Generating Links for Patent Documents: an Automatic Approach using Computational Intelligence

C. M. Souza, M. E. Santos, M. R. G. Meireles

Pontifical Catholic University of Minas Gerais, Brazil
cinthia.mikaela@sga.pucminas.br
matheus.santos.1004060@sga.pucminas.br
magali@pucminas.br

Abstract. Patents are organized into classification systems, which assist offices and users in the process of seeking and retrieving such documents. A wide variety of users use the patent systems and the information contained in these documents. However, patents are complex legal documents with a significant number of technical and descriptive details, which makes it difficult to identify and analyze the information contained in these documents. An automatic link system associated with some of the terms found in the patents would provide quick access to the concepts contained in specific knowledge bases. This work presents results of a project in which the objective is the automatic generation of links in patent documents. The experiments were conducted with four subgroups of the United States Patent and Trademark Office (USPTO), which uses the Cooperative Patent Classification (CPC) system. As the patent documents did not have keywords, the meaningful terms were selected using the algorithm χ^2, for which the contents of the entire patent document were used. Some keywords with more than one meaning were disambiguated using a specific algorithm, generating a file with useful information used in the experiments. The links were generated based on Wikipedia articles and the USPTO patent database. The use of the patent database as a possible destination for the link is intended to cover cases in which Wikipedia has no articles on certain terms and also to provide an alternative source that may assist readers in understanding those documents. It is expected, with the creation of automated links, to make it easier to access concepts related to the terms presented by the documents and to understand the information disclosed by the inventors.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

Keywords: Word Sense Disambiguation, Keywords Extraction, Link creation, Patents

1. INTRODUCTION

Patents are an important knowledge source and, therefore, their analysis has been considered a useful tool for research and for management development. Ouellette [2017] conducted a survey of 832 researchers to assess the importance of patent study. Most researchers in different fields of knowledge have stated that they have found useful information in the documents, but acknowledged that there is room for improvement, particularly regarding the accessibility and understanding of information contained in patents. Many of the interviewees stated that it is possible to find information unavailable in the scientific literature and that patents are an underutilized complement to the dissemination of scientific knowledge.

In general, patent documents are divided into sections, defined by patent offices. The main sections of these documents are, title, abstract, claims and descriptions. Each of these sections have par-

The authors thank the financial support of the Pontifical Catholic University of Minas Gerais, the National Council for Scientific and Technological Development (CNPq, grant 429144/2016-4) and the Foundation for Research Support of the State of Minas Gerais (FAPEMIG, grant APQ 01454-17).

ticular characteristics. Abstracts are characterized by their complex syntactic structure and generic vocabulary. The claims have more specific information of the invention. Usually, the language used is more formal, and sometimes more obscure, than everyday language. The description section has more distinctive information describing the features of the invention [Codina-Filbà et al. 2017; Durham 2018].These characteristics make it difficult, in many cases, to consult and understand the information, thus generating an underutilization of patent knowledge [Seneviratne 2018].

The increase in the number of patenting processes along with the need to access the information contained in the patent databases motivate researchers to develop efficient techniques for extracting knowledge from such documents [Khode and Jambhorkar 2017]. One of the factors hindering the process of analyzing and extracting knowledge from the information contained in patents is related to the fact that patents are complex documents with technological details, legal language and exhaustive descriptions [Meireles et al. 2016]. In this context, conventional approaches for information retrieval are difficult to apply. The in-depth study of patents has yet to become more accessible, by identifying potential areas of research for the scientific community and generating useful information in the processes of decision-making in the area of competitive intelligence. The proposal of creating an automatic link generation system, which allows simplified access to the concepts related to the terms presented by patents, solves this problem, beckoning with the possibility of quickly accessing the knowledge base related to the theme proposed by the patent.

The automatic determination of a link includes the identification of possible text fragments that should be associated with knowledge bases. In most cases, keywords are selected and their extraction can be done by supervised or unsupervised methods [Mihalcea and Csomai 2007]. Before selecting the texts that will be associated with certain words, it is necessary to define the context in which the word is being used. This is because several language units have different meanings, which is a common feature in many languages and a problem that needs to be addressed in depth.

This article presents results of a project that pursues the automatic generation of links in patent documents. Reginaldo et al. [2017] state, in their work, that the algorithm χ^2 achieved the best results for keywork extraction in this context. To achieve a more satisfactory performance of the keyword extraction process, all sections of the patent were explored. This destination of the links are generated using two separate databases, the Wikipedia article database and the USPTO patent database.

This work was divided into 5 sections. Section 2 presents the main concepts used, as well as a description of the algorithms implemented in the processes of keyword extraction of links and Word Sense Disambiguation (WSD). Section 3 presents the proposed methodology, the used database and the methodological steps of the work. Sections 4 and 5 show the results, analysis and final considerations.

2. AUTOMATIC LINK GENERATION

The automatic link generation process can be divided into three distinct steps: identifying the source of a link, determining the appropriate document to be associated with the terms selected as the source, and generating the link in the document. In the second step, we must solve the problem of WSD. Given this, this section will be divided into four subsections. The first three sections address the main concepts of the work and the last one presents some related works.

2.1 Keywords Extraction

Keywords are a set of relevant terms that sufficiently describe a given text document. Because it is considered a compact form of document representation, keywords are often used in the task of locating information in large databases. Due to these characteristics, keywords are considered an important resource in text mining tasks, natural language processing and information retrieval [Onan et al. 2016; Duari and Bhatnagar 2019]. In some cases, as in patent documents, the keywords are not defined by

the authors of the document and therefore it is necessary to develop a method or select an algorithm that extracts the words considered significant for the document and can represent it in a system of information retrieval.

The χ^2 algorithm, defined by Manning and Schütze [1999], evaluates the independence between two variables and compares observed and expected values, evaluating how far apart they are. According to Manning and Schütze [1999] the chi-squared statistic sums the difference between observed and expected values in all squares of the table, scaled by the magnitude of the expected values. This algorithm is used to order the words according to their dependence on the patent, so that the greater the score given χ^2 to a word, the greater its dependence on the document. Even if the algorithm accepts that a word is independent, the note given by it to the word is simply added to the ordering in lower positions [Reginaldo et al. 2017]. The χ^2 is defined by Equation 1:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{1}$$

where, O_{ij} is the observed value, which represents the number of occurrences of the word in the document, i is the line and j is the column, and E_{ij} is the expected value calculated using Equation 2 and Table I:

$$E_{ij} = \frac{\sum(column_j) \times \sum(line_i)}{\sum(total)}. \tag{2}$$

Table I. Contingency Table

Words/Documents	P_1	...	P_m	
$Word_1$	Occurrences of the $Word_1$ in Document P_1	...	Occurrences of the $Word_1$ in Document P_m	Sum (line)
.	.		.	.
.	.	..	.	.
.	.		.	.
$Word_n$	Occurrences of the $Word_n$ in Document P_1	...	Occurrences of the $Word_n$ in Document P_m	Sum (line)
	Sum (column)	...	Sum (column)	Sum (Internal cells)

Source: Adapted from Mihalcea and Csomai [2007]

2.2 Word Sense Disambiguation

The techniques of WSD aim to computationally identify the meaning of a word, taking into account the context in which it is inserted. Therefore, given a document with a sequence of words $T = w_1, w_2, ..., w_n$, this technique aims to give meaning to all or some words of that document. This task can be performed with only one lexical sample or with all the words in the document. Generally, the lexical sample is more used, because a wide coverage of domains is necessary to carry out the sense disambiguation of all the words [Corrêa Jr et al. 2018].

JobimText is a WSD algorithm proposed by Panchenko et al. [2017]. This algorithm receives, as input, the word that will be disambiguated and the context to which it belongs. In addition, it is necessary to define some parameters, such as the model used and the output format. Among the

available models, the ensemble model is the most complete, since it searches for the meaning of the ambiguous word in the inventory of word meanings. If a word is outside this vocabulary, then it is disambiguated by using the super meaning inventory. This template was created from a text corpus which is a combination of Wikipedia, ukWaC, corpus LCC News and Gigaword. For the realization of WSD, this algorithm performs the following steps:

—Extraction of context features computing word and feature similarities;
—Word meaning induction;
—Labeling of clusters with hypernyms and images (hypernym is a word with a broad meaning constituting a category in which words with more specific meanings fall [2]);
—Sense disambiguation of words in context based on the induced inventory.

In the end, the algorithm returns the meaning of the word, its hypernyms, the set of related words taken from the dictionary, and a set of phrases to exemplify the meaning of the disambiguated word. The context words are those that co-occur with the word ambiguous destination in the given meaning, and they are also returned with the words related to the disambiguated word, taken from the text itself, and the level of trust of the WSD. The confidence level is a metric that evaluates the WSD result. It is calculated from the extraction of hypernyms. For this, the algorithm ranks the hypernyms using functions that relate the word to the set of words of the cluster and to the hypernym. Figure 1 shows a diagram with the steps of the WSD algorithm.

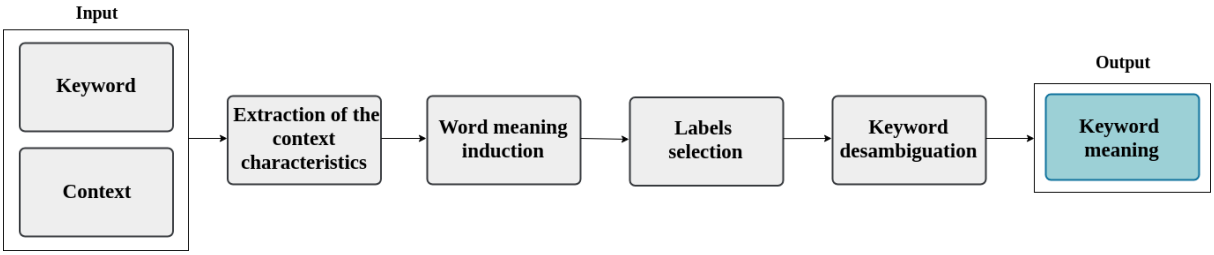

Fig. 1. Steps of the WSD process

2.3 Identification of the link's destination

Identification of the link's destination consists in finding and selecting appropriate documents to be used as the destination for the link [Erbs et al. 2011]. There are different approaches that intend to solve this problem, for example by using the text content, the title of the document and also the links present in the possible destination documents [Seneviratne 2018].

Many papers use the Wikipedia article database as the destination for the link, due to its large number of articles and also the quality of these articles in technical domains. The task of associating terms with Wikipedia articles is called Wikification [Tsunakawa and Kaji 2015]. However, in some cases, it may occur that the Wikipedia database was not sufficient, and it becomes necessary to use other alternative databases.

2.4 Related Works

Linking web data to relevant knowledge base articles has become popular. Some studies on automatic linking of text to important database has captured the interest of the research community [Gardner

[2]en.oxforddictionaries.com

and Xiong 2009]. The discovery of automatic links has been widely discussed in the literature and has comprehensive solutions in Wikipedia texts and scientific articles [Seneviratne 2018]. The majority of the work is focused in linking Wikipedia texts to their referent Wikipedia pages [Mihalcea and Csomai 2007; Cucerzan 2007; Jana et al. 2017]. Han et al. [2011] and Ratinov et al. [2011] focus on generating links in documents that already have the keywords defined. However, in some documents, keywords are not previously defined, such as in patents. Thus, it is necessary to use techniques capable of extracting these words. Works such as of Mihalcea and Csomai [2007], Erbs, Zesch and Gurevych [2011] and Jana et al. [2017] are described in this subsection, and provide a good perspective on how the automatic link generation task is explored.

Mihalcea and Csomai [2007] used Wikipedia for the automatic extraction of keywords and for the WSD process. The system developed by the authors automatically extracts the keywords, makes the WSD process, and generates the link with the Wikipedia page. For the extraction of the keyword, three methods were tested, tf-idf, χ^2, and Keyphraseness. For the keyword sense disambiguation, the authors tested two methods. The first one, a knowledge-based approach, and the other one, based on data-driven. In the end, the Wikify! system presented superior results compared to the competitive baselines.

Erbs, Zesch and Gurevych [2011] evaluated, in their work, the performance of the link discovery task, for Wikipedia database, using the content of the text, the title of the document and also the links present in the possible destination documents. In the end, the authors concluded that, in documents that have a large amount of links added manually, the approach based on links has a good result. However, most documents do not have a large amounts of links. In these cases, the approach using text content performs better.

Jana et al. [2017] presented a project to generate links in abstracts of scientific documents with Wikipedia articles. They performed the extraction of the important mentions of the scientific text using tf-idf, together with a set of intelligent filters. Afterwards, for each mention, they extracted a list of candidate entities (Wikipedia links). These entities were classified and punctuated according to their similarity, and finally, based on this score, the entity for link generation was selected. The results show that the methodology used helps to improve the performance of the wikification task in scientific articles.

Although the task of link generation is widely discussed, it is still little explored in the domain of patent knowledge. There are currently several patent databases and search systems that aim to facilitate the search for information in these documents. However, none of them offer features such as Wikipedia's to facilitate understanding of these documents. Moreover, the fact that patents do not have previously defined keywords is a major problem in this task, since there is no great discussion in the literature about the performance of keyword extraction methods in the patent domain. There is a gap between existing technology solutions and desired solutions to make patent knowledge accessible [Seneviratne 2018].

3. METHODOLOGY

This section presents the database created for the experiments and the description of the methotological steps.

3.1 Database

The database used in the experiment is provided by the United States Patent and Trademark Office (USPTO), whose classification system is the Cooperative Patent Classification (CPC). CPC classifies patents into sections, classes, subclasses, groups, and subgroups. For this work, four subgroups, G06K 7/1443, G06K 7/1447, G06K 7/1452 and G06K 7/1456 of the G06K subclass, named "recognition of

data, presentation of data, record carriers, handling record carriers", were randomly selected. Figure 2 illustrates the hierarchical organization of CPC system. The subgroups used in this work are represented by their suffixes, 43, 47, 52, 56, and were highlighted in Figure 2. These subgroups are at the lowest level of the CPC hierarchy.

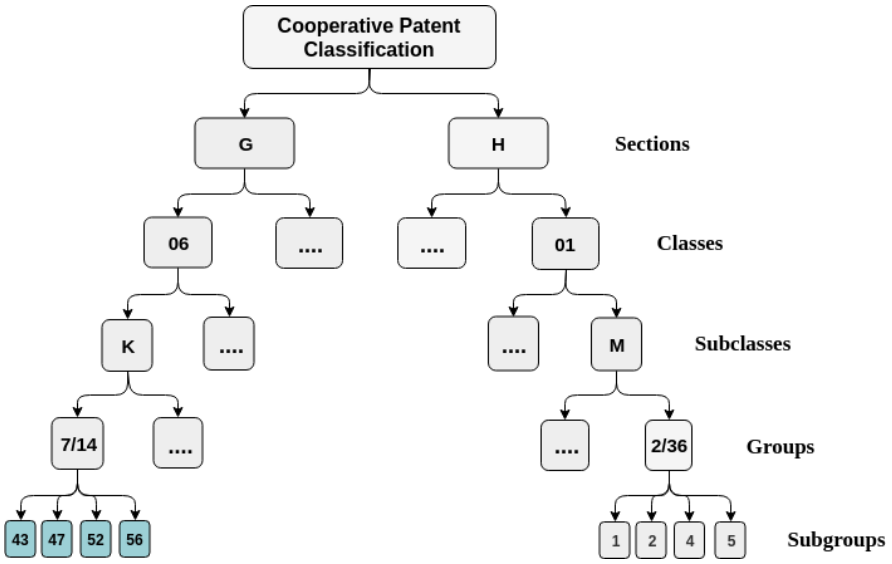

Fig. 2. Hierarchical organization of CPC system

The used database is composed of 999 patents and was updated on April 30, 2019. To validate the methodology in this work, 10 patents of each subgroup were selected. Table II shows the name of the subgroup in the CPC classification system and the distribution of the patents in each of them.

Table II. Patent database

CPC Codes	Number of patents
G06K 7/1443	490
G06K 7/1447	297
G06K 7/1452	82
G06K 7/1456	130

3.2 Methodological Steps

The proposed methodology is presented in Figure 3, showing the main steps of link generation. The first step is the preprocessing of the patent document. The used algorithm performs the removal of stopwords, special characters and the stemization of words. In addition, the algorithm uses a vocabulary based on Wikipedia titles to generate significant n-grams. Thus, only the words which are titles of Wikipedia articles are identified as n-grams, unlike those generated by algorithms that consider only the number of occurrences of two or more words together. This algorithm receives as input a patent document and returns a matrix of occurrences, which contains information on how many times a given word occurs in the document. All sections of the document were used to generate this matrix.

After the preprocessing has been executed, there is an array of documents by words where the occurrence of the words in the document is computed. The second step consists of extracting keywords

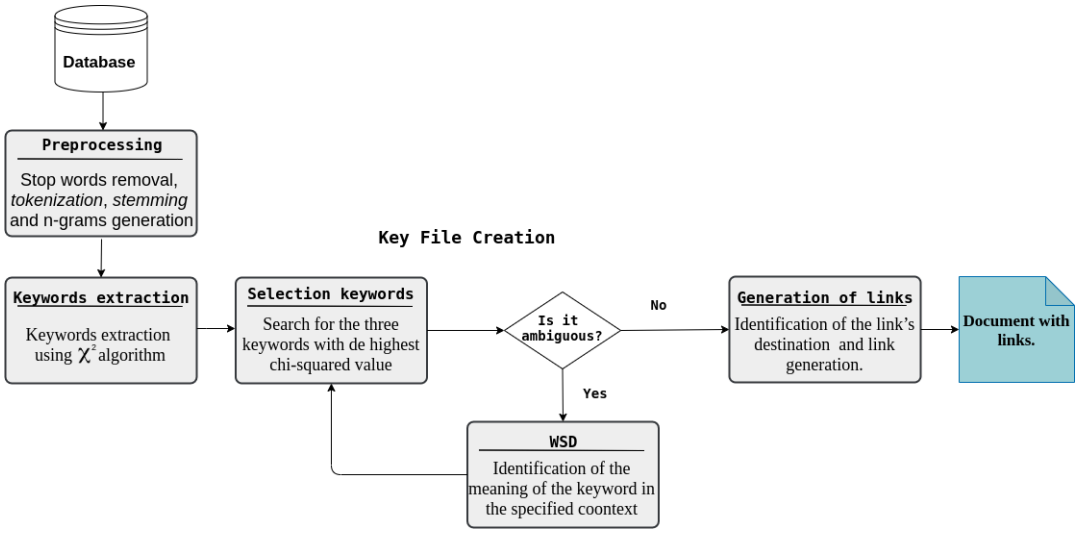

Fig. 3. Proposed methodology

using the algorithm χ^2. This algorithm receives as input the matrix generated by the preprocessing and returns a list of keywords.

The third step creates the key files for each one of subgroups used in the experiments. Each key file consists of the patent identification code, three keywords with the highest χ^2 index, along with their respective index of χ^2 and the meaning of the keyword if it is ambiguous. For keywords that have ambiguous meaning, the first four hypernyms were ordered and stored according to their reliability index. To identify if the keyword has ambiguous meaning, the Wikipedia API was used. Key files are generated using unprocessed keywords as they are found in the original text. So, after finding the keywords for each preprocessed text, the keywords are selected from the unprocessed text.

The WSD is performed, if necessary, in the fourth step. The used WSD algorithm receives the keyword and a paragraph from the patent text that owns the keyword. Both inputs are not preprocessed. In this case, this paragraph is the context that the algorithm used to find the meaning of the word. For this work, the ensemble model was chosen. This algorithm returns the meaning of the word and four hypernyms. The reliability of the WSD process is then verified with the use of the metric provided by the algorithm. The trust value (0 to 100%) indicates whether the meaning of the keyword and the content of the page selected for the link's destination are associated with the same context.

In the fifth step, the link generation was performed. Two distinct databases were used in this task, the Wikipedia article database, and the USPTO patent database presented in Table II. From the list of keywords generated in the second methodological step, the three keywords with the highest value of χ^2 were extracted for each document. To identify the destination of the link, initially, the implemented algorithm receives a keyword and searches for a Wikipedia page with the corresponding content and returns the page link. This algorithm uses a Python library called "wikipedia", which encapsulates the MediaWiki API[1] for this purpose. This library allows access to Wikipedia data and metadata via API. In order to access this data, the user provides an input data and the algorithm provides a Wikipedia page. However, in some cases, ambiguity in the meaning of the keyword may occur. In this case, the algorithm makes an exception, notifying that there is more than one content-related page. For the treatment of the exception, we use the keyword meaning in the document, present in the key file created in the third methodological step. Using the meaning, the search for the keyword at the base of Wikipedia is carried out again. If a page that has the keyword in the specified meaning is still

[1]Application Programming Interface

not found, the algorithm uses the other meaning returned by the WSD algorithm. At the end, if no page is found, the algorithm returns, stating that it was not possible to find a Wikipedia page.

To identify the patent document used as the destination for the link, the implemented algorithm searches, in the key file, the patents that contain the source keyword of the link with the same meaning. In some cases, more than one patent meets this criterion. In this case, the algorithm chooses the patent, in which the selected keyword has the highest value of χ^2. The algorithm starts from the principle that the best destination for the link is the patent that has the keyword with the same meaning and the highest value of χ^2. Keywords in one of the four subgroups only links to those in the other three subgroups. As the proposed work is focused on the lowest level of the CPC hierarchy, when creating a link with a document of the same subgroup, the generated link can target a very similar document, not adding value to the explanation of the keyword.

After defining the link destination patent, a code in Python language was implemented using Web Scraping techniques to access the USPTO database and to choose for the link of the selected patent. For the effective creation of the link in the patent document, the keyword is replaced by the URL in HTML code of the page. Figure 4 presents a diagram of the algorithm implemented to link the patent with the two selected database.

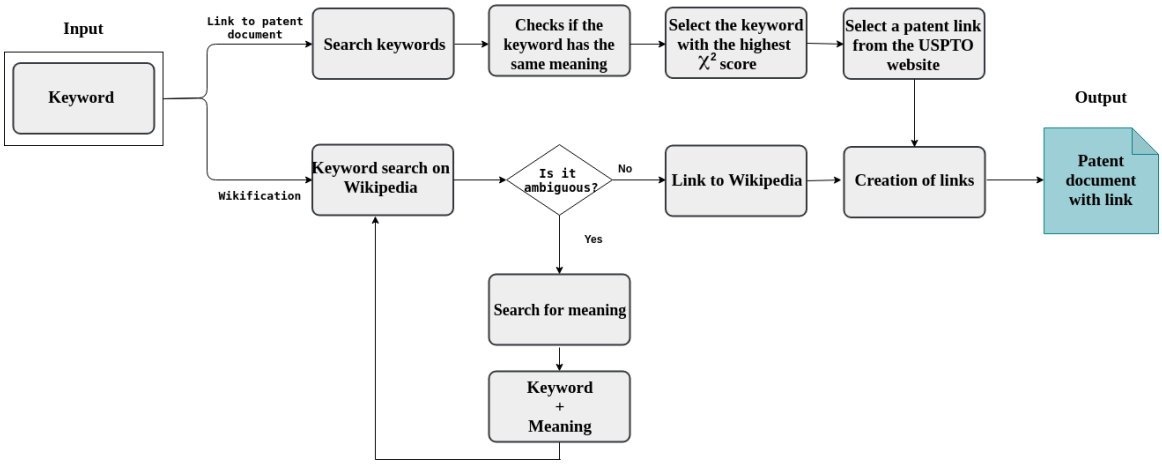

Fig. 4. Identification of the link's destination algorithm steps

4. RESULTS AND DISCUSSION

In the majority of the cases, the proposed methodology was able to correctly identify the Wikipedia pages associated with the keyword origin of the link, when it had no ambiguous meaning. As the vocabulary used for preprocessing is from Wikipedia titles, all extracted keywords represent a page. This makes it easier to correctly identify the destination of the link. However, when the keyword has ambiguous meaning, it is necessary to identify a page that is able to explain the term according to the specified meaning. In this case, the identification of the correct destination for the link is directly associated with the reliability of the WSD algorithm. Therefore, the better the performance of the WSD algorithm, the better is the result obtained by the link destination identification algorithm. It is worth mentioning that identifying the correct link for the keyword is not only related to identifying a page that has the title equal to keyword, but, rather to identifying a page that somehow addresses keyword-related content, giving readers the possibility to understand more clearly the subject matter covered in the patent.

Overall, out of the 40 patents selected for the test database, a total of 120 keywords were extracted. Of these, 80 were identified as unambiguous. Analyzing this group of keywords, it was observed that it is made up of keywords that really have no ambiguity, such as **microcontroller**, **digital asset management**, **greyscale**, **geometric transformation**, **color image sensor**, **image acquisition**, **homography**, **input device**, among others. In addition, 30 keywords were identified as ambiguous. They were **pipe**, **product**, **tag**, **relationship**, **cell**, **filter**, **feature**, among others. The average reliability of sense disambiguation for the 30 keywords was of 76.19%, with values between 41.82% and 100%. Figure 5 shows the amount of keywords distributed in each realiability interval. Of the keywords extracted, 10 were not found in the wikipedia database, showing the need to use another database. In this group of words, we found words such as **dynamic range**, **center line**, **audible**, among others.

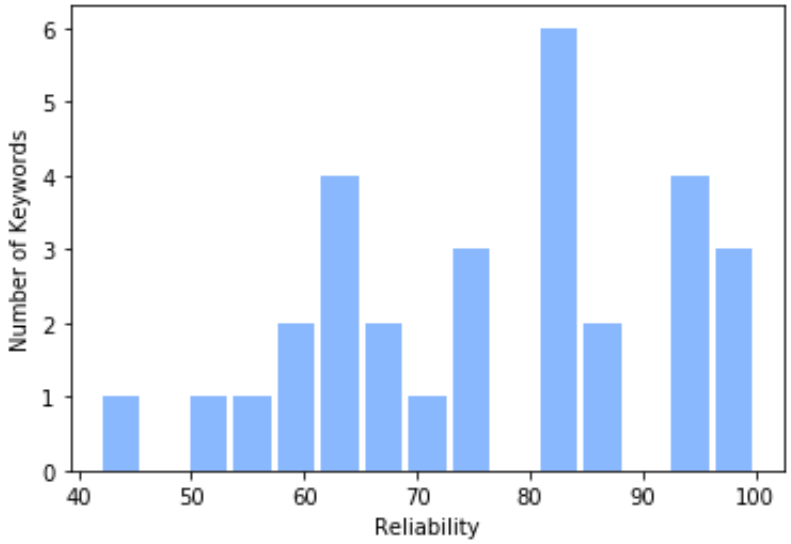

Fig. 5. Realiability of the WSD process

To exemplify the results obtained, we selected four patents that have at least one ambiguous keyword. Tables III, IV, V and VI present some of the results obtained in the WSD step. The first column presents the extracted keywords that have ambiguous meaning, the second one, the hypernyms, the third one, the meaning of the word in the specific context of the patent and the fourth presents the reliability calculated by WSD algorithm. Before each table, the three extracted keywords and the context for sense disambiguation of the selected keywords are presented. The selected patents were identified as P_1, P_2, P_3 and P_4.

Patent P_1
Keywords: compression, **filters**, input buffer.
Context for sense disambiguation of the keyword **filters**:

"An apparatus, and corresponding method for finding an area of interest in an input image, **filters** the input image with at least one compression filter to generate a final compressed image. The apparatus determines the location of an artifact in the final compressed image and then determines the location of the area of interest in the input image according to the location of the artifact in the final compressed image. An apparatus, and corresponding method for finding one or more artifacts in a two-dimensional image, receives a first row of the image and generates a list of regions in accordance with the first row. The apparatus receives a next row of the image and updates the list of regions

in accordance with the next row. The apparatus determines whether a region in the list of regions corresponds to an artifact in the image and, if so, selects the region as the artifact."

Table III. Results of the WSD process for P_1

Keyword	Hypernyms	Meaning	Reliability
filters	device equipament technique application	device	87,35%

Patent P_2
Keyword: **magnitude**, candidate, fft.
Context for sense disambiguation of the keyword **magnitude**:

" 11. The method of claim 1, wherein determining whether the candidate start region includes at least a portion of a candidate oriented encoded signal includes: determining a number of pixels in the candidate start region that have an edge **magnitude** exceeding an edge **magnitude** threshold; and comparing the number of pixels to a pixel count threshold."

Table IV. Results of the WSD process for P_2

Keyword	Hypernyms	Meaning	Reliability
magnitude	factor parameter property condition	factor	74,48%

Patent P_3
Keyword: qr, **registration**, dynamic image.
Context for sense disambiguation of the keyword **registration**:

"The present disclosure relates to systems and methods for decoding intrinsic matrixed bar codes, such as Quick Response ("QR") codes. In one implementation, system for decoding an intrinsic matrixed bar code may include an image-receiving device, a processor configured to execute instructions, and a memory storing the instructions. The instructions may include instructions to: receive an image using the image-receiving device; determine whether the received image contains at least one **registration** mark; when the received image is determined to have at least one **registration** mark: determine, from the at least one **registration** mark, one or more coordinates; when the received image is determined not to have at least one **registration** mark: determine one or more coordinates based on a detected color shifting of a portion of the image; and extract one or more matrixed bar codes overlaid on the received image and located at the one or more coordinates."

Table V. Results of the WSD process for P_3

Keyword	Hypernyms	Meaning	Reliability
registration	service information issue document	service	75,71%

Patent P_4

Keyword: image capturing device, **product**, linear slide.
Context for sense disambiguation of the keyword **product**:

"A system for acquiring multi-angle images of a **product** includes a workstation having a working surface for placing a **product**, a camera supporting member having a vertical axis, and an image capturing device movably attached to the camera supporting member so that it may move along the vertical axis of the camera supporting member. The system captures and analyzes a digital image of a **product** to detect the vertical center of the **product**, and adjusts the position of the image capturing device along the vertical axis so that the vertical center of the **product** is proximate to the vertical center of the image. The system may also have a turntable and additionally rotate the turntable at multiple capturing angles and capture one or more additional digital images of the **product** at various capturing angles and store the one or more additional images in a **product** database."

Table VI. Results of the WSD process for P_4

Keyword	Hypernyms	Meaning	Reliability
product	item industry service material	item	92,40%

By analyzing the presented results, it can be seen that the reliability values found by the WSD algorithm were satisfactory. In general, ambiguous keywords are not necessarily complex terms, but they are of great relevance to the document. When analyzing the keyword and its meaning obtained through the WSD algorithm, it was possible to verify that, in many cases, this word does not have a common meaning. Therefore, it becomes necessary to have a great understanding of the content of the patent to correctly identify the meaning and to evaluate if it is coherent. The possibility of generating links to other databases besides Wikipedia comes with the need to supply the cases in which it is not possible to find a page in Wikipedia that addresses the subject of the keyword in a specific meaning. The experiments showed that the algorithm used to identify the link to the patent database presents a satisfactory result, since it was possible to select a patent capable of explaining the content of the keyword. However, using the key file with only three of the best keywords limited the number of possible patents to be used as the link target. Figure 6 shows the word **metadata** identified, in the patent of subgroup G06K 7/1443 as a keyword. This keyword was linked to another patent from subgroup G06K 7/1447 and to a Wikipedia article.

5. FINAL CONSIDERATIONS

Patents are legal documents with a significant number of technical and descriptive details, which makes their analysis very complex. Access to information in such documents is often laborious, due to the difficulty imposed by technical language and poorly designed writing styles, contrary to the main objective of a patent system of sharing knowledge.

The proposal to create an automatic link generation system is an alternative to provide a simpler access to the knowledge bases related to the content of the patent document. With this proposal, we expect to contribute to the study of links in patents, facilitating the understanding of the information contained in these documents and promoting the dissemination of scientific knowledge associated with the technological advances proposed by these inventions. As a continuation of this work, an user experiment will be conducted to determine which terms in the patents used in the experiments require reference to external knowledge. As users with different types of background may indicate different words to link, these selected words can be considered gold standards to evaluate the performance of our experiments.

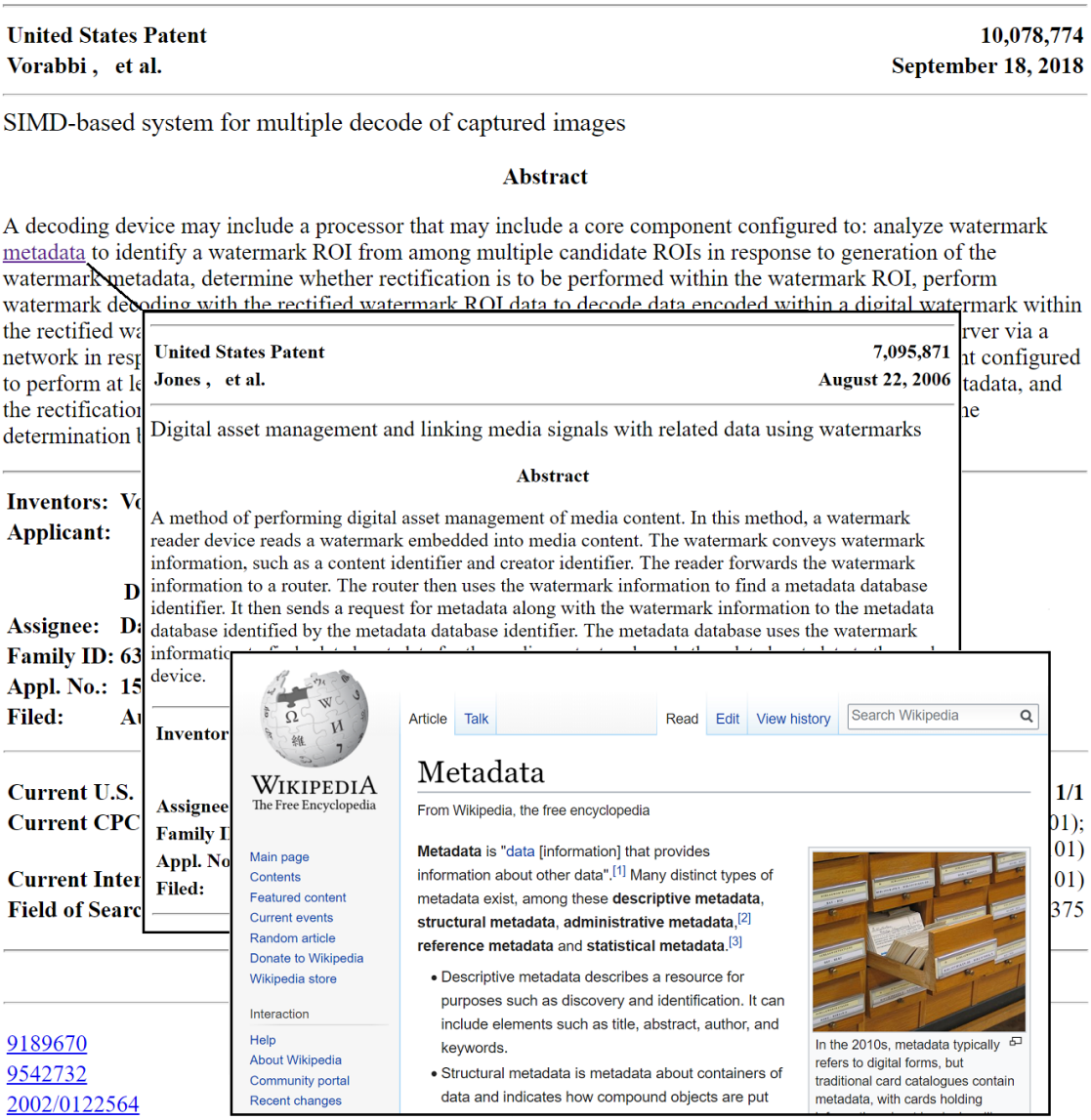

Fig. 6. Example of a patent with the link to two databases

REFERENCES

Codina-Filbà, J., Bouayad-Agha, N., Burga, A., Casamayor, G., Mille, S., Müller, A., Saggion, H., and Wanner, L. Using genre-specific features for patent summaries. *Information Processing & Management* 53 (1): 151–174, 2017.

Corrêa Jr, E. A., Lopes, A. A., and Amancio, D. R. Word sense disambiguation: A complex network approach. *Information Sciences* vol. 442-443, pp. 103–113, 2018.

Cucerzan, S. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. pp. 708–716, 2007.

Duari, S. and Bhatnagar, V. scake: Semantic connectivity aware keyword extraction. *Information Sciences* vol. 477, pp. 100–117, 2019.

Durham, A. L. *Patent law essentials: A concise guide*. ABC-CLIO, 2018.

Erbs, N., Zesch, T., and Gurevych, I. Link discovery: A comprehensive analysis. In *2011 IEEE Fifth International Conference on Semantic Computing*. IEEE, pp. 83–86, 2011.

Gardner, J. J. and Xiong, L. 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, pp. 1701–1704, 2009.

Han, X., Sun, L., and Zhao, J. Collective entity linking in web text: a graph-based method. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, pp. 765–774, 2011.

Jana, A., Mooriyath, S., Mukherjee, A., and Goyal, P. Wikim: metapaths based wikification of scientific abstracts. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, pp. 1–10, 2017.

Khode, A. and Jambhorkar, S. A literature review on patent information retrieval techniques. *Indian Journal of Science & Technology* 10 (37), 2017.

Manning, C. D. and Schütze, H. *Foundations of statistical natural language processing*. MIT press, 1999.

Meireles, M. R. G., Ferraro, G., and Geva, S. Classification and information management for patent collections: a literature review and some research questions. *Information Research* 21 (1), 2016.

Mihalcea, R. and Csomai, A. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. CIKM '07. ACM, pp. 233–242, 2007.

Onan, A., Korukoğlu, S., and Bulut, H. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications* vol. 57, pp. 232–247, 2016.

Ouellette, L. L. Who reads patents? *Nature biotechnology* 35 (5): 421–424, 2017.

Panchenko, A., Ruppert, E., Faralli, S., Ponzetto, S. P., and Biemann, C. Unsupervised does not mean uninterpretable: The case for word sense induction and disambiguation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Vol. 1. pp. 86–98, 2017.

Ratinov, L., Roth, D., Downey, D., and Anderson, M. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pp. 1375–1384, 2011.

Reginaldo, T. V., Lucindo, D. L. B., Meireles, M. R. G., Patrocínio Júnior, Z. K. G., and Almeida, P. E. M. A comparison of algorithms for the extraction of keywords in a patent database. *Proceedings of the XXXVIII Iberian Latin-American Congress on Computational Methods in Engineering*, 2017.

Seneviratne, D. *Patent Link Discovery*. Ph.D. thesis, Queensland University of Technology, 2018.

Tsunakawa, T. and Kaji, H. Towards cross-lingual patent wikification. *Proceedings of 6th Workshp on Patent and Scientific Literature Translation (PSLT6)* vol. 6, pp. 89, 2015.