Applications and Challenges of Text Mining with Patents

Hidir Aras, René Hackl-Sommer, Michael Schwantner and Mustafa Sofean
FIZ Karlsruhe
Hermann-von-Helmholtz-Platz 1, D-76344 Eggenstein-Leopoldshafen
firstname.lastname@fiz-karlsruhe.de

ABSTRACT

This paper gives insight into our current research on three text mining tools for patents designed for information professionals. The first tool identifies numeric properties in the patent text and normalises them, the second extracts a list of keywords that are relevant and reveal the invention in the patent text, and the third tool attempts to segment the patent's description into it's sections. Our tools are used in the industry and could be applied in research as well.

1. INTRODUCTION

Patents are a very complex and difficult to analyse type of text. As described in [10], their linguistic structure differs very much from common language. Patents, as a corpus and as a single document, are both very heterogeneous. They belong to subject areas as diverse as chemistry, pharmacology, mining and all areas of engineering, with the consequence that all kinds of terminology can be found in a patent corpus. A patent corpus usually covers a long time span, often from the 1950s to the present. Patents from the principal patent authorities amount to more than 70 million publications. Typographical errors are not uncommon, since many patents in their machine-readable form are derived from OCR-processing and machine-translation. Patents are on the average two up to five times longer than scientific articles. Their textual part is composed mainly of the detailed description of the invention and the claims. The former is often similar to scientific articles, whereas the latter is characterised by a legal language.

Users of patent information usually are information professionals, who cooperate with the research departments or the legal department of their companies. They have very high requirements on the correctness and completeness of the data, on the efficiency of the search interface, and on the trustworthiness of the provider. The cause of their search is normally business critical, the endeavour compares to a search for a needle in a haystack. Their search strategy is by far different from a typical Google search; it uses complex Boolean queries, the diligent usage of proximity operators, and vast lists of synonyms. New functionality, which helps them in searching and analysing the result set, is therefore greatly appreciated. Tools and methods for ordinary documents are manifold, the challenge is to adapt or to re-design them in such a manner that they work with patents.

In this paper, we introduce three text mining tools specifically designed for patent texts we have implemented or are investigating on, respectively. Section 2 describes the numeric property extraction, which allows for recognising numbers, measurements, and intervals. This feature enables the user to integrate a search for numeric properties, e.g. for temperature measurements ranging from 150K to 200K, into his query to enhance the precision. Section 3 shows the challenges of automatic keyword extraction with focus on the invention, giving the user the opportunity to get a quicker overview of the content of a single document or an answer set. Section 4 outlines the patent description segmentation, a tool for identifying the several parts which constitute a patent description. With that, the user can limit his search to specific parts of the description, again for a higher precision. Finally, we conclude this work with our main findings and future work.

2. NUMERIC PROPERTY EXTRACTION

In many technical fields, key information is provided in the form of figures and units of measurement. However, when these data appear in full text, they are almost certainly lost for search and retrieval purposes. The reason for this is that full text is indexed in a way that makes it searchable with strings. In that manner, only the string representation of a numeric property would be searchable, which is, of course, wholly unsatisfactory.

2.1 Related Work

To date, some attempts have been made to extract such data automatically from text. A tentative approach in GATE where the identification of numeric properties from patents was addressed as a sub-task is described in [1]. [4] examine the detection of units of measurement in English and Croatian newspaper articles over a small sample of 1745 articles per language using NooJ. [9] investigate the issue from a Belarussian/Russian perspective with many unique language-related challenges relying on NooJ, too. These approaches lack either the generalisability to an extensive corpus or deal mainly with the Russian language. There is also

Copyright l' 2014 for the individual papers by the papers' authors. Copying permitted for private and academic purposes. This volume is published and copyrighted by its editors. Published at Ceur-ws.org Proceedings of the First International Workshop on Patent Mining and Its Applications (IPAMIN) 2014. Hildesheim. Oct. 7th. 2014. At KONVENS'14, October 8-10, 2014, Hildesheim, Germany.

a commercial tool available from quantalyze[1], however, this tool appears to identify a much more limited variety of units than ours and it also lacks the identification of enumerations, which are abundant in patents and therefore indispensable.

2.2 Requirements and Tasks

The following sections describe the requirements and relevant tasks in numeric property extraction.

Identification of numbers

Clearly, a number consisting of digits only can be easily identified. For numbers with decimal points we have observed that in our data both numbers following English as well as German convention are present. Numbers do also appear in scientific notation, and there is a range of characters that is used to denote a multiplication or a exponentiation. We also note the use of the HTML sup-tag indicating superscript. Examples of valid expressions therefore include:

```
1,300.5 (English convention);  1.300,5 (German);
3.6 x 10-4; 10^5;  4.5x10sup"5;  8.44 x 10 sup* 10
```

Frequently, in patents numbers are spelled-out, as in *ten mg* instead of *10 mg*. These instances are recognised and converted into their respective numerical values.

Identification of units of measurements

This task, looking simple at first sight, requires some attention with respect to spelling (in particular uppercase vs. lowercase), spacing, and disambiguation.

- Upper-/lower case: There are some instances, in which capital letters and small letters refer to different entities, e.g. *S* stands for *Siemens*, a unit for electric conductance, whereas *s* stands for *second*.
- Spacing: There is some diversity regarding blank characters in spellings of units of measurement consisting of more than one word, e.g. *J per mol-K*. Therefore, the longest possible sequence in a series of tokens has to be matched.
- Ambiguity: For a few units, their abbreviated spellings might refer to different entities, e.g. *C* might stand for *Degrees Celsius* or *Coulomb*; *A* might mean Ampere or Angström (cf. Noise Reduction).

The vast majority of units appear after numbers; however, there are some units that only appear before numbers, like the pH value or the refractive index.

Unit normalisation

Many measurements of physical properties can be expressed with various units. For example, *800W* is equivalent to *800 Joules/second*, and *180°C* to *453 degrees Kelvin*. For the measurement of pressure, the following non-exhaustive list of units can be used: *kg/m2; N/m2; Pa; Torr; atm; cm Hg; ounces per square yard*. Additionally, a great number of prefixes like *nano*, *μ*, *kilo, tera* and their abbreviations have to be considered. Hence, to get a hit with standard indexing, a user would need to include all sorts of variations in order to achieve even a modicum of accuracy and recall. Clearly, a superior way to address these issues is to define a common base unit for all units which describe the same physical property and convert all instances from the full text to that base unit for indexing and searching. Therefore, all instances of units from the full text are converted into their corresponding base units as they are defined in the International System of Units (SI).

[1]https://www.quantalyze.com/

Identification of Intervals

There are two main ways in which intervals can be construed. One relies on context words, in which the words surrounding numeric entities indicate an interval, e.g. *between 12 and 100 Watts*. Another way comprises the use of symbols, e.g. *5–6 mg* or *>12 hours*. While there are only some phrases that are often encountered which indicate intervals with bounds on both sides, there are many more when it comes to intervals unbounded on one side. The latter can appear before or behind the numeric entities to which they refer, e.g. *more than 200ml* or *200ml or more*. Negated formulations like *not more than* have to be taken into account as well. Frequently, there are also adverbs present which add no specific information to the context, but just need to be filtered out, e.g. *about, around, roughly*.

Enumerations and Ratios

Enumerations of numbers or even intervals are very common in patents. They usually follow a comma-separated pattern: *a thickness of 1, 2, 3, 4, or 5 mm*. The identification of enumerations is rather straightforward as there is only a small number of variations that together cover >90% of occurrences.

Ratios are used to describe the proportionate relationship between two or more entities from a common physical dimension. A sample expression from an everyday background might be *make sure the ratio between sugar and flour is 1:3*. This being a simple example, the recognition of ratios is actually a difficult endeavour. The reason is the immense heterogeneity in which ratios can be expressed. Simple ratio formulations are typically separated by colons or slashes. They take general forms like "Number:Number" or "Number-to-Number". An approach relying solely on these patterns will invariably locate many false positives.

Noise Reduction

The aim of noise reduction is to eliminate false positives. This is a critical task especially for units of measurements consisting of only one letter, the most frequent being the aforementioned A and C.

2.3 Implementation

We are using the Apache UIMA framework for the presented analysis of data. It provides a robust infrastructure for developing modular components and deploying them in a production environment. Finite State Automata (FSA) are used throughout for pattern matching. They perform much better than Java-patterns and regular expressions, and even small improvements add up quickly when it comes to processing data in the terabyte range. For the identification of numbers, intervals, and enumerations valid sequences of phrase parts and type-related placeholders (both configurable) are expressed in a FSA-based grammar.

Adapted to the English language, our system currently recognises more than 15,000 unit variants belonging to 80 base units. Included are all commonly used dimensions like time, temperature, or weight, but also many dimensions that are more relevant in professional use, e.g. dynamic viscosity, solubility, or thermal conductivity. We are using a windowing technique for ratio recognition. From any occurrence of the word *ratio* in the text, up to five words to the left and 15 words to the right are evaluated. While this approach manages to identify many valid ratios, many cases still remain in which ratios are not recognised, like ratios for more than two entities or ratios in alternative formulations (e.g. *10 parts carbon black and 4 to 6 parts oil extender*). These will be dealt with in future versions.

Conversion between units is a straightforward task. The units, their variants and conversion rules are kept in a configuration file. Three more configuration files are provided for rules to recognise intervals and for the noise reduction, respectively. By this means, changes or extensions can be effected without the need to change the source code and redeploy the software. For the noise reduction task, two lists have been defined. The first list applies to all units. It contains terms like *figure* or *example*. If one of those global terms precedes a numeric entity, that entity is judged as noise and removed (examples: *figure 1A* or *drawing 2C*). The second list is specific to certain units only. If a term contained therein follows a numeric entity, this text passage will be ignored as well (e.g. *13C NMR*). Extracted and converted entities are added to our search engine indexes.

Regarding evaluation, we followed an iterative development cycle with many intellectual assessments. In the process, we have set up extensive JUnit-tests for software development and continuous integration. When a test person or, later, one of our customers found a specific piece of text that required improvement, we included it. As a result, given the size of our data it has over time become increasingly difficult to find text snippets that are not or faultily recognised. We have not carried out extensive formal recall/precision evaluations, because the effort required building a gold standard with significant sample size and real world data (as opposed to manually construed "difficult" data) is not offset by the projected gains. All our customer feedback indicates that our results are very good.

3. KEYWORD EXTRACTION

Keywords extracted from a document are of great benefit for search and content analysis. In the patent domain important keywords can be utilised for searching as well as getting an overview of the topics and the focus of a single patent document or an answer set. In both cases they can avoid unnecessary time-consuming and costly analysis e.g. in prior art or freedom to operate scenarios. Existing methods for keyword extraction – be it automatic or supervised – use either statistical features for detecting keywords based on the distribution of words, sub-words and multi-words, or exploit linguistic information (e.g. part-of-speech) over a lexical, syntactic or discourse analysis. Furthermore, hybrid approaches exist, which try to combine the various types of algorithms and apply additional heuristic rules, e.g. based on position, length or layout.

3.1 Related Work

[2] used term frequency, phrase frequency and the frequency of the head noun for identifying the relevant keywords from a candidate set. The phrase candidates are sorted according to the head noun frequency. Afterwards additional statistical filters are applied. [7] reported that technical terms mainly consist of multi-words, e.g. noun phrases with a noun, adjective and the preposition "of" in English texts. Single words in general are less appropriate for representing terminology. Most word combinations describing terminology are noun phrases with adjective-noun combinations. Experiments also indicate the impact of the term position, e.g. in title or a special section. It was also shown that proper nouns rarely represent good keywords for representing terminology.

3.2 Challenges and Tasks

One main challenge in keyword extraction is related to the subjectivity of keywords for a particular user, whose expertise, common knowledge about the regarded technical domains and the focus of interest can vary with respect to manifold aspects. Besides that, patent full texts describe general aspects, state of the art that experts are familiar with and make use of expressions and terms that are rarely used in classic texts (*neologisms*). Hence, separating the wheat from the chaff can be difficult. Moreover, as the description part of a patent can be very heterogeneous, mixed with tables, figures, examples, mathematical or chemical formulas, etc., identifying relevant sections that contain keywords that are directly related to the invention can be a tricky task as well. All these challenges call for deeper analysis of the content, in order to better understand patent texts and improve searching specific aspects or entities in the patent texts.

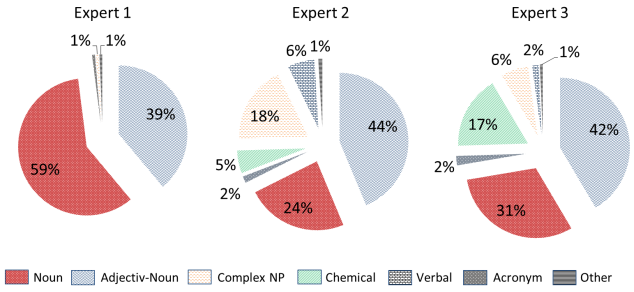

Figure 1: Phrase pattern distribution of top keywords from three experts (Analysis of EPO patents).

Analyses show that most of the relevant linguistic phrases in patent texts are noun-sequences and noun-adjective combinations (Figure 1). Despite this, depending on the domain of interest, complex noun phrases that are used to describe, e.g. a process, chemical entity or formula, and verbal phrases can be observed. The role of the verbal phrases seems to be debatable, as recent results [8] show.

Investigation of evaluation data from experts indicate that extracting phrases of length ≤ 5 is reasonable in case of linguistic technical terms, which might be different when considering also domain-specific entities from the chemical, bio-pharma, or other domains. Figure 2 shows the frequency distribution of the phrase lengths up to 9 words in the an-

notated corpus. For example, in the descriptions part, the experts annotated more than 350 times phrases consisting of only two words. Focusing on automatic keyword extraction, a further prerequisite is to deal with similar phrases with different morphological and syntactical structure. For keyword search or for generating content overviews this syntactic variations [5] must be normalised and mapped to one canonical form. For example: circular or rectangular patterns → circular pattern, rectangular pattern, method for combating spam → spam combating method, etc.

Another important task that also concerns patent search in general is semantic normalisation to aggregate semantically equivalent or similar phrases which can vary in wording considerably. The recognition of specific entities – be it simple or complex forms, identifying taxonomic relations, synonyms, chemical entities, enumerations, etc. represent other challenges in the course of understanding a given patent text beyond general linguistic phrases or terms. In classic keyword extraction, keywords in title or abstract are automatically regarded as important, while for patents a sophisticated weighting scheme based on analysing keyword occurrence and co-occurrence with respect to different sections is required. A further task is to decide how the final keyword set is presented to the user. While in classic keyword extraction rarely more than 10 keywords are returned to the user, in the patent domain information professionals indicate that displaying 50, even 100 keywords would be desirable.

3.3 Implementation and Evaluation

A proof-of-concept prototype based on linguistic and statistical analysis was implemented in order to evaluate some of the described tasks. The general procedure comprised the steps for linguistic and statistical pre-processing, noun phrase extraction and analysis and phrase weighting based on features such as length, position, TF-iDF weight or section. A typical linguistic pre-processing includes sentence detection, tokenisation, POS-tagging and noun phrase chunking. The noun phrase extraction allows to identify basic patterns of important noun phrase chunks, while applying a filtering method for removing irrelevant (stop-)words at start and end. As many syntactic variations of the extracted keywords may occur besides a syntactic normalisation method, linguistic and statistical analysis must be applied in order to reduce the candidate set for ranking. A candidate phrase is evaluated by means of a scoring formula that takes into account the respective parameters. In order to avoid loss of information, a conservative method is preferred over utilising harsh frequency thresholds. Rather, the overall ranking is affected by an elaborated weighting scheme considering besides intra-section features also field-based analysis for the sections title, abstract, claims and the descriptions text.

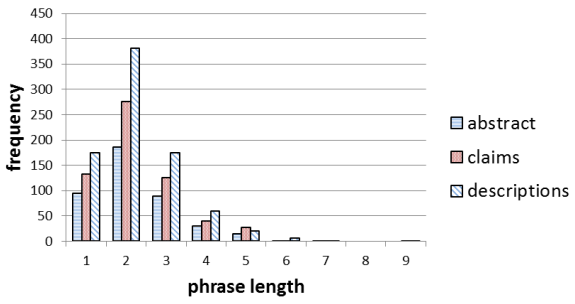

Figure 2: Phrase length distribution of top keywords for abstract, claims and descriptions.

3.3.1 Dataset and Evaluation

The implemented approach was evaluated based on a corpus with 20.000 documents from several domains, e.g. chemical, bio-pharma as well as engineering, from the European patent database comprising granted patent documents having title, abstract, claims and descriptions text. An expert-based study served to create a test corpus of 70 patent documents annotated with keywords in the aforementioned main sections of the patent text. Therefore, the two participating experts marked up to 20 most relevant keywords in a patent document that characterise the topic and the focus of the described invention. The main textual sections comprising the combined title-abstract, claims and descriptions were evaluated separately, i.e. keywords sets were not mixed. The created (annotated) datasets were used for evaluating the keyword extraction. For evaluating the implemented baselines based on the TF-iDF weighting scheme, the rank-based evaluation metrics precision@k, recall@k and F-Score have been used.

For the field combination title-abstract, the exact keyword match results for precision varied between 34% for the top 10 keywords and 20% up to 30% for the top 20. Looking at recall considering a wider range of up to 50 keywords, a score around 40% was calculated. As exact match does not consider syntactic variations for the extracted key phrases, a fuzzy matching method was applied as well. Depending on the fuzziness parameter, false positives may also be returned, which only can be detected by manual expert-based inspection. The results after applying the fuzzy matching method were much better for precision (~75% for the top 10 keywords and 46% for the top 20 keywords) and recall (~87%). For the claims the precision varied between 27% and 30% for the top 20 keywords in case of exact match, while again the recall for the extracted keywords increased from 27% to approx. 46% when taking a wider range of up to 50 keywords. For fuzzy matching, a precision score above 75% for the top 10 keywords and 70% for the top 20 was achieved. In claims, the recall for the top 50 keywords was about 92%. Due to the heterogeneity and the amount of text present in the descriptions part, the challenges seem here much higher. For the TF-iDF baseline the exact match results for precision varied between 14%-15%, while the recall for the top 10-50 keywords increases from 8% to 25%. Applying fuzzy matching, the scores for precision were again much better. Depending on the fuzziness parameter for the matching similarity that varied between 0.5 (50% match) and 0.9 (90% match), the precision score was between 80% and 50% for the top 50 keywords for the regarded dataset.

4. TEXT SEGMENTATION

Patent documents are lengthy, abundant, and full of details, such that it may hinder the topic analysis for humans and for machines as well. One of the text mining techniques which can ease these intricacies is text segmentation [3]. The automatic structuring of patent texts into pre-defined sec-

Table 1: A list of sections in description text of the patent.

Section Types	Example
Detailed Description	Best Mode of the Invention, Embodiments of the Invention
Background	Background of Invention, Prior Art
Summary	Summary of the Invention, Objectives of the Invention, Disclosures
Methods	Procedures, Operations, Experiments
Drawing and Figures	Detailed Description of the Drawing
Applicability	Industrial Applications, Applications of the Invention
Technical Field	Technical Field of the Invention, Field of Technology
Examples	Embodiment Example, Experimental Example
Sequences	List of Sequences, Numerical Sequence
References	List of References, Literatures
Statements	Statement of Government Rights, Acknowledgement

tions will serve as a pre-processing step to patent retrieval and information extraction, as well as enable the interested people to understand easily the structure of a patent that leads to fast, efficient, and easy access to specific information which they are looking for. Furthermore, noun phrases of important sections in the patent texts could be used as main features for patent classification and clustering to achieve a good performance.

The textual part of a patent contains title, abstract, claims, and the detailed description (DetD) of the invention. The latter includes the summary, embodiment, and the description of figures and drawings of the invention. As of the amount of information in DetD, there is a need for automated tools, which can determine the document-level structure of the DetD, identify the different sections and map them automatically to known section types. There has been previous work which showed that the semantic of the patent document structure is valuable in patent retrieval [6], but it only focused on structured patent text which is labelled by specific tags in the original text. The work in [1] presented a rule-based information extraction system to automatically annotate patents with relevant metadata including section titles. In this section, we describe our text segmentation method which is used to recognise the structure of the DetD.

There are many challenges that arise in patent text segmentation, for example measuring the similarity between the sentences is difficult to use because there are a lot of identical terms in the sentences. Another challenge is that the patent contains a lot of new technical terminologies which are hard to collect when using a term matching technique. To meet these challenges, we currently develop a patent text segmentation tool which automatically segments the patent text into semantic sections by discovering the headers inside the texts, identifying the text content which is related to each header, and determining the meaning of the header.

4.1 Dataset and Preprocessing

Our dataset consists of a random sample of 139,233 patents from the European Patent Office (EPO) and converted by FIZ Karlsruhe[2] into a proprietary XML format with tagged paragraphs. Processing techniques have been applied to understand the type, style, and format of headings inside patent texts. We started by parsing XML files to get a list of headings in the description part. The headers pass through a cleansing process that is represented by removing undesired tokens in each header (e.g.; numbers, special characters, words containing special symbols, words starting with numbers, math equations, and formulas) via a tokenisation process. Then, we created the positive-list which contains terms that appear more than five times in all headings of the dataset, and the first-token list which includes terms from the headers which appear more than five times as the first word of a header.

[2]http://www.fiz-karlsruhe.de

4.2 Header Detection and Meaning

In cooperation with a patent expert, we identified segmentation guidelines. These guidelines help us to understand the section types (Table 1) in the DetD. In order to discover the headers inside the DetD, we need to get the boundary of the headers. i.e., the header's start and end. We call this operation *Header Detection*. Then, we identify the text content which is related to each header. The header meaning on the other hand is represented by assigning the header to an appropriate section type (e.g.; summary, example, background, method, etc). Here, a rule-based approach is more suitable because in the patent domain, there is no sufficient training data for a machine learning algorithm to be successful. To do so, we develop a rule-based algorithm to identify headers and their boundaries. The output consists of all headers and their positions inside the DetD. Our algorithm works as follows: As input we take the DetD as a sequence of paraghraphs. Then, we test the following features to decide whether a paragraph is a header or not:

A. The number of words in the paragraph.

B. The number of characters in the paragraph.

C. True, if all letters in the current paragraph are in upper case; false otherwise.

D. True, if all words in the paragraph start with an upper case letter; false otherwise.

E. True, if the current paragraph contains words from the positive-list, false otherwise.

F. True, if in the current paragraph more words start with a capital letter than with a lowercase; false otherwise.

G. True, if the current paragraph starts with a bullet; false otherwise.

H. True, if the previous or the next paragraph starts with a bullet; false otherwise.

I. True, if the first token in the paragraph appears in the first-token list; false otherwise.

J. True, if the current text paragraph contains a simple chemical text; false otherwise.

K. The average header length in the dataset's headers.

L. The average number of words in the dataset's headers.

We use these features on each input paragraph of the DetD to build decision rules for the header detection. Some of the decision rules are listed below:

i. C is true and G is false and A≥1 and J is false

ii. D is true, E is true, A≥1, G is false, and J is false

iii. G is true, H is false, A<L, J is false, B<K, and A≥1

iv. I is true, F is true, J is false, A≥1, and G is false.

After detecting the headers, we identify the start and end position of each header in the DetD. The detection of the text belonging to the header is performed by identifying the paragraphs between the current header and the next header. After the detection of headers and their boundaries, each header should be assigned to one of the appropriate predefined section types by using a prediction model from the machine learning step. This task was modelled as a classification task via constructing a training dataset by labelling manually a representative sample of 1377 headers of section types that are shown in the Table 1. The labelling process is done by applying the guidelines created by the patent expert. Pre-processing steps were performed to remove undesired tokens like numbers, special symbols, and stopwords, as well as to compute the weight vector for the training set. We used Support Vector Machines (SVMs) as a multi-classification technique to train the dataset. The evaluation was done by using 5-fold cross validation, and the performance of the categorisation achieved up to 91%, 90%, and 91% of accuracy, recall, and precision respectively.

5. CONCLUSION AND FUTURE WORK

In this paper we presented our research on three text mining tools tailored to the singularities of patent documents. Though patents are very different from normal texts in length, structure, language, and terminology, though the requirements of patent information searchers are much more strict than those of other users, and though no gold-standards for these tasks are available, which reflect a realistic retrieval situation, we could show, that solutions exist which can cope with these challenges. The results of our numeric entity extractor are since long available to our clients and are well accepted. When designing functionality like keyword extraction or description segmentation, we seek at an early stage the feedback of our customers. For the numeric property extraction, there are still some areas of potential for further research. Disambiguation is one of them: the symbols *A* and *C* were already mentioned; *in* might be a preposition or denote *inch*. Other topics concern the extraction of relations. For instance, it might be useful to identify what kind of a temperature is discussed in a text. Is it a melting point or a boiling point? To which substance or process does the temperature refer? Oftentimes in patents, whole recipe-like paragraphs are available from which a lot of factual data could be extracted. For keyword extraction, besides the challenges discussed before, learning keywords by considering domain-specific knowledge from controlled vocabularies is required to identify most relevant facts about an invention more precisely. It is also reasonable to extract keywords rather on the basis of semantic information tailored for a specific domain and use, e.g. treatment of diseases, medical substances, etc. than in an isolated manner. Possible enhanced methods for keyword context analysis could rely on semantic analysis based on the co-occurrence method, (latent) semantic analysis or other dedicated semi-supervised and unsupervised machine learning techniques. Furthermore, a more enhanced method for semantic segmentation of patent text needs to deal with patents that do not have any heading inside their texts and address the overlap problem between section types. Our final goal is to develop a flexible, scalable and automatic tool, which has the ability to facilitate the reading of a patent, keyword extraction, summary extraction, and classification and clustering of patent texts.

6. REFERENCES

[1] M. Agatonovic, N. Aswani, K. Bontcheva, H. Cunningham, T. Heitz, Y. Li, I. Roberts, and V. Tablan. Large-scale, Parallel Automatic Patent Annotation. In *Proceedings of the 1st ACM Workshop on Patent Information Retrieval*, PaIR '08, pages 1–8, New York, NY, USA, 2008. ACM.

[2] K. Barker and N. Cornacchia. Using Noun Phrase Heads to Extract Document Keyphrases. In H. Hamilton, editor, *Advances in Artificial Intelligence*, volume 1822 of *LNCS*, pages 40–52. Springer Berlin Heidelberg, 2000.

[3] D. Beeferman, A. Berger, and J. Lafferty. Statistical Models for Text Segmentation. *Machine Learning*, 34(1-3):177–210, Feb. 1999.

[4] B. Bekavac, Z. Agic, K. Sojat, and M. Tadic. Detecting Measurement Expressions using NooJ. In *Proceedings of the Conference on NooJ*, pages 121–127, 2009.

[5] R. Bhagat and E. H. Hovy. What Is a Paraphrase? *Computational Linguistics*, 39(3):463–472, 2013.

[6] H.-Y. J. Jae-Ho Kim, Jin-Xia Huang and K.-S. Choi. Patent Document Retrieval and Classification at KAIST. "Proceedings of NTCIR-5 Workshop Meeting, December 6-9, Tokyo, Japan".

[7] J. S. Justeson and S. M. Katz. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering*, 1(1):9–27, 1995.

[8] J. M. Schulz, D. Becks, C. Womser-Hacker, and T. Mandl. A Resource-light Approach to Phrase Extraction for English and German Documents from the Patent Domain and User Generated Content. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *LREC*, pages 538–543. European Language Resources Association (ELRA), 2012.

[9] A. Skopinava and Y. Hetsevich. Identification of Expressions with Units of Measurement in Scientific, Technical & Legal Texts in Belarusian and Russian. In *Proceedings of the Integrating IR technologies for Professional Search Workshop*, pages 26–34, 2013.

[10] J. M. Struss, D. Becks, T. Mandl, M. Schwantner, and C. Womser-Hacker. Patent Retrieval und Patent Mining: Sind die Anforderungen eingelöst? In *3. DGI-Konferenz. Informationsqualität und Wissensgenerierung.*, pages 25–36, 2014.