An Initial Study of Anchor Selection in Patent Link Discovery

Dilesha Seneviratne
Queensland University of Technology
Brisbane, Australia

Shlomo Geva
Queensland University of Technology
Brisbane, Australia

Guido Zuccon
Queensland University of Technology
Brisbane, Australia

Andrew Trotman
University of Otago
Dunedin, Newzealand

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 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 and 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 to link.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; **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, microblogs, etc.) [3, 6], 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 elsewhere [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., researchers - to learn about existing technologies, inventors - to ensure their idea is novel [12]. All these create the need for technological solutions to facilitate the comprehension of the patent content.

Anchor selection is a key process in link discovery. While it is possible to link all text to relevant information, it has some disadvantages: (1) Finding accurate targets for all words in the text poses a computational burden on the link discovery system; (2) Over linking should be avoided as it does not aid user understanding[1] and it would overly complicate the interface. 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. However, unsupervised methods are applicable to collections without prior links. 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 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 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 anchor text selection; explored methods include: Naive Bayes, Decision Trees, Conditional Random Fields and Support Vector Machines [7, 14]. Their disadvantage is they are highly dependent on both the 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 [18] is the only study that attempted patent link discovery. They have employed a domain terminology extraction system known as Termex [15]. In contrast to our system, where users-suggested anchor texts are used as ground truth, they have considered the existing anchor texts in Wikipedia as ground truth when evaluating their system. The main drawback of this type of evaluation is that 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. The 72 patents were from two major patent sections of Mechanical Engineering (ME) and Information Technology (IT), 36 from

[1] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking

ADCS 2017, December 7–8, 2017, Brisbane, QLD, Australia

ACM ISBN 978-1-4503-6391-4/17/12…$15.00
https://doi.org/10.1145/3166072.3166078

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

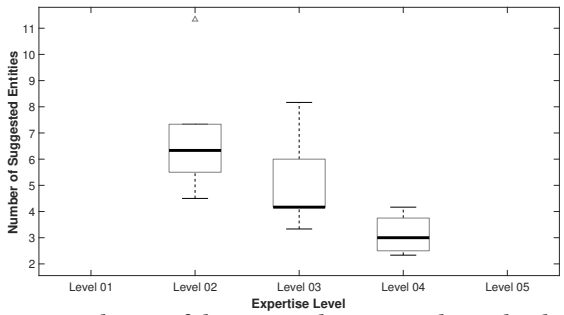

Figure 1: Distribution of the suggested entities with user level.

each section. These domains were selected because we could gain access to users with suitable domain expertise who represented a reasonable selection of potential readers. The objective of our work was to identify link-worthy entities that required 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, 21 were PhD students (14 - Computer Science, 7 - other Engineering disciplines). The other 3 were Engineering professionals. 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 included the extracted text from the sections of title, abstract, claims 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 categorized users over five expertise levels from 1 (lowest expertise level) to 5 (highest), 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 was very different from that of level 4 users, while these ranges were 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 [2]) on the selected patents, considering the anchor texts suggested by the user-study participants as ground truth. The literature revealed that the first three algorithms are promising for Wikipedia anchor text detection and Termex was used by Tsunakawa and Kaji for patent anchor text extraction [13, 18]. However none of these systems have been evaluated considering 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 and DBpedia which was generated by Bryl et al. [2]. Using a controlled vocabulary has shown success in past link discovery work [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 was 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 was employed in a slightly different way compared to 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 resulted 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 experiments which is implemented as a combination of the Okapi's TF and Sparck- Jones' IDF.

BM25: Okapi BM25 weighting model implemented in the Terrier 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

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

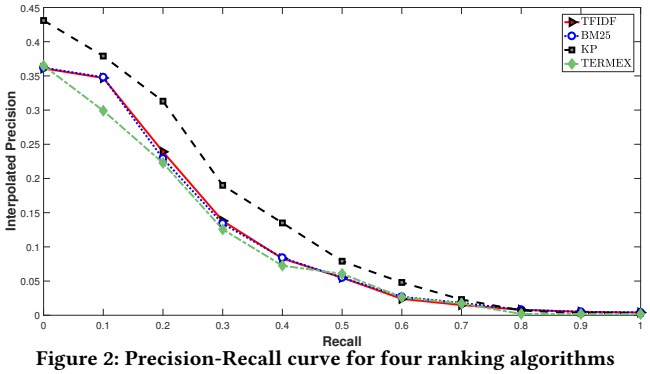

Figure 2: Precision-Recall curve for four ranking algorithms

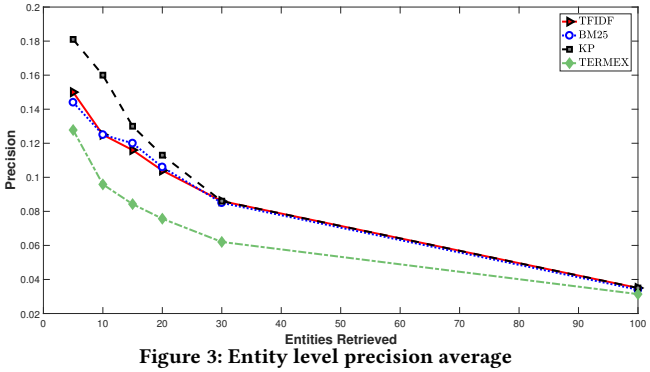

Figure 3: Entity level precision average

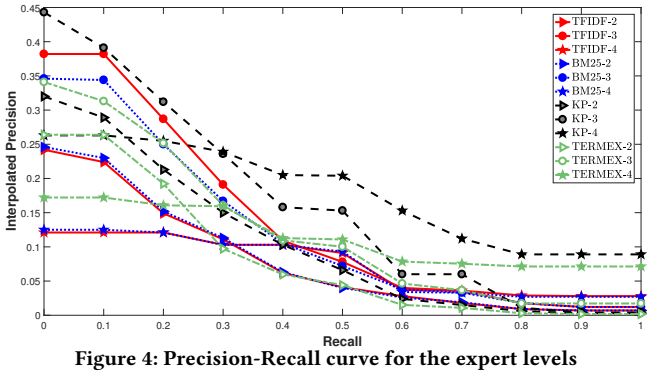

Figure 4: Precision-Recall curve for the expert levels

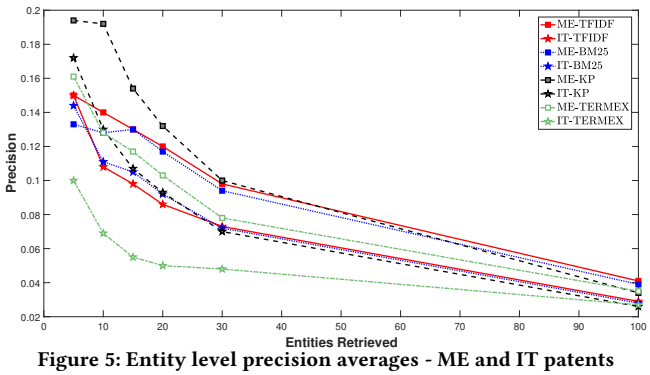

Figure 5: Entity level precision averages - ME and IT patents

where the term appeared. In our study we used publicly available Keyphraseness values[3]. These Keyphraseness values were 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]. Similar to the approach of Tsunakawa and Kaji, we used the output candidate list without applying any filtering based on scores [18].

4.2 Performance of ranking algorithms

We compared the ranked list retrieved by the algorithms 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 considering the entire dataset. Figure 2 illustrates the composite precision-recall curves for each algorithm. The precision values for entity detection was considerably lower than the values received in past link discovery evaluations such as INEX [5, 8] . A possible reason for the low precision of ranking algorithms is that user-suggested entities were used as ground truth. This is a small number of entities compared to the large number of relevant entities used in earlier experiments, which used extensively linked Wikipedia documents. We also found that only 284 out of 653 user-suggested entities appeared in the controlled vocabulary. This suggests that the anchor texts required in the patent domain are distinct from the Wikipedia domain. Agreeing with prior work, Keyphraseness (KP) outperformed TFIDF and BM25. The performance of Termex showed 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 four algorithms. Here again, KP outperformed the other ranking algorithms and the highest precision was achieved when retrieving only 5 entities. The highest R-precision (the precision after R relevant entities retrieved) was 0.176 and it was obtained by KP while both TFIDF and BM25 scored a value of 0.130 and Termex scored a value of 0.110.

Retrieval performance of the algorithms considering the two different sets of patents (ME and IT) is shown in Figure 5. All ranking algorithms exhibited better performance with the ME patents than with the IT patents. According to Table 1 the ME patents had more link-worthy entities than IT patents. A large part of our users in our experiments were people with a strong IT background and these results are likely to be related to this fact.

We grouped the user-suggested entities that were obtained from the user study according to the user-specified expertise levels. We only had 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 the four ranking algorithms. Figure 4 illustrates how interpolated precision of the ranking algorithms varied with the recall level for the different user expertise levels considered. As shown in Figure 4, all ranking algorithms had better performance with level 03 users than

[3] *http://www.ntu.edu.sg/home/axsun/datasets.html*

level 02 and level 04 users. These results suggests that the ranking algorithms performed 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

We gave free text to the users to suggest anchor texts and thus the size of the potential anchor list was not fixed. 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:

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

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 was 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 considered the first R ranks in the retrieval list of each system (ranking algorithm). Then we calculated Precision @ R or R-precision for each system.

The calculated agreement values (average and standard deviation) are illustrated in Table 2, where case 01 refers to the use of the complete patent set. Case 02 and case 03 values were calculated using only patents with more than 5 suggested anchors and more than 10 suggested anchors, respectively. The values suggest that each user had 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 did 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 had higher agreement with users than pairs of user had, in situations where two users expected more than five links for a patent. This behaviour is much more evident in situations where the union of user suggested entities were more than 10. However, it was found that the agreement between Termex and users was very low in all the situations when compared to the other three ranking algorithms.

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

Agreement	Case 01		Case 02		Case 03	
	Avg	Std	Avg	Std	Avg	Std
User-User (All)	0.19	0.19	0.14	0.12	0.16	0.10
User-TFIDF (All)	0.13	0.12	0.14	0.12	0.19	0.11
User-BM25 (All)	0.13	0.11	0.14	0.11	0.19	0.11
User-KP (All)	0.17	0.15	0.16	0.13	0.20	0.10
User-TERMEX (All)	0.11	0.12	0.11	0.10	0.11	0.08
User-User (ME)	0.19	0.16	0.17	0.14	0.19	0.11
User-TFIDF (ME)	0.14	0.12	0.15	0.11	0.21	0.10
User-BM25 (ME)	0.13	0.11	0.14	0.11	0.20	0.10
User-KP (ME)	0.17	0.14	0.19	0.13	0.21	0.09
User-TERMEX (ME)	0.16	0.12	0.15	0.10	0.13	0.07
User-User (IT)	0.19	0.22	0.10	0.09	0.13	0.06
User-TFIDF (IT)	0.12	0.12	0.13	0.12	0.17	0.12
User-BM25 (IT)	0.12	0.12	0.13	0.12	0.17	0.12
User-TERMEX (IT)	0.17	0.16	0.13	0.13	0.18	0.11
User-TERMEX (IT)	0.06	0.10	0.05	0.07	0.67	0.07

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] D. Alberts, C. Yang, D. Fobare-DePonio, K. Koubek, S. Robins, M. Rodgers, E. Simmons, and D. DeMarco. 2011. Introduction to Patent Searching. In *Current Challenges in Patent Information Retrieval.* Vol. 29. 3–43.

[2] V. Bryl, C. Bizer, and H. Paulheim. 2015. Gathering Alternative Surface Forms for DBpedia Entities. In *NLP-DBPEDIA*. 13–24.

[3] H. Ceylan, I. Arapakis, P. Donmez, and M. Lalmas. 2012. Automatically Embedding Newsworthy Links to Articles. In *Proc. of CIKM'12.* 1502–1506.

[4] J. J. Gardner and L. Xiong. 2009. Automatic Link Detection: A Sequence Labeling Approach. In *Proc. of CIKM'09.* 1701–1704.

[5] S. Geva, D. Huang, A. Trotman, and Y. Xu. 2008. Overview of INEX 2007 Link the Wiki Track. In *Proc. of INEX'08.* 373–387.

[6] S. Guo, M. Chang, and E. Kiciman. 2013. To Link or Not to Link? A Study on End-to-End Tweet Entity Linking. In *NAACL-HLT'13.*

[7] J. He and M. de Rijke. 2010. An Exploration of Learning to Link with Wikipedia: Features, Methods and Training Collection. In *Proc. of INEX'09.* 324–330.

[8] D. Huang, S. Geva, and A. Trotman. 2009. Overview of the INEX 2008 Link the Wiki Track. In *Proc. of INEX'09.* 314–325.

[9] K. Itakura and C. Clarke. 2009. University of Waterloo at INEX 2008: Adhoc, Book, and Link-the-Wiki Tracks. In *Proc. of INEX'09.* 132–139.

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

[11] D. Jenkinson, K. Leung, and A. Trotman. 2009. Wikisearching and Wikilinking. In *Proc. of INEX'09.* 374–388.

[12] H. Joho, L. Azzopardi, and W. Vanderbauwhede. 2010. A survey of patent users: an analysis of tasks, behavior, search functionality and system requirements. In *Proc. of IIiX'10.* 13–24.

[13] R. Mihalcea and A. Csomai. 2007. Wikify!: Linking Documents to Encyclopedic Knowledge. In *Proc. of CIKM'07.* 233–242.

[14] D. Milne and I. Witten. 2008. Learning to Link with Wikipedia. In *Proc. of CIKM'08.* 509–518.

[15] H. Nakagawa and T. Mori. 2003. Automatic term recognition based on statistics of compound nouns and their components. *Terminology* 9, 2 (2003), 201–219.

[16] L. 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 *Proc. of OSIR'06.*

[18] T. Tsunakawa and H. Kaji. 2015. Towards Cross-lingual Patent Wikification. In *Proc. of PSLT'6.* 89.