INTERNATIONAL JOURNAL OF TRANSLATION
VOL. 28, NO. 1-2, JAN-DEC 2016

A Comparison of Statistical Models for Information Retrieval in Patent Translation: A Probability Distribution Approach

DARREN HSIN-HUNG LIN
YVES LEPAGE
Waseda University, Japan

ABSTRACT

By investigating the lexical information of patent translation from USPTO (United States Patent and Trademark Office) and JPO (Japan Patent Office), we firstly calculate the substitution probability and matching strength to measure the average retrieval probability based on the Syntagmatic Paradigmatic Model (SPM; Dennis & Harrington 2001). We further apply Spearman's rank correlation, based on the Pooled Adjacent Context Model (PAC; Redington, Chater & Finch 1998), to identify the similarity between words. Our proposed model demonstrates the specific domain of the patent corpora. The PAC model shows a substantive advantage over the SPM model for the application on prepositional phrases. The SPM model performed similarly with the same substitution probability (0.5). Both models, however, performed equivalently with associative structures of syntagmatic similarity.

Keywords: Patent translation, statistical model, probability, distributional hypothesis

1. INTRODUCTION

Patent documents are structural documents with their own distinguishing characteristics from general documents. Kim & Choi (2007: 1201) indicate that such structural characteristics of

patent in previous works were not properly considered. Therefore, the semantics of patent structure is one of the most important features for categorization purposes.

As patent include much technical terminology, applicants may define and use their original terms not used in other patents. Chen & Chiu (2011) utilize the special characteristics of patent documents to generate the indexing vocabulary for presenting all the patent documents. Chen & Chiu (2012), however, reveal that patent documents havesome characteristics that make it difficult to apply traditional feature selection methods directly.

Together with scientific articles, patent data represents the main source of technical information in the world (Lupu 2014: 204). It covers all technical, scientific, and engineering fields. To characterize the specialized vocabulary in modern patent language, Symonds, Bruza & Sitbon (2014) evaluated the efficiency of corpus-based distributional models for literature-based discovery on patent information. The computational complexity analysis finds that distributional approaches can allow faster computation of associations between vocabulary terms.

To rich the terminological knowledge of word associations in patent retrieval, we propose the method for extracting distribution information in patent translation to address the distributional characteristics of the term candidate.

2. RELATED WORK

There has been continuing interest in distributional information related to patent studies (Tsai 2010, 2012; Tseng 2011). As patent writers often invent novel terms rather than using standard ones to make finding a patent hard (Judea, Schutze & Brugmann 2014: 297), Lin & Hsieh (2010a) compiled the patent technical wordlist to induce the distributional structures of semantic associations in contemporary English patents. Lin & Hsieh (2010b) statistically retrieved the distributional patterns of patent claims from LexisNexis, database for legal professions. Their study supports that statistical approach is useful for measuring the linguistic features of technical documents. In Lin & Hsieh

(2011), they characterized the distributional features of independent claim using statistical-retrieval approaches. Their work suggests statistics plays a crucial role to identify the distributional information of technical documents. However, as English becomes the lingua franca, the number of translated abstracts of technical documents has incredibly increased. To further investigate the distributional similarity and difference, there is a need to study on distributional statistics of patent translation for native characterization.

Lin and Lepage (2014) studied on native and non-native writings of technical documents. They firstly compare the technical terms distribution of patent technical words in USPTO and JPO. In turn, they extracted the co-occurrence information of transitional phrases for genre analysis. The findings based on the distribution statistics implies that technical genre reveals more distributional and meaning similarity. To make a further step to refine the preliminaries of their work, the present study aims to research on distributional properties of the co-occurrence of words in patent translation to define the role of the context and its influence on the representation of syntagmatic and paradigmatic relations.

3. DISTRIBUTIONAL HYPOTHESIS

Distributional hypothesis, according to Sahlgren (2005), defining two words that constantly occurs with the same contexts are justified in assuming that they mean similar thing. Sahlgren (2008: 33) maintains that distributional approaches to meaning acquisition utilize distributional properties of linguistic entities as the building blocks of semantics. In other words, there is a correlation between distributional similarity and meaning similarity, which allows us to utilize the former in order to estimate the latter. In Cohen & Widdows (2009)'s explanation, distributional hypothesis is often cited as a theoretical motivation for distributional semantics of sub-language theory is also attributed to Zellig Harris, whom provides a framework to characterize the additional constraints of the language used in a specialized domain.

Distributional methods could themselves embody innate knowledge. Redington, Chater & Finch (1998: 464) imply that the use of distributional information is consistent with any point on the native-empiricist continuum. They point out that words of the same category tend to have a large number of distributional regularities in common can be used as a cue to syntactic category. Distributional methods is thus considered possible source of information about syntactic structure.

There is a growing body of research suggesting that distributional information plays a more powerful role than previously thought in a number of aspects of language processing (McDonald & Ramscar 2001: 611). The purpose of the present study is therefore to examine the distributional characteristics of regularities and irregularities in patent translation. Our investigations are framed under the Syntagmatic Paradigmatic Model (SPM; Dennis & Harrington 2001) and the Pooled Adjacent Context Model (PAC; Redington, Chater & Finch 1998).

4. STATISTICAL MODELS

Dennis (2003) compares the statistical models for the extraction of lexical information from text corpora. In this work, two statistical models were applied. They are the Syntagmatic Paradigmatic Model (SPM) and the Pooled Adjacent Context Model (PAC).

The application of Syntagmatic Paradigmatic Model is to calculate the average retrieval probability. It defines a fragment as a sequence of words bounded by very high frequency words and assigns fragments with the same high frequency words patterns to the same equivalence. We give an instance in Figure 1.

The matching strength (hereafter, Match) is the count of the number of the words in position that the fragments had in common. The substitution probability (hereafter, P) is the calculation of each fragment within an equivalence class was matched against each other fragment in that class. The average retrieval probability is thus 0.415.

	Match	P(Retrieval)
A picture OF THE		
A copy OF THE	3	0.33
A description OF THE	3	0.33
A side OF THE	3	0.33
ONTO THE picture [E]		
ONTO THE copy [E]	3	0.5
ONTO THE table [E]	3	0.5

P(<picture, copy>) = (0.5+0.33)/2 = 0.415

Figure 1. *Illustration of SPM model calculation (Dennis 2003)*

The application of the Pooled Adjacent Context Model (PAC), on the other hand, is to calculate the similarities between words are determined using Spearman's rank correlation. Redington, Chater & Finch (1998) suggest that PAC constructs a representation of a word by accumulating frequency counts of the words that appeared in the two positions immediately before and immediately after the target word. Figure 2 demonstrates the use of PAC.

Example Windows of Text

found	a	picture	of	the
found	a	picture	in	her
a	pretty	picture	of	her
found	a	copy	of	a
found	a	copy	below	the
destroyed	the	copy	of	the

Corresponding Pooled Vectors

picture	2	0	1	1	2	0	2	1	0	1	2	0
copy	2	1	0	0	2	1	2	0	1	2	0	1
	found	destroyed	a	pretty	a	the	of	in	below	the	her	a
	Pos -2			Pos -1			Pos 1			Pos 2		

Figure 2. *Illustration of PAC model calculation (Dennis 2003)*

According to Dennis (2003), the SPM model contexts are kept separate and similarities are pooled, whereas the PAC model contexts are pooled and then similarities calculated. However, both models rely on the immediately surrounding words to act as a form of context.

5. THE CORPORA

We employ USPTO glossary[1] as a reference tool as the first step to calculate the term distribution in PatFT[2] and PAJ[3] in last decade, year 2000 to 2009.

Table 1. *Distribution of technical words in USPTO and JPO*

Country	**USA**	**JAPAN**
Patent office	USPTO	JPO
Database	PatFT	PAJ
Times of occurrence	23,646,035	2,681,473

Of patent claims, "independent claim" best describes the invention in adding essential features. Among the patent claims, "independent claim" ranked the 2nd in USPTO and occurred 587,926 times; ranked the 1st in JPO and occurred 15,513 times. Table 2 elaborates the term distribution information.

Table 2. *Term distribution of patent claims in USPTO and JPO*

Patent Claim	**Times of Occurrence**	
	USPTO	**JPO**
dependent claim	625,886	3,739
independent claim	587,926	15,513
benefit claim	437,599	1,256
priority claim	381,352	11,858
withdrawn claim	227,433	3,259
canceled claim	32,306	14,793
multiple dependent claim	494	60
TOTAL	2,292,996	50,478

We therefore retrieve "independent claim" as the candidate term from LexisNexis[4] and further compile the corpora of English abstracts in USPTO and JPO within year 2014.

Table 3. *The comparable corpora of independent claim*

	USPTO	JPO
Number of patents	932	850
Citation	1,206	649
Size (kb)	100,605	53,924

The information of the comparable corpora is given in Table 3. It is shown that the number as well as the citations of *independent claim* in USPTO is outperformed in JPO. To discriminate the characteristics of term distribution information, we conduct experiments based on the Syntagmatic Paradigmatic Model and the Pooled Adjacent Context Model.

6. RESULTS

6.1. *SPM Model*

The relevance for prepositional phrases of the technical term "independent claim" in USPTO has a matching of 2. As there are two fragments of equal strength the retrieval probability of each fragment is 0.5, and so the substitution probability as calculated from the fragment between "preamble" and "feature" is 0.5.

Table 4. *Substitution probability of prepositional phrases in USPTO*

Equivalence Class	Frequency	M	P(Retrieval)
the preamble of	106	2	0.5
the features of	100	2	0.5
the characterizing part of	38	2	0.5

The second instance of the word "feature" appears in the context of the fragment "the feature of" in Table 5. Retrieval using this fragment results in a substitution probability of 0.5, so that the average retrieval probability is also 0.5.

Table 5. *Substitution probability of prepositional phrases in JPO*

Equivalence Class	Frequency	M	P(Retrieval)
the features of	40	2	0.5
the subject of	8	2	0.5

6.2. *PAC Model*

The correlation of prepositional phrases can be thought of prevalently dominance in PAC model, where it shows a substantive advantage over the SPM model. We present the distribution of prepositions in USPTO and JPO in Table 6 and the PAC model in Figure 3.

As can be seen, there is a variety of prepositional phrases used in USPTO. The top 5 preposition are "of," "to," "in," "by," and "with." JPO, on the other hand, presents less variety.

The PAC model, in brief, shows a substantive advantage over the SPM model, in particular, the prepositional phrases.

Table 6. *Distribution of prepositions in USPTO and JPO*

Preposition	**Frequency**	
	USPTO	**JPO**
at	36	10
by	149	77
for	69	317
from	-	13
in	382	175
into	16	-
of	532	198
on	32	9
over	5	-
to	460	61
under	78	-
with	86	14

The PAC model of prepositions in USPTO and JPO is given in Figure 3. As the syntagmatic associations are thought to exist between words that often occur together, it can be found in the preposition of "of", "in" and "by." By contrast, it can be seen that the irregularities are "from" of JPO, and "into," "over," and "under" of USPTO.

The instance of PAC model for *independent claim* is given in Table 7. The six position vectors created in this way are then concatenated to form the representation of such technical term in Table 8. The position immediately after the term shows a variety

of syntagmatic associations with prepositional and conjunctive usages.

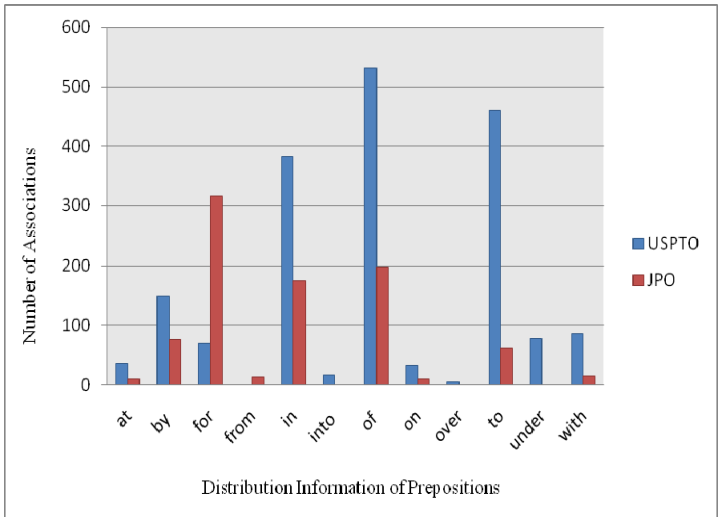

Figure 3. *PAC model of prepositions in USPTO and JPO*

Table 7. *The representation of independent claim*

	Examples windows of text
USPTO	descried in the *independent claim* of the present descried in an *independent claim* of the present descried in the *independent claim* and a process
JPO	descried in the *independent claim* of the corresponding descried in the *independent claim* of the evaporator descried in the *independent claim* and a method

Table 9 and 10 illustrates a number of examples drawn from the similarity matrices of both the SPM and PAC model to demonstrate the different sorts of information. The examples in Table 9 show the sensitivity of the models to syntactic categories. JPO shows a substantive advantage of the syntagmatic associations over USPTO in terms of "number" and "ordinal numbers" of the technical term of *define*. USPTO, on the other hand, shows a substantive advantage over JPO of the syntagmatic associations of "auxiliary" of *include*. Both models perform

equivalently with syntagmatic similarities of "passive voice" and "the progressive."

Table 8. *Illustration of PAC model calculation of independent claim*

USPTO	3	3	2	1	2	1	2	1	0	0	0	2	1
JPO	3	3	3	0	2	1	2	1	1	1	1	0	0
Corresponding pooled vectors	descried	in	the	an	of	and	the	a	corresponding	evaporator	method	present	process
Position	Pos-3	Pos-2	Pos-1		Pos1		Pos2		Pos3				

Table 9. *Similarity examples: Syntactic*

Context	**Ten Most Similar Words**	
	define	**include**
USPTO	may which is used be can are second first said	can may be combined does is are limited described determining
JPO	used is first which are two second may third respectively	may can is are limited be configured described receiving intended

As can be seen, Table 10 shows the sensitivity to semantic and associative information. The paradigmatic associations exist

between words that may not appear together but can appear in similar sentential context, such as “computer” and “data.”

The most similar words for “method” contain many words that are clearly semantically related. Such substitutional relations, for example, is distinguished by having “claim” and “invention,” “step” and “system,” and “embodiment” and “example” in both contexts.

Same substitution between words in both contexts for “invention” is recognized by “embodiment” and “example”, and “object/scope” and “method” “Example” is one of the most frequent used words. “Summary”, on the other hand, is used mostly by USPTO for native characterization.

To summarize, SPM model and PAC model are capable of capturing a significant proportion of the syntactic structure, at least for high frequency words. To solve semantic problems on words, phrases, or sentences, make use of paradigmatic associations to further refine the lexical information.

Table 10. *Similarity examples: Associative and semantic*

Context	**Ten Most Similar Words**	
	method	**invention**
USPTO	claim invention embodiment system apparatus device step computer example data	embodiment method aspect device example object accordance scope summary apparatus
JPO	claim invention device step embodiment image system picture example use	embodiment method device aspect composition purpose example system compound molecule

7. CONCLUSION AND FUTURE WORK

In summary, both models show evidence of distinguishing verbs, nouns, prepositions, passive voices, and the progressives. The SPM model performed similarly for the average retrieval probability. The PAC model shows an impressive advantage of the application on prepositional phrases over the SPM model, especially "number" and "ordinal number."

For future research, we seek to measure syntagmatic and paradigmatic associations. For example, apply semantic matching based on Moran's (1973) proposal on word associations to examine semantic mapping of "disjointness," "equivalence," "more specific," and "less specific" to further refine the preliminaries of the present study.

NOTES

1. USPTO Glossary is available at: <http://www.uspto.gov/main/glossary/index.html>.
2. USPTO Patent Full-Text and Image Database (PatFT) is available at <http://patft.uspto.gov/netahtml/PTO/search-adv.htm>.
3. Patent Abstract of Japan (PAJ) is available at <http://patft.uspto.gov/netahtml/PTO/search-adv.htm>.
4. LexisNexis is available at <https://www.lexisnexis.com>.

REFERENCES

Chen, Y. L. & Chiu, Y. T. 2011. An IPC-based vector space model for patent retrieval. *Information Processing and Managemenet*, 47, 309-322.

——., & Chiu, Y. T. 2012. Vector space model for patent document with hierarchial class labels. *Journal of Information Science*, 38/3, 222-233.

Cohen, T. & Widdows, D. 2009. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42, 390-405.

Dennis, D. 2003. A comparison of statistical models for the extraction of lexical information from text corpora. *The 25th Annual Meeting of the Cognitive Science Society*, Boston, MA.

——. & Harrington, M. 2001. The syntagmatic paradigmatic model: A distributed instance-based model of sentence processing. In M. Isahara & Q. Ma (Eds.), *Proceedings of the 2nd Workshop on Natural Language Processing and Neural Networks* (pp. 38-45).

Judea, A., Schutze, H. & Brugmann. 2014. Unsupervised training set generation for automatic acquisition of technical terminology in patents. In *Proceedings of COLING2014, the25th International Conference on Computational Linguistics: Technical Papers* (pp. 290-300).

Kim, J. O. & Choi, K. S. 2007. Patent document categorization based on semantic structural information. *Information Processing and Management*, 43, 1200-1215.

Lin, H. H. & Hsieh, C. Y. 2010a. The specialized vocabulary of modern patent language: Semantic association in patent lexis. *PACLIC 24: Pacific Asia Conference on Language, Information, and Computation* (pp. 417-424). Japan: Waseda University Press.

——. & Hsieh, C. Y. 2010b. Collocation features of independent claim in US patent documents: Information retrieval from LexisNexis. *ROCLING XXII: Conference on Computational Linguistics and Speech Processing* (pp. 296-310). Taiwan: Academia Sinica.

——. & Hsieh, C. Y. 2011. Characteristics of independent claim: A corpus-linguistic approach to contemporary English patents. *International Journal of Computational Linguistics and Chinese Language Processing*, 16/3-4, 77-106.

——. & Lepage, Y. 2014. Testing distibutional hypothesis in patent translation. *ROCLING XXVI: Conference on Computational Linguistics and Speech Processing* (pp. 185-192). Taiwan: Academia Sinica.

Lupu, M. 2014. On the usability of random indexing in patent retrieval. In N. Hernandez, R. Jäschke, & M. Croitoru (Eds.), *Graph-Based Representation and Reasoning* (pp. 202-216). Switzerland: Springer International Publishing.

McDonald, S. & Ramscar, M. 2001. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 611-616), Edinburgh, Scotland.

Moran, L. J. 1973. Comparative growth of Japanese and North American cognitive dictionaries. *Child Development*, 44, 862-865.

Redington, M., Chater, N., & Finch, S. 1998. Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22/4, 425-469.

Sahlgren, M. 2005. An introduction to random indexing. In *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering (TKE)*, Copenhagen, Denmark.

——. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20/1, 33-53.

Symonds, M., Bruza, P. & Sitbon, L. 2014. The efficiency of corpus-based distributional models for literature-based discovery on large data sets. In *Proceedings of the 2nd Australasian Web Conference* (pp. 49-57), Auckland, New Zealand.

Tsai, Y. 2010. Text analysis of patent abstracts. *The Journal of Specialized Translation*, 13, 61-80.

——. 2012. The use and misuse of high frequency nouns in English translation of Chinese patent abstracts. *FORUM: International Journal of Interpretation and Translation*, 10/2, 161-186.

Tseng, T. Y. 2011. A Study of Patent Claim Translation: Function-Plus-Loyalty Principle and Contrastive Functional Analysis. Unpublished master thesis, National Taiwan University of Science and Technology, Taipei, Taiwan.

DARREN HSIN-HUNG LIN
GRADUATE SCHOOL OF INFORMATION,
PRODUCTION, AND SYSTEMS,
WASEDA UNIVERSITY, JAPAN.
E-MAIL: <NOBUHIRO602@TOKI.WASEDA.JP>

YVES LEPAGE
GRADUATE SCHOOL OF INFORMATION,
PRODUCTION, AND SYSTEMS,
WASEDA UNIVERSITY, JAPAN.
E-MAIL: <YVES.LEPAGE@WASEDA.JP>