An Intial Study of Anchor Selection in Patent Link Discovery

Anonymised
for peer review
Anonymous@Anonymous

ABSTRACT

Patents are a source of technical knowledge, but often difficult to understand. Technological solutions that would help understand the knowledge expressed in patents can assist the creation of new knowledge, and inventions. This paper explores anchor text selection in patents for linking patents to external knowledge sources such as web pages and prior patents. While link discovery has been investigated in other domains, e.g. Wikipedia and the medical domain, the application of linking patents has received little attention; it presents some unique challenges as this paper shows. The paper contributes (1) a test collection investigating the identification of anchor text (entities) in patent link discovery, (2) a user experiment studying the selection of anchors by users, and (3) an evaluation of four popular unsupervised keyword ranking methods (TFIDF, BM25, Keyphraseness, Termex) to identify potential anchors for linking.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; **Personalization**; **Evaluation of retrieval results**; **Test collections**; *Specialized information retrieval*;

1 INTRODUCTION

Link discovery aims to link phrases in text to knowledge bases like Wikipedia in order to ease the understandability of the text. A considerable amount of literature exists on the theme of link discovery and the majority of such work is focused on Wikipedia articles [4, 10, 13]. Though recent work has shown interest in linking texts in the domains other than Wikipedia (bio-medical documents, micro blogs, etc) [3, 6], up to now far too little attention has been paid to linking patent documents to knowledge bases.

Patent link discovery is an important and distinct task for the following reasons. Firstly, patents include exhaustive scientific descriptions and valuable technological information which may not be available else where [1]. Secondly, knowledge disclosed is often trapped in the complexity of technical and scientific language; In addition patent writers often obfuscate the actual details of the invention [16]. Consequently, the information disclosed in patents is often inaccessible and not understandable, thus compromising the chief aim behind the patenting system. Finally, in contrast to users in other domains, patent users often are highly motivated to understand patents. They include researchers, inventors, patent analysts and investors, e.g. inventors - to ensure their idea is novel, researchers - to learn about existing technologies [12]. All these create the need for technological solutions to facilitate the comprehension of the patent content.

Anchor texts to be selected in patents are often technical terms while in some domains like Wikipedia anchor texts can contain named entities too [13, 18]. Both unsupervised and supervised approaches have been taken in the past for anchor text selection with similar success [13]. However, unsupervised methods are applicable to collections without prior links [13]. Most unsupervised methods consist of two main stages: (1) candidate extraction, and (2) ranking [13]. Mihalcea and Csomai have used TFIDF, χ^2 and Keyphraseness (see section 4 for definition) to identify link worthy terms. They found that Keyphraseness which is based on link probabilities obtained by sampling Wikipedia's articles is the most accurate for link detection. Itakura and Clarke's approach of link strength used in INEX is a slight variation of this [9] and it was further enhanced by Jenkinson et al [11]. Machine learning methods have also shown to be effective for the anchor text selection; explored methods include: Naive Bayes, decision trees, Conditional Random Fields and Support Vector Machines [7, 14]. The disadvantage of them is they are highly dependent on both domain and the availability of a good quality labeled set to be used for training.

As far as we know, the recent study by Tsunakawa and Kaji is the only study that attempted patent link discovery. They have employed a domain terminology extraction system known as Termex [15, 18]. In contrast to our system, where real user suggested anchor texts are used as ground truth, they have considered the existing anchor texts in Wikipedia as their ground truth when evaluating their system . The main drawback of this type of evaluation is, it fails to identify the unique needs of patent users for anchor texts.

2 DATA COLLECTION

We randomly selected 72 English language patents from the WIPO-alpha train collection which is a publicly available patent collection. These 72 patents are from two major patent sections of Mechanical Engineering (ME) and Information Technology (IT), 36 from each section. These domains were selected because we could gain access to users with suitable domain expertise who represent a reasonable selection of potential readers. The objective of our work is to identify link-worthy entities that require reference to external knowledge. Twenty four participants were recruited for the user study to be representative of a plausible set of patent readers. Among the selected, the majority were higher degree research students, some with strong IT expertise. We did not have access to patent experts (such as patent examiners or inventors). While these are important, our general target audience of link discovery is not this group. We grouped the selected patents into 12 sets, each set consisting of 3 ME patents and 3 IT patents. The grouped patents include the extracted text from the sections of title, abstract, claims

ADCS'17, December 2017, Brisbane, Queensland

ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

Table 1: Statistics of the collected entities

Patents	Total entities	Mean entities per patent	Max entities per patent	Min entities per patent
All	653	9.07	22	2
ME	348	9.66	21	2
IT	305	8.47	23	3

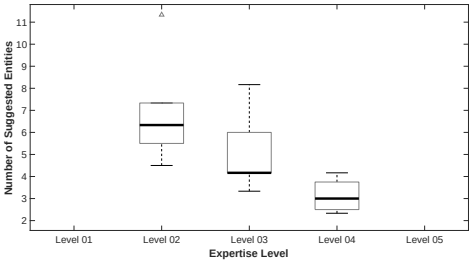

Figure 1: Distribution of the suggested entities with user expert level

and description. Each patent set was given to two users who were provided with a custom computer interface. Users were asked to open the given patents and highlight anchors (a word or a phrase) that they considered to require a hyperlink for better understanding. Usually these anchors described a process, an artifact or a field of study. Participants were asked not to highlight anchors that they could easily understand. When an anchor was nested or had an overlap with another entity, we asked participants to highlight the entity which was more specific and informative. We were able to receive anchors from two different users for every patent in the selected patent set. In order to identify user background including their education and expertise level, all users were asked to answer a questionnaire at the end of the study.

3 DATA ANALYSIS

Table 1 shows the statistics of the collected entities. Interestingly some participants highlighted a very small number of entities to be linked in patents, suggesting strong confidence in their understanding. We categorized the suggested entities according to the users' self-selected expertise level. Although we categorize users over five expertise levels from 1 (lowest expertise level) to 5 (highest expertise level), there were no participants who indicated level 1 or level 5 in our study. The distribution of entities for all patents according to the user expertise level is shown in the Figure 1.

The range of number of entities required by level 2 users is very different from that of level 4 users, while these ranges are overlapping for level 2 and level 3 users. We performed t-tests to compare the averages of number of expected entities for level 2 and level 4 users and t-test analysis showed significant difference between the responses of the two groups ($p < 0.001$).

4 RANKING ALGORITHMS FOR ANCHOR TEXT DETECTION

We explore the performance of four well known ranking algorithms (TFIDF, BM25, Keyphraseness, and Termex [1]) on the selected patents, considering the anchor texts suggested by the user-study participants as ground truth. The literature reveals that the first three algorithms are promising for Wikipedia anchor text detection and Termex is used by Tsunakawa and Kaji for patent anchor text extraction [13, 18]. However none of these systems have been evaluated considering real user suggested anchor texts.

4.1 Candidate Extraction and Ranking

We extracted the alphabetic text from the patent title, abstract, claims and description. Then n-grams, from n=1 to n=5 were extracted from each patent. The extracted n-grams were filtered using a list of surface forms extracted from Wikipedia which was generated by Bryl and others [2]. Using a controlled vocabulary has shown success in the past link discovery [13]. These surface forms are extracted from labels, redirects and disambiguations, and from anchor texts of internal Wikipedia links. We used these surface forms as the controlled vocabulary and the size of the controlled vocabulary is 36,035,294 terms. The filtered n-gram list is considered as the candidate anchor list for a given patent. We used the longest entity when there were overlapped entities. We employed the ranking algorithms (TFIDF, BM25 and Keyphraseness) on the candidate anchor list. Termex is employed in a slightly different way than the other three algorithms. N-gram extraction was not necessary for Termex. Thus each patent text was directly given to Termex for retrieving candidate anchor texts and those candidates were filtered using the list of surface forms.

Ranking algorithms assigned a numeric rank score to each candidate anchor text. This process results in a ranking of potential anchor texts, ordered in decreasing numeric score for each patent in the study. The details of the ranking algorithms used in the study were as follows.

TFIDF- a traditional weighting model used in information retrieval for estimating the importance of a term in a given document. We used TFIDF ranking model implemented in the Terrier IR package [17] for our experiment which is implemented as a combination of the Okapi's TF and Sparck- Jones' IDF.

BM25- Okapi BM25 weighting model implemented in the Terrier IR package with the default settings ($k_1 = 1.2, b = 0.75$).

Keyphraseness- Keyphraseness is a measure introduced by Michalcea and Csomai [13] and it exploits the information contained in already linked Wikipedia articles. The score of Keyphraseness for a given entity is defined as

$$P(Keyword|W) \approx \frac{|D_{key}|}{|D_w|}$$

where $|D_{key}|$ is the number of documents where the term was already selected as a keyword and $|D_w|$ is the total number of documents where the term appeared. In our study we used publicly available Keyphraseness values[2]. These Keyphraseness values are calculated from the English Wikipedia dump created on January 30, 2010 and contained about 1.9 million phrases with non-zero Keyphraseness values.

Termex Termex is a publicly available domain terminology extraction system which was developed by Nakagawa and Mori [15]. Termex takes the given text as input and outputs a list of terms (words and phrases) ranked by a termhood score values. Each score value for a term is calculated based on occurrence and concatenation frequencies of simple and compound nouns [15]. As similar to the approach of Tsunakawa and Kaji, we used the output candidate list with out applying any filtering based on scores [15].

[1] $http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb\_eng.html$

[2] $http://www.ntu.edu.sg/home/axsun/datasets.html$

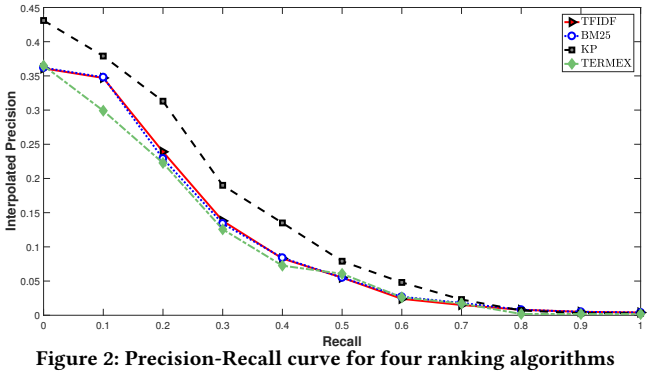

Figure 2: Precision-Recall curve for four ranking algorithms

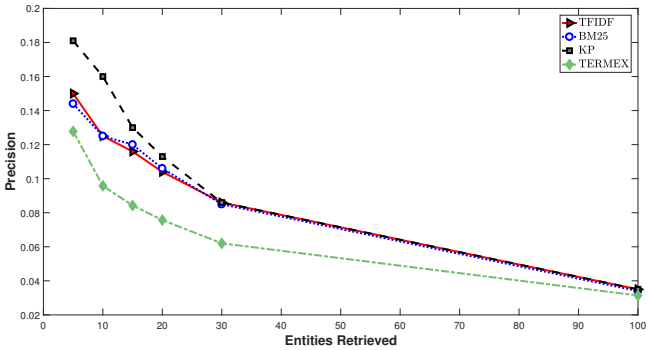

Figure 3: Entity level precision average

4.2 Performance of ranking algorithms

We compare the ranked list retrieved by the algorithm for each patent with the ground truth. The ground truth was defined as the union of the suggested entities by two users.

We measured mean interpolated precision at 11 recall levels of 0.0, 0.1, ..., 1.0 considering the entire dataset. Figure 2 illustrates the composite precision-recall curves for each algorithm. The precision values we retrieved for entity detection is considerably lower than the values received at past link discovery approaches such as INEX [5, 8] . The possible reason for the low precision of ranking algorithms is that we are using user suggested entities as our ground truth which is a very small number of entities compared to the large number of relevant entities used in the earlier experiments, using extensively linked Wikipedia documents. As well, only 293 out of 653 user sugeested entities appear in the controlled vocabulary suggesting that required anchor texts in patent domain are distinct from the Wikipedia domain. Agreeing with prior work Keyphraseness (KP) outperforms TFIDF and BM25. The performance of Termex shows a similar trend to the BM25 and TFIDF (see the Figure 2).

Figure 3 shows how precision changes with the number of entities retrieved for the three algorithms. Here again, KP outperform the other ranking algorithms and we can see the best precision when retrieving only 5 entities. The highest R precision (The Precision after R relevant entities retrieved) score is 0.176 and it was obtained by KP while both TFIDF and BM25 score a value of 0.130. and Termex score a value of 0.110.

Retrieval performance of the algorithms considering the two different set of patents (ME and IT) is shown in Figure 5 and all the ranking algorithms show better performance with the ME patents than with the IT patents. According to Table 1 the ME patents have

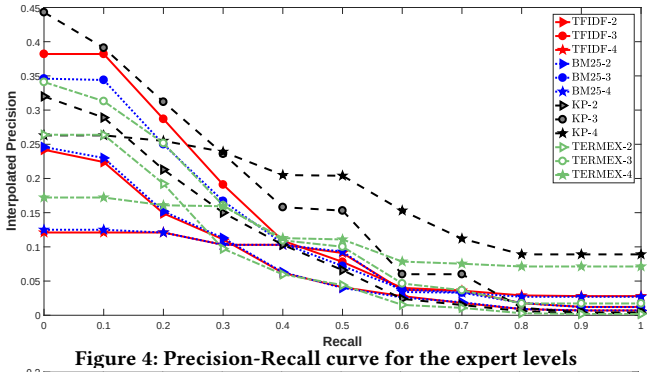

Figure 4: Precision-Recall curve for the expert levels

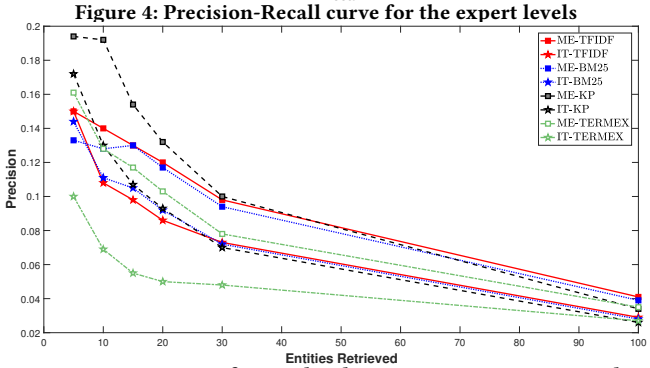

Figure 5: Comparison of entity level precision averages - ME and IT patents

more link-worthy entities than IT patents. A large part of our users are people with a strong IT background and these results are likely to be related with this fact.

We grouped the user suggested entities that were obtained from the user study according to the user-specified expertise levels. We only have participants from 02, 03 and 04 expertise levels. Four participants indicated level 04, eleven participants indicated level 03, and nine indicated level 02. Each participant highlighted entities in six patents. There were 66 patents assessed by level 03, 54 patents assessed by level 02, and 24 patents assessed by level 04. We used the anchors proposed by each participant as the ground truth for that specific patent and again evaluated the performance of three ranking algorithms. Figure 4 illustrates how interpolated precision of three ranking algorithms vary with the recall level for the different user expertise levels that are considered. As shown in the Figure 4, all ranking algorithms have better performance with level 03 users than level 02 users and level 04 users. These results suggests that the ranking algorithms perform better with average expert users than with the users who are above and below average expertise.

4.3 User-User agreement and User-System agreement

As existing user agreement measures including Cohen's Kappa can not be used in the situations where there is no defined baseline agreement and no plausible way to define chance agreement in any consistent manner, we developed our own criteria for measuring user agreement for a given patent.

$$Agreement_{u1,u2} = \frac{|E_{u1} \cap E_{u2}|}{|E_{u1} \cup E_{u2}|}$$

Table 2: Agreements (user-user and user-system)

Agreement	Case 01	Case 02	Case 03
user-user (All)	0.187	0.143	0.166
user-TFIDF (All)	0.129	0.142	0.192
user-BM25 (All)	0.127	0.14	0.189
user-KP (All)	0.168	0.165	0.200
user-TERMEX (All)	0.109	0.108	0.115
user-user (ME)	0.185	0.142	0.197
user-TFIDF (ME)	0.135	0.178	0.206
user-BM25(ME)	0.13	0.141	0.202
user-KP (ME)	0.168	0.188	0.214
user-TERMEX (ME)	0.158	0.154	0.151
user-user (IT)	0.188	0.105	0.131
user-TFIDF (IT)	0.122	0.135	0.175
user-BM25(IT)	0.125	0.139	0.175
user-KP (IT)	0.168	0.135	0.185
user-TERMEX (IT)	0.062	0.050	0.069

Here E_{u1} is the set of entities suggested by user1 and E_{u2} is the set of entities suggested by user2. Agreement between user1 and user2 is calculated by dividing the intersection of both users' suggested link-worthy entities by the union. The agreement between the system and the users was calculated considering the union of user suggested entities as ground truth. Given the number of entities in the union of two users' suggested entities or relevant entities is equal to R, we consider the first R ranks in the retrieval list of each system (ranking algorithm). Then we calculate Precision @ R or R-precision for each system.

The calculated agreement values are illustrated in Table 2 where case 01 contains the values calculated using the complete patent set, case 02 and case 03 are calculated using only patents with more than 5 suggested anchors and more than 10 suggested anchors respectively. The values suggest that each user has different expectations about the entities that should be linked. A possible explanation for this might be the high subjectivity of the task and the difference in expertise levels.

Also user-user agreement does not show any improvement with an increase of user suggestions. What stands out in the table is that the automated anchor suggestion systems based on TFIDF, BM25 and KP have better agreement with users than pairs of user have, in situations where two users expect more than five links for a patent. This behaviour is much more evident in situations where the union of user suggested entities are more than 10. However, it is found that the agreement between Termex and users is very low in all the situations when compared to the other three ranking algorithms.

5 CONCLUSIONS AND FUTURE WORK

This paper explored hyperlink anchor selection in patent documents. We conducted a user study to examine user agreement over which entities should be linked to improve the understandability of patents. To the best of our knowledge this is the first study which conducts a user study to explore anchor text selection for link discovery in the patent domain. In previous studies of link discovery, notably in the INEX Link the Wiki track, user agreement with the ground truth of the Wikipedia, and automated methods, was quite high. Notwithstanding the relatively limited number of users in our study (24) the very low user agreement results in this domain clearly suggest that a personalization component for patent link discovery may be necessary to improve the performance of established methods. Future work will involve a larger number of participants and will be extended to include link disambiguation - the linking of anchors to target resources.

REFERENCES

[1] Doreen Alberts, CynthiaBarcelon Yang, Denise Fobare-DePonio, Ken Koubek, Suzanne Robins, Matthew Rodgers, Edlyn Simmons, and Dominic DeMarco. 2011. Introduction to Patent Searching. In *Current Challenges in Patent Information Retrieval*, Mihai Lupu, Katja Mayer, John Tait, and Anthony J. Trippe (Eds.). The Information Retrieval Series, Vol. 29. Springer Berlin Heidelberg, 3–43. https://doi.org/10.1007/978-3-642-19231-9_1

[2] Volha Bryl, Christian Bizer, and Heiko Paulheim. 2015. Gathering Alternative Surface Forms for DBpedia Entities.. In *NLP-DBPEDIA@ ISWC*. 13–24.

[3] Hakan Ceylan, Ioannis Arapakis, Pinar Donmez, and Mounia Lalmas. 2012. Automatically Embedding Newsworthy Links to Articles. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*. ACM, New York, NY, USA, 1502–1506. https://doi.org/10.1145/2396761.2398461

[4] James J. Gardner and Li Xiong. 2009. Automatic Link Detection: A Sequence Labeling Approach. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*. ACM, New York, NY, USA, 1701–1704. https://doi.org/10.1145/1645953.1646208

[5] Shlomo Geva, Darren Huang, Andrew Trotman, and Yue Xu. 2008. Overview of INEX 2007 Link the Wiki Track. In *Focused Access to XML Documents: Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007*, N Fuhr, J Kamps, M Lalmas, and A Trotman (Eds.). Springer, Germany, Saarland, Dagstuhl Castle, 373–387. http://eprints.qut.edu.au/30560/

[6] Stephen Guo, Ming-Wei Chang, and Emre Kiciman. 2013. To Link or Not to Link? A Study on End-to-End Tweet Entity Linking. In *NAACL-HLT 2013*. https://www.microsoft.com/en-us/research/publication/to-link-or-not-to-link-a-study-on-end-to-end-tweet-entity-linking/

[7] Jiyin He and Maarten de Rijke. 2010. *An Exploration of Learning to Link with Wikipedia: Features, Methods and Training Collection*. Springer Berlin Heidelberg, Berlin, Heidelberg, 324–330. https://doi.org/10.1007/978-3-642-14556-8_32

[8] Wei Che (Darren) Huang, Shlomo Geva, and Andrew Trotman. 2009. *Overview of the INEX 2008 Link the Wiki Track*. Springer Berlin Heidelberg, Berlin, Heidelberg, 314–325. https://doi.org/10.1007/978-3-642-03761-0_32

[9] Kelly Y. Itakura and Charles L. A. Clarke. 2009. *University of Waterloo at INEX 2008: Adhoc, Book, and Link-the-Wiki Tracks*. Springer Berlin Heidelberg, Berlin, Heidelberg, 132–139. https://doi.org/10.1007/978-3-642-03761-0_14

[10] Abhik Jana, Sruthi Mooriyath, Animesh Mukherjee, and Pawan Goyal. 2017. WikiM: Metapaths based Wikification of Scientific Abstracts. *CoRR* abs/1705.03264 (2017). http://arxiv.org/abs/1705.03264

[11] Dylan Jenkinson, Kai-Cheung Leung, and Andrew Trotman. 2009. *Wikisearching and Wikilinking*. Springer Berlin Heidelberg, Berlin, Heidelberg, 374–388. https://doi.org/10.1007/978-3-642-03761-0_38

[12] Hideo Joho, Leif A. Azzopardi, and Wim Vanderbauwhede. 2010. *A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements*. 13–24. https://doi.org/10.1145/1840784.1840789

[13] Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking Documents to Encyclopedic Knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (CIKM '07)*. ACM, New York, NY, USA, 233–242. https://doi.org/10.1145/1321440.1321475

[14] David Milne and Ian H. Witten. 2008. Learning to Link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*. ACM, New York, NY, USA, 509–518. https://doi.org/10.1145/1458082.1458150

[15] Hirosi Nakagawa and Tatsunori Mori. 2003. Automatic term recognition based on statistics of compound nouns and their components. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 9, 2 (2003), 201–219. https://doi.org/10.1075/term.9.2.04nak

[16] Lisa Larrimore Ouellette. 2012. Do patents disclose useful information? *Harvard Journal of Law & Technology* 25 (2012).

[17] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. 2006. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*.

[18] Takashi Tsunakawa and Hiroyuki Kaji. 2015. Towards Cross-lingual Patent Wikification. *Proceedings of 6th Workshp on Patent and Scientific Literature Translation (PSLT6) Miami* (2015), 89.