An Examination of the Validity of General Word Embedding Models for Processing Japanese Legal Texts

Linyuan Tang
linyuan-tang@g.ecc.u-tokyo.ac.jp
Graduate School of Interdisciplinary Information Studies,
The University of Tokyo
Tokyo, Japan

Kyo Kageura
kyo@p.u-tokyo.ac.jp
Interfaculty Initiative in Information Studies,
The University of Tokyo
Tokyo, Japan

ABSTRACT

Thanks to the recent developments in distributed representation learning and the large amounts of published and digitized legal texts, computational linguistic analysis of legal language becomes possible and efficient. However, most of these open language resources and shared tasks are in English. For the languages that have little open legal texts like Japanese, a word embedding model trained on the specific language usages is accompanied by the concern of less accuracy and representativeness. Based on the observation that legal language shares a modest common vocabulary with general language, we examined the validity of using the pre-trained general word embedding model for processing legal texts by an intrinsic evaluation constructed on pairs of synonyms and related terms which were extracted from a legal term dictionary. We first investigated the settings of hyperparameters of the embedding models trained on legal texts. Then we compared the performances of our domain-specific models with general models. The pre-trained Wikipedia model conducted a better performance than domain-specific models on detecting semantic relations. This model also showed a higher compatibility with legal texts than the general model trained on newspaper articles. Although researchers tend to indicate the importance of domain-specific representation models, a general model can still be an alternative solution when there is little language resource.

1 INTRODUCTION

Owing to the emergence of *Word2Vec* [7, 8] and the following explosive improvements in distributed representation learning, use of distributed representation models as features becomes a paradigm in automated semantic analysis. In general, for the construction of such models, trainings on large-scale balanced corpora are ideal and necessary, and for evaluation, shared downstream tasks and robust evaluation measures are required.

Resources of general language usages are abundant in major languages. However, when processing texts in specialised domains, vocabularies of these domains can be very different from general language. Besides those so-called "technical terms" appeared in every specialised domain, there are also words called "sub-technical terms" that "activate a specialised meaning in the legal field, being frequently used as general words in everyday language" [5]. The high ratio of sub-technical terms in legal English vocabulary differentiates it "from the lexicon of other LSP (Language for Specific Purposes) varieties" [6]. The existence of technical terms and sub-technical terms indicates that words in the specialised domain are not only different from general language in the aspect of vocabulary, but also in the aspect of semantics. That can make the application of general word embedding models to specialised domains inefficient and unreasonable due to the inconsistency of the semantic spaces. Thus, to let the compositions of the processing texts stay along with the embedding models in the same semantic space, the domain-specific models are preferable.

Unfortunately, in comparison with English, there are less open source data of legal texts in Japanese. Either the documents are not in machine-readable data format, or they are not even open to public. As estimated in [7] that "both using more data and higher dimensional word vectors will improve the accuracy", less data will cause lower accuracy conversely. Nevertheless, that legal language has a high ratio of overlapping of the vocabulary with general language provides a possibility for us to apply general embedding models.

Therefore, in this paper, we examine whether general embedding models, specifically, a Japanese word embedding model pre-trained on Japanese Wikipedia and a model trained on newspaper articles, can be used when processing legal texts. We start with constructing a similarity and relatedness task as an intrinsic evaluation of trained embedding models. Pairs of synonyms and related terms are extracted from a Japanese legal term dictionary. We train domain-specific embedding models on two legal text datasets with different settings of hyperparameters and investigate the best configurations. The comparison of the general embedding models and the domain-specific models are then conducted by the intrinsic evaluation.

Although the performance of a model mostly depends on downstream tasks, we believe that it is also important for researchers to have an awareness of the distributed representations inside of the embedding models when trying to use them to achieve better scores in specific tasks and to solve the real world problems.

2 RELATED WORK

NLP tasks related to legal issues, including legal information retrieval, document classification, question answering methods and so on, have been increasingly attracting attention from both computational linguists and legal professionals.

To improve the performances of theses tasks with the assistance of semantic analysis, there were two word embedding models specifically trained on legal texts. One was the pre-trained model

In: Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2019), June 21, 2019, Montreal, QC, Canada.

Published at http://ceur-ws.org

built in a Python library called LexNLP [1]. LexNLP focused on natural language processing and machine learning for legal and regulatory text. The pre-trained models were based on thousands of real documents and various judicial and regulatory proceedings. The other one was *Law2Vec* [1] provided by LIST[2]. This model "oriented to legal text trained on large corpora comprised of legislation from UK, EU, Canada, Australia, USA, and Japan among other legal documents." Although legal texts of Japan seemed to be used for achieving semantic representations of words in the legal domain, the used texts were English-translated and the models were for legal English.

COLIEE (the legal question answering Competition on Legal Information Extraction/Entailment) [4] is the only competition about Japanese legal texts and providing law articles both in Japanese and English as a knowledge resource. COLIEE 2017 focused on extraction and entailment identification aspects of legal information processing related to answering yes/no questions from Japanese legal bar exams. Carvalho et el. [2] and Nanda et el. [9] both tested the Google News dataset pre-trained vectors[3] in information retrieval, and the former team also found that the "pure common text embedding" resulted in poor performance, "most probably due to the absence of legal vocabulary and corresponding semantics."

The evaluation of the word embeddings trained from different textual resources has been conducting in the biomedical domain. Roberts [12] revealed that combinations of corpora led to a better performance. Wang et al. [13] concluded that the word embeddings trained on the biomedical domain did not necessarily have better performance than those trained on the general domain. While they both agreed that the efficiency of a word embedding model was task-dependent, Gu et al. [3] argued that even smaller domain-specific corpora may be preferable to pre-trained word embeddings built on a general corpus if the diversity of vocabulary was low.

In general, related work tends to indicate the importance of domain-specific distributed representation models for processing specialised texts.

3 DATA

Our dataset consists of three corpus, a dictionary of legal terms (hereinafter, referred to as *dictionary*), "the fact of the crime" parts of the judgements obtained from *Westlaw Japan* [4] judicial precedent corpus (referred to as *judgements*), and newspaper articles contained in *Mainichi Newspaper Corpus* (referred to as *newspaper*). Basic statistics of our corpus are given in Table 1. Detailed descriptions of each corpus are given below.

Dictionary. The technical term dictionary adopted in this work was *Yuhikaku Legal Term Dictionary (4th edition)*. The dictionary consists of 13,812 entry words with the definitions written by experts and carefully edited. We simply referred the word "legal term" (or "term") to the entry words that were recorded in the dictionary instead of getting involved in the sophisticated discussion about the meaning of the word. In the dictionary, a synonym of term t is given when t has no definition and is labeled with a See tag, while a related term of t is labeled with a See Also tag when the experts thought more information was needed

Table 1: Basic statistics of our dataset.

Corpus	#Token	#Type
dictionary	781,027	20,328
judgements	790,665	17,423
dictionary+judgements	1,571,692	30,915
newspaper	22,928,051	242,630

Judgements. We obtained 2,306 judgements passed on criminal cases in district courts nationwide from 2008 to 2017. Legal English is known as legalese because of its tedious and puzzling language usage. Legal Japanese also shared these problems. Therefore, in order to conduct a moderate comparison with newspaper articles in the aspect of contents and document lengths, we extracted "the fact of the crime" part from each judgement.

Newspaper. When a case happened, it is often reported as an article in the social section of the newpaper. Additionally, the language usage in a newspaper article can be considered as a general usage, or at least less specialized than legalese used in legal texts. We obtained all the articles from a one-year (2015) corpus. 1748 legal terms were observed in these articles.

4 METHODS

Before processing, we applied Japanese morphological analyzer *Chasen*[5] to split the sentences into words and remove signals and numbers. The similarity measured between two vectors in this paper were all cosine similarity.

The examining procedure was in two steps. First, we built a term pair inventory for performance evaluation. Term pairs were separated into synonym pairs and related pairs. Domain-specific models were then trained with hyper parameter tuning on this inventory. Second, we focused on the common term pairs existed in both general models and domain-specific models. The performances of the models were examined on both synonym detection and related term detection.

4.1 Task Design

We extracted 1440 pairs of synonyms and 6641 pairs of related terms by exploiting the indicative tags provided in the dictionary. These pairs constructed the gold standards of synonym detection task and related term detection task for evaluating each model's ability of catching semantic relations between terms.

We evaluated the performances of models by counting how many semantic relations were correctly caught by each model. Specifically, we first obtained top n most similar words of term t from the model. n was set to {1, 5, 10}. If the synonym or the related term was in these most similar words, we treated the trial as a correct one. The performance was represented by *accuracy* as the ratio of correctly predicted pairs to all synonym or related term pairs.

[1] https://archive.org/details/Law2Vec.
[2] http://www.luxli.lu/university-of-athens/.
[3] https://code.google.com/archive/p/word2vec.
[4] https://www.westlawjapan.com/.
[5] version: 0.996, neologd 102.

Table 2: Vocabulary sizes of word embedding models. The sizes of domain-specific models are presented in the order of min.count = {2, 3, 5}. The min.count value of general models were 3.

Source	#Vocabulary		
dictionary	8,803	7,202	5,697
judgements	9,539	7,747	5,991
dictionary+judgements	14,292	11,769	9,318
Wikipedia		1,463,528	
newspaper		242,630	

Table 3: Hyperparameter tuning for domain-specific models.

Parameter	Value
dimension	50, 100, 200, 300, 400
window size	2, 3, 5, 10, 15
min.count	2, 3, 5
negative sample	3, 5, 10, 15

Table 4: The best accuracy scores on synonym detection under different configurations. (1440 synonym pairs)

Model	Synonym (%)		
	top 1	top 5	top 10
dictionary	3 (0.2%)	5 (0.3%)	6 (0.4%)
dictionary+judgements	3 (0.2%)	4 (0.3%)	5 (0.3%)
Wikipedia	**15 (1.0%)**	**56 (3.9%)**	**79 (5.5%)**
newspaper	5 (0.3%)	18 (1.3%)	27 (1.8%)

4.2 Model Training

We applied pre-trained *Wikipedia Entity Vectors* as our general word embedding model [6]. It is a 300-dimension Skip-Gram Negative Sampling (SGNS) model. With the same training configuration of it, we trained another general model on newspaper articles for the comparison within general models. We then trained our domain-specific models on the dictionary and the judgements, respectively and together. The size of the source data and vocabularies are given in Table 2.

The performance of word embedding models can be improved by hyperparameter tuning. Since the effects of different configurations can be diverse, we investigated hyperparameter settings as in [10]. We exploited `gensim` [11] for model training. Examined parameters and values are shown in Table 3. Each model had five chances on each task.

5 RESULTS

5.1 Model Tuning

The best accuracy scores of models on the two tasks under different configurations are shown in Table 4, 5. The models trained on judgements failed in detecting both synonym and relatedness relations. The best accuracy of those judgement models was 0 (0.0%) for synonym pairs, and 19 (0.3%) for related term pairs. In both tasks, additional legal texts (i.e., judgements) did not improve the performance of our domain-specific models, which indicated that our legal text dataset is biased to the dictionary dataset and the more data does not always lead to the better performance.

The default training configuration of `gensim` is {dimension = 100, window size = 5, min.count = 5, negative sample = 5}. The selected configuration after a hyperparameter tuning on an English domain-specific model training [10] was {dimension = 400, window size = 5, min.count = 5, negative sample = 5}. However, we found that *window size* or *negative sample* that was lower than 10 would led to worse performances in all circumstances. Due to the relatively tiny data size, *min.count* that larger than 3 also had a negative effect on the performances.

The most suitable configuration for the models trained on the dictionary across the variation of *top_n* was {dimension = 300, window size = 15, min.count = 3, negative sample = 10}. It is similar to the configuration of the Wikipedia model which is {dimension = 300, window size = 10, min.count = 3, negative sample = 10}.

We selected the same values of parameters as the Wikipedia model as the training configuration of our domain-specific model with which the general models would be compared on the next stage.

[6] https://github.com/singletongue/WikiEntVec. Wikipedia data until 2018.10.01.

Table 5: The best accuracy scores on related term detection under different configurations. (6641 related term pairs)

Model	Related term (%)		
	top 1	top 5	top 10
dictionary	83 (1.2%)	163 (2.5%)	206 (3.1%)
dictionary+judgements	78 (1.2%)	159 (2.4%)	208 (3.1%)
Wikipedia	**142 (2.1%)**	**371 (5.6%)**	**472 (7.1%)**
newspaper	43 (0.6%)	107 (1.6%)	153 (2.3%)

Table 6: Results of synonym detection. (18 synonym pairs)

Model	Synonym (%)		
	top 1	top 5	top 10
dictionary	1 (5.6%)	1 (5.6%)	1 (5.6%)
Wikipedia	**3 (16.7%)**	**8 (44.4%)**	**8 (44.4%)**
newspaper	1 (5.6%)	2 (11.1%)	4 (22.2%)

Table 7: Results of related term detection. (564 related term pairs)

Model	Related term (%)		
	top 1	top 5	top 10
dictionary	44 (7.8%)	84 (14.9%)	108 (19.1%)
Wikipedia	**58 (10.3%)**	**135 (24.0%)**	**165 (29.3%)**
newspaper	18 (3.2%)	40 (7.1%)	49 (8.7%)

5.2 Intrinsic Evaluation

As shown in Table 4, 5, the Wikipedia model achieved higher performances on detecting semantic relations of legal terms, even those relations were obtained from the legal domain. This result can be due to the absence of low frequency terms in the dictionary corpus. Therefore, we further conducted two detection tasks on the common pairs among the domain-specific model, the Wikipedia model and the newspaper model. There were 18 common synonym pairs and 465 common related term pairs. Results of the experiment are shown in Table 6, 7.

The Wikipedia model achieved the best accuracy score among three models, while the same general embedding model, the newspaper model, was the worst. The performance difference between the Wikipedia model and the newspaper model also confirmed that the performance of general models are effected by the diversity of general language resources. The similar results of the examination on the common term pairs to the examination on all term pairs indicated that the Wikipedia model is superior to the domain-specific dictionary model for catching the intrinsic semantic relations of legal terms.

6 CONCLUSION

Since the usefulness of an embedding model mostly depends on the downstream tasks, we don't argue that which embedding model is better or worse for legal NLP tasks. The purpose of this research is to investigate whether a general corpus could be used when the training on the specific domain is not practicable. The word embedding model built on Wikipedia showed a considerable performance on the intrinsic evaluation. The legal domain is different from other specialised domains in the aspect of the ratio of overlapping words with general language. This characteristic is helpful when there are not enough domain-specific language resources. In this paper, we provided some evidence that domain-specific word embedding models are not always outperform general models and not all the domain-specific texts are useful when constructing the semantic relations among technical terms. The using of general word embedding models, especially the models trained on a balanced large-scale corpus, therefore can be considered as an alternative way to processing those domain-specific texts.

ACKNOWLEDGMENTS

The authors would like to thank *YUHIKAKU Publishing Co., Ltd.* for providing the legal dictionary dataset. We are also grateful to the reviewers for their valuable comments and suggestions.

REFERENCES

[1] Michael James Bommarito, Daniel Martin Katz, and Eric Detterman. 2018. LexNLP: Natural Language Processing and Information Extraction For Legal and Regulatory Texts. *SSRN Electronic Journal* (2018). https://doi.org/10.2139/ssrn.3192101

[2] Danilo S. Carvalho, Vu Tran, Khanh Van Tran, and Nguyen Le Minh. 2017. Improving Legal Information Retrieval by Distributional Composition with Term Order Probabilities. In *COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment (EPiC Series in Computing)*, Ken Satoh, Mi-Young Kim, Yoshinobu Kano, Randy Goebel, and Tiago Oliveira (Eds.), Vol. 47. EasyChair, 43–56. https://doi.org/10.29007/2xzw

[3] Yang Gu, Gondy Leroy, Sydney Pettygrove, Maureen Kelly Galindo, and Margaret Kurzius-Spencer. 2018. Optimizing Corpus Creation for Training Word Embedding in Low Resource Domains: A Case Study in Autism Spectrum Disorder (ASD). *AMIA Annual Symposium proceedings* (2018), 508–517.

[4] Yoshinobu Kano, Mi-Young Kim, Randy Goebel, and Ken Satoh. 2017. Overview of COLIEE 2017. In *COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment (EPiC Series in Computing)*, Ken Satoh, Mi-Young Kim, Yoshinobu Kano, Randy Goebel, and Tiago Oliveira (Eds.), Vol. 47. EasyChair, 1–8. https://doi.org/10.29007/fm8f

[5] María José Marín and Camino Rea. 2014. Researching Legal Terminology: A Corpus-based Proposal for the Analysis of Sub-technical Legal Terms. *ASp* 66 (nov 2014), 61–82. https://doi.org/10.4000/asp.4572

[6] María José Marín Pérez. 2016. Measuring the Degree of Specialisation of Sub-technical Legal Terms through Corpus Comparison: A Domain-independent Method. *Terminology* 22, 1 (2016), 80–102. https://doi.org/10.1075/term.22.1.04mar

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. (2013). arXiv:cs.CL/1301.3781v3

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems*. 3111–3119.

[9] Rohan Nanda, Adebayo Kolawole John, Luigi Di Caro, Guido Boella, and Livio Robaldo. 2017. Legal Information Retrieval Using Topic Clustering and Neural Networks. In *COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment (EPiC Series in Computing)*, Ken Satoh, Mi-Young Kim, Yoshinobu Kano, Randy Goebel, and Tiago Oliveira (Eds.), Vol. 47. EasyChair, 68–78. https://doi.org/10.29007/psgx

[10] Farhad Nooralahzadeh, Lilja Øvrelid, and Jan Tore Lønning. 2018. Evaluation of Domain-specific Word Embeddings using Knowledge Resources. In *Proceedings of the 11th Language Resources and Evaluation Conference*. European Language Resource Association, Miyazaki, Japan. https://www.aclweb.org/anthology/L18-1228

[11] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[12] Kirk Roberts. 2016. Assessing the Corpus Size vs. Similarity Trade-off for Word Embeddings in Clinical NLP. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*. The COLING 2016 Organizing Committee, Osaka, Japan, 54–63. https://www.aclweb.org/anthology/W16-4208

[13] Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. 2018. A Comparison of Word Embeddings for the Biomedical Natural Language Processing. *Journal of Biomedical Informatics* 87 (nov 2018), 12–20. https://doi.org/10.1016/j.jbi.2018.09.008