac.jp

Abstract

Measuring semantic similarity of technical terms is not a trivial task especially for multi-word terminology. Matching term equivalents in documents for decision-making in fields like medicine or law requires high precision, therefore instead of using popular statistical methods, synonym lists are being manually created, which is the most certain but costly solution. Although many similarity-based methods have been proposed, we have not found any reports on how useful are human-made redirections included in Wikipedia pages. In this paper we report results of our experiments for discovering synonyms of terms related to trade security in Japanese language by using redirections and compare them with simple entity-linking approach. We perform a series of experiments using a synonym list prepared by the government specialists as the gold set. Strictness of the evaluation setup resulted in low scores, but we confirmed which heuristics have more potential than others. We discuss several findings which shed some light on how they could be utilized to solve the difficult task of extracting similar technical terms.

Keywords: Technical Terms, Wikipedia, Multiword Level Similarity, Trade Security, Export Control

1. Introduction

In recent years, trade security has gathered international attention not only on the level of governments or industries but also at universities. However, with some exceptions (Rzepka et al., 2021; Matsuzawa and Hayasaka, 2022) almost no NLP research deals with this topic. One of the biggest difficulties is the fact that except the legal jargon, the trade-security terminology covers various fields from nuclear physics through biology to engineering. Official vocabulary often differs from what the one that researchers use, and it becomes a problem when an artificial agent communicates with a user in order to navigate her or him to a proper passage of the regulatory text. Some terms can carry different weigh in terms of danger, therefore the usage of common similarity-based approaches becomes problematic. Hence, avoiding a loose fuzzy matching whenever possible can limit the number of agent's erroneous advices. In the case of Japanese language which does not use spaces, there is an additional problem of word segmentation which is not trivial with technical terms across several fields. When we tested a glossary list prepared by the CISTEC[1] organization (1,660 terms), Japanese morphological tool Janome has divided 72.5% (1,204/1,660) of all technical terms which are theoretically single semantic entities according to the glossary authors. On the other hand the Japanese government provides a list of official synonyms (or rather synonymical phrases) for no more than 6% of these terms. The research question that appears is "how could we automatically enrich such a list with a high confidence?" and in this research explores the possibility to utilize Wikipedia for discovering closely related terms. Our paper is structured as follows: in Section 2. we list some common approaches to discovering technical terms and measuring their similarities; in Section 3. we introduce eight approaches we tested; in Sections 4. and 5. we describe the experiments and their results. We discuss the results in Section 6. and conclude our paper with Section 7.

[1]https://www.cistec.or.jp/english/export/

2. Related Work

Technical terms have been approached in the field of natural language processing from different angles (Butters and Ciravegna, 2008). One obvious scenario is to discover them with their equivalents in other language to achieve higher quality machine translation. Bilingual corpora or datasets are utilized and matching the closest terms is the main target. For example, (Bollegala et al., 2015) propose prototype vector projection (PVP) which is a non-negative lower-dimensional vector projection method to help compiling large-scale bilingual dictionaries for technical domains. In the same paper they propose a method to learn a mapping between the feature spaces in the source and target language using partial least squares regression (PLSR) which requires only a small number of training instances to learn a cross-lingual similarity measure. The proposed method outperforms several other feature projection methods in biomedical term translation prediction tasks.

In the context of trade security, (Matsuzawa and Hayasaka, 2022) propose a method for associating technical documents and legal statements of export control in English and Japanese. Although the main goal is to find dissimilarities on the sentence level, the authors underline the importance

of proper matching technical terms which are crucial for avoiding misunderstanding of regulatory texts.

In their recent study, (Liwei, 2022) tackle the problem of technical term matching between English and Chinese patents. After testing many approaches including deep learning-based ones, they discovered that adapting C-value (Frantzi et al., 2000), a hybrid terminology extraction method combining linguistic rules and statistical theory, to specific domains, yields the best results. This newly proposed DC-value method combined with information entropy successfully extracted Chinese technical terms outperforming the original C-value method, the log-likelihood ratio method and the mutual information method (Church and Hanks, 1990).

Another relatively popular target is to find acronyms or abbreviations of technical terms. For example (Yagahara and Sato, 2020) automatically extract full forms from abbreviations by using word2vec for terminology expansion in the "image diagnosis"-related abstracts retrieved from PubMed. They determine the optimal word2vec parameters that ensure the highest accuracy, which was Skip-gram with 200 dimensions and 10 iterations achieving 74.3%. Although recognizing acronyms like "ldr" stating for "low dose rate" seem simple enough to use heuristics, errors like "bb" assigned to *biobreeding* instead of the correct *black blood*, underline the importance of context processing.

Simplifying the technical documents is another task where technical terms are important. The task is to identify them and replace with simple equivalents to make a document easier to comprehend for a layperson. (Abrahamsson et al., 2014) have improved an existing method for assessing difficulty of words in Swedish text. The difficulty of a word was assessed not only by measuring the frequency of the word in a general corpus, but also by measuring the frequency of substrings of words. By doing so they adapted the method to the compounding nature of Swedish, signaling that language specific approaches are important to develop bilingual thesauri.

Wikipedia is a valuable source for finding similar terms (Hwang et al., 2011). An early example of how to use inter-wiki links to extract named entities and rank synonyms is the work of (Bøhn and Nørvåg, 2010), who used, except heuristics, the frequencies of inter-wiki links which inspired our use of thresholds. More recent is an approach proposed by (Jagannatha et al., 2015), who use Wikipedia for automatic extraction of synonyms related to the biomedical domain. By using inter-wiki links, they extract the candidate synonyms (which are not technical terms) of an anchor-text in a Wikipedia page and the title of its corresponding linked page. They rank synonym candidates with word embedding and PRF (pseudo-relevance feedback). They found that PRF-based re-ranking outperforms word embedding based approach and a strong baseline using inter-wiki link frequency. Furthermore, their results showed that a hybrid method (namely Rank Score Combination), achieved the best results and upon this finding we also tested combinations of our implemented methods.

3. Tested Methods

3.1. Redirect-based Methods

In usual Wikipedia terminology, a redirect indicates a type of article that sends the reader to another article when there are different names for the same subject. For example, when "USA" is input in the search box, Wikipedia displays the page of "United States". In this research, we utilize redirects in a slightly different manner. Using the example of the "United States"-related page example, it contains a phrase "*scientific* force" as a linked string, and its link redirects to "Science and technology in the United States" page. It is not uncommon that the linked string and the title of the linked page are different, and the link in Wikipedia's HTML contains the title of the linked page[2]. For the purpose of this approach we assume that linked phrase and the title of the linked have similar meaning, and such pairs can form a thesaurus. However, this heuristic is not perfect due to ofter arbitrary way how the Wiki creators create such links. For instance, in the example above, only the word "scientific" is linked although the linked page is related to "scientific force". We assume that if a phrase is linked to a target page only once, there is a high probability that it is an unusual combination and it may cause noise. To confirm this hypothesis, we propose an additional method which collects pairs only if a phrase is linked two or more times to a give target title. We call these methods REDIRECTING and REDIRECTING WT (**W**ith **T**hreshold). To investigate the effectiveness of redirect's opposite direction, namely when the redirected page (here "Science and technology in the United States") sends back to the target word page ("United States" in our example), we add a pair of algorithms implementing this approach and name them REDIRECTED and REDIRECTED WT.

3.2. Inner-Link-based Method

In this method we use a Wikipedia page of the target word (if it exists), and assume that all linked words are related and probably synonymous. For example, in the Wiki page of "photodetector" (*hikari kenshutsu-ki*[3]), we can find inks to Japanese terms for "photomultiplier tube" or "solar cell". Similarly to the REDIRECTING and REDIRECTING WT methods, we add the same threshold and call the methods LINKING and LINKING WT, respectively. Furthermore, we also construct algorithms for checking the opposite direction ("photomultiplier tube" linking back to "photodetector") and call the additional methods LINKED and LINKED WT.

4. Experiment

In this section we explain how we tested the above-described heuristics by matching the results with expert-created thesaurus.

[2] `https://en.wikipedia.org/wiki/Science_and_technology_in_the_United_States` in this example

[3] `https://bit.ly/3j7ZBhP`

Approach	Precision	Recall	F-score
(1) REDIRECTING	0.0700 (14/200)	0.0927 (14/151)	0.0798
(2) REDIRECTING WT	0.1558 (12/77)	0.0795 (12/151)	**0.1053**
(3) REDIRECTED	0.2188 (7/32)	0.0464 (7/151)	0.0765
(4) REDIRECTED WT	0.2143 (3/14)	0.0199 (3/151)	0.0364
(1)+(3)	0.0806 (17/211)	0.1126 (17/151)	0.0939
(1)+(4)	0.0683 (14/205)	0.0927 (14/151)	0.0787
(2)+(3)	0.1868 (17/91)	0.1126 (17/151)	**0.1405**
(2)+(4)	0.1667 (14/84)	0.0927 (14/151)	0.1191

Table 1: Experimental results for the **redirect**-based approach and the combinations of its methods. Bold font is used for top F-scores in both single and hybrid approaches, highest **precision** scores are underlined. Numbers in brackets indicate number of terms matched with gold set / numbers of found synonyms (precision) and number of terms matched with gold set / number of all synonyms in the gold set (recall).

Approach	Precision	Recall	F-score
(5) LINKING	0.0033 (14/4305)	0.0927 (14/151)	0.0063
(6) LINKING WT	0.0040 (11/2720)	0.0728 (11/151)	0.0077
(7) LINKED	0.0033 (10/2995)	0.0662 (10/151)	0.0064
(8) LINKED WT	0.0103 (3/290)	0.0199 (3/151)	**0.0136**
(5)+(7)	0.0031 (14/4494)	0.0927 (14/151)	0.0060
(5)+(8)	0.0032 (14/4318)	0.0927 (14/151)	0.0063
(6)+(7)	0.0035 (11/3134)	0.0728 (11/151)	0.0067
(6)+(8)	0.0040 (11/2743)	0.0728 (11/151)	**0.0076**

Table 2: Experimental results for the **inner-links**-based approach and the combinations of its methods. Bold font is used for top F-scores in both single and hybrid approaches, highest **recall** scores are underlined. Numbers in brackets indicate number of terms matched with gold set / numbers of found synonyms (precision) and number of terms matched with gold set / number of all synonyms in the gold set (recall).

4.1. Data

Here we describe the data used for experiments – the source of links and the test dataset of synonyms.

4.1.1. Japanese Wikipedia

For the experiments we have downloaded latest dump of Japanese Wikipedia[4] with *wikiextractor* tool[5]. Redirects, linked phrases and target page titles have been extracted from HTML code with the BeautifulSoup library for Python.

4.1.2. Test Set

The gold set of term examples with their synonyms (*Yomikaehyou*) has been downloaded from the Export Control page of Japanese Ministry of Economy, Trade and Industry[6]. There are currently (as for January 25, 2023) 83 examples in the set. Because it contains sentences as "measuring equipment that uses linear variable differential transformers (LVDTs)", we removed all entries including verbs, as they are not technical terms but rather descriptions of their categories that cannot be precisely matched (77 is the number of terms after removing sentences). Most of the gold set terms have more than one synonym, for example "Solid-state cameras: CCD cameras, CMOS cameras". In some cases differences are in type of writings. For example term "photodetector" has three separate synonyms: "phototransistor", "photodiode" written in Chinese characters and "photodiode" written in katakana syllables used for loan words.

[4] https://dumps.wikimedia.org/jawiki/, version 20230101.

[5] https://github.com/attardi/wikiextractor

[6] https://www.meti.go.jp/policy/anpo/matrix_intro.html

4.2. Experimental Setup

We used every target word from the thesaurus described above and run the algorithms explained in Section 3.

5. Experimental Results

The results presented in Tables 1 and 2 show that redirect-based approach yields much better results than utilizing inner-links. The highest F-score for single methods is achieved by REDIRECTING WT but improved when this method is combined with REDIRECTED. When it comes to precision, also redirect functionality obtained better scores, but this time a single method (REDIRECTED) appeared to be higher than the best combination (REDIRECTING WT with REDIRECTED). On the other hand, overall scores of inner-links-based methods were minuscule with over 10 times lower F-score when compared to the redirect-based ones, meaning that recall has not improved the results as expected. None of the combinations scored higher than the single LINKED WT method, showing that implementing threshold removed many problematic synonym candidates.

Target Term	$Synonym_1$	$Synonym_2$	$Synonym_3$	$Synonym_4$	$Synonym_5$
asshuki (compressor)	*dendo kuuki asshuki (electric air compressor)*	*kuuki asshuki (air compressor)*	*konpuressaa (compresser)*	*eakonpuressaa (air compressor)*	***konpuressa (compressor)***
uran (uranium)	*U*	***uraniumu (uranium)***	*uran-235 (uranium-235)*		
genshiro atsuryoku youki (reactor pressure vessel)	*atsuryoku youki (pressure vessel)*	***genshiro youki (reactor vessel)***			
kotai satsuzou soshi (solid state image sensor)	*satsuzou soshi (image sensor)*	*imeeji sensaa (image sensor)*	*imeeji sensa (image sensor)*	*satsuei soshi (image sensor)*	
jikuuke (bearing)	***bearingu (bearing)***	*jikuu-ke (bearing)*	*rooraa bearingu (rolling-element bearing)*		
shuuseki kairo (integrated circuit)	***IC***	***LSI***	*chippu (chip)*	*IC chippu (IC chip)*	*VLSI*
shinkuu ponpu (vacuum pump)	***bakyuumu ponpu (vacuum pump)***	*kou-shinkuu ponpu (high vacuum pump)*			
tanso sen'i (carbon fiber)	***kaabon faibaa (carbon fiber)***	*kaabon (carbon)*	*kaabon-faibaa (carbon fiber)*	*tanso sen'i kyouka purasuchikku (carbon fiber reinforced plastics)*	*tanso-kei (carbon related)*
hakkou daioodo (light emitting diode)	***LED***	*furu karaa LED (full color LED)*	*LED-shiki (LED type)*	*aoiro hakkou daioodo (blue light emitting diode)*	*LED raito (LED light)*
hikari kenshutsu-ki (photodetector)	*hikari sensaa (light sensor)*	*kenshutsu-ki (detector)*	*hikari sensa (light sensor)*		
ben (valve)	***barubu (valve)***				
mujin koukuu-ki (unmanned aerial vehicle)	***UAV***	***doroon (drone)***	*mujin-ki (unmanned vehicle)*	*mujin teisatsu-ki (unmanned reconnaissance vehicle)*	*mujin (unmanned)*
rejisuto (resist)	***fotorejisuto (photoresist)***				

Table 3: Synonyms for target words extracted by the REDIRECTING WT method. Technical terms which exist in the gold set are marked with bold font. Due to the space constraints only up to five synonyms are given (10 out of total 54 have been truncated).

However, decreasing the number of candidates from 2,995 to 290 also lead to decreasing correct discoveries from 10 to 3.

6. Discussion

While we expected the inner-links methods too be weak as it treats all linked words as potentially related words, the methods based on the Wikipedia's redirect functionality, even if much better than inner-links, appeared to be far from perfect. Our assumption was that because Wikipedia creators use their knowledge to create meaningful links between pages, it will be possible to achieve a relatively high precision. As we deal with very specific expert knowledge which is not so widely represented in Wikipedia as, for instance, field of medicine or biology (Yang and Colavizza, 2022), the recall was not expected to be high. Moreover, the thesaurus used as the gold set is meant for experts, while Wikipedia is made by and targeted mostly by non-experts. This lead to the situation where a target word is very often redirected to more popular synonyms, while the gold set contains also less obvious equivalents. For example, in Table 3 which presents part of results of REDIRECTING WT method, popular synonyms of "integrated circuit" like "IC" or "LSI" are found, while in the thesaurus made by export control experts we can find synonyms like "monolithic IC" or "hybrid IC" written entirely in English. For certain, small size of the thesaurus and poor coverage of the terms in Wikipedia led to very low scores. Out of 77 terms in the gold set, only 21 had their pages in Wikipedia and the total number of gold thesaurus synonyms for these

terms with dedicated pages was 47. Of these, 14 were correctly extracted using the LINKING method, which means recall of 29.8% if only the terms with pages are considered. The lack of relevant content in both datasets seem to be a major problem, however it must be noted that our testing method is very strict. For example, when we showed full version of Table 3 to an export control expert, he ruled out only 10 out of 54 extracted synonyms as most probably improper to be included in the official list. If we had access to many experts we could perform more suitable evaluation experiment, unfortunately there are only few of them in Japan.

7. Conclusion

In this work we tested how inner-links and redirect functionality of Wikipedia can help to find synonyms of technical terms regarding export control regulations for the trade security. We discovered that although redirect-based methods yield much better results than inner-links, the expert-made thesaurus used for evaluation has too few overlaps with Wikipedia to achieve satisfactory F-score. However, a small evaluation performed by a single expert suggest that the tested methods have much bigger potential than the scores indicate.

8. Future Work

To improve the results, in the near future we are planning to implement similarity measures of linked pages and combine older approaches which utilize context clustering (Courseault Trumbach and Payne, 2007; Judea et al., 2014). We will also test various extraction methods and tools to enlarge the number of synonym candidates also for lay-person term equivalents similarly to the work performed by (Sandoval et al., 2019). By generating high-quality synonym candidates list we will aim to lessen the burden of experts who have to manually check the appropriateness of the technical terms. When the goal is achieved, we plan to extend the government-created thesaurus by finding all possible synonym candidates for the glossary published by the Japan Machinery Center for Trade and Investment.

9. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 20K12556.

References

Abrahamsson, Emil, Timothy Forni, Maria Skeppstedt, and Maria Kvist, 2014. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*.

Bøhn, Christian and Kjetil Nørvåg, 2010. Extracting named entities and synonyms from wikipedia. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications*. IEEE.

Bollegala, Danushka, Georgios Kontonatsios, and Sophia Ananiadou, 2015. A cross-lingual similarity measure for detecting biomedical term translations. *PloS one*, 10(6):e0126196.

Butters, Jonathan and Fabio Ciravegna, 2008. Using similarity metrics for terminology recognition. In *LREC*.

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

Courseault Trumbach, Cherie and Dinah Payne, 2007. Identifying synonymous concepts in preparation for technology mining. *Journal of Information Science*, 33(6):660–677.

Frantzi, Katerina, Sophia Ananiadou, and Hideki Mima, 2000. Automatic recognition of multi-word terms:. the c-value/nc-value method. *International journal on digital libraries*, 3:115–130.

Hwang, Myunggwon, Do-Heon Jeong, Seungwoo Lee, and Hanmin Jung, 2011. Measuring similarities between technical terms based on wikipedia. In *International Conference on Internet of Things and on Cyber, Physical and Social Computing*.

Jagannatha, Abhyuday, Jinying Chen, and Hong Yu, 2015. Mining and ranking biomedical synonym candidates from Wikipedia. In *Proceedings of the sixth international workshop on health text mining and information analysis*.

Judea, Alex, Hinrich Schütze, and Sören Brügmann, 2014. Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical Papers*.

Liwei, Zhang, 2022. Chinese technical terminology extraction based on dc-value and information entropy. *Scientific Reports*, 12(1):20044.

Matsuzawa, Keiichi and Mitsuo Hayasaka, 2022. Associating technical documents and export control laws using Japanese-English translation for regulatory compliant data access control (in Japanese). In *Proceedings of The 36th Annual Conference of the JSAI, 1D5-GS-11-03*.

Rzepka, Rafal, Daiki Shirafuji, and Akihiko Obayashi, 2021. Limits and challenges of embedding-based question answering in export control expert system. In *Proceedings of the 25th International Conference on Knowledge-Based and Intelligent Information & Engineering System*. Szczecin, Poland: Springer.

Sandoval, Antonio Moreno, Julia Díaz, Leonardo Campillos Llanos, and Teófilo Redondo, 2019. Biomedical term extraction: NLP techniques in computational medicine. *IJIMAI*, 5(4):51–59.

Yagahara, Ayako and Tetta Sato, 2020. Evaluation of the automatic full form retrieval method from abbreviation using word2vec for terminology expansion (in Japanese). *Nihon Hoshasen Gijutsu Gakkai Zasshi*, 76(11):1118–1124.

Yang, Puyu and Giovanni Colavizza, 2022. A map of science in Wikipedia. In *Companion Proceedings of the Web Conference 2022*.