Available online at www.sciencedirect.com

ScienceDirect

Information Processing and Management 00 (2016) 1–29

Inf. P&M

Using Genre-Specific Features for Patent Summaries

Joan Codina[a], Nadjet Bouayad-Agha[a], Alicia Burga[a], Gerard Casamayor[a], Simon Mille[a], Andreas Müller[b], Horacio Saggion[a], Leo Wanner[c,a,*]

[a]*Natural Language Processing Group, Dept of Communication and Information Technologies, Pompeu Fabra University*
[b]*Institute for Natural Language Processing, University of Stuttgart*
[c]*Catalan Institute for Research and Advanced Studies (ICREA)*

Abstract

Patent search is recall-driven, which goes hand in hand with at least a partial sacrifice of precision. As a consequence, patent analysts have to regularly view and examine a large amount of patents. This implies a very high workload. Interactive analysis aids that help to minimize this workload are thus of high demand. Still, these aids do not reduce the amount of the material to be examined, they only facilitate its examination. Its reduction can be achieved working with patent summaries instead of full patent documents. So far, high quality patent summaries are produced mainly manually and only a few research works address the problem of automatic patent summarization. Most often, these works either replicate the summarization metrics known from general discourse summarization or focus on the *claims* of a patent. However, it can be observed that neither of the strategies is adequate: general discourse state-of-the-art summarization techniques are of limited use due to the idiosyncrasies of the patent genre, and techniques that focus on claims only miss in their summaries important details provided in the other sections on the components of the invention introduced in the claims. We propose a patent summarization technique that takes the idiosyncrasies of the patent genre (such as the unbalanced distribution of the content across the different sections of a patent, excessive length of the sentences in the claims, abstract vocabuly, etc.) into account to obtain a comprehensive summary of the invention. In particular, we make use of lexical chains in the claims and in the description of the invention and of aligned claim–description segments at the subsentential level to assess the relevance of the individual fragments of the document for the summary. The most relevant fragments are selected and merged using full-fledged natural language generation techniques.

Keywords:
summarization, patents, lexical chains, segmentation, segment-based summarization, sentence aggregation

1. Introduction

Patents are the treasure of the modern economies. They protect intellectual property rights, serve as source of inspiration, define business models of companies, and are instruments for securing market shares and controlling competitors. It is thus of outmost importance for any player in the patent market to monitor the increasingly dynamic patent landscape, without missing any patent that might be of importance to them. Therefore, it is not surprising that

*Corresponding author, Address: C/ Roc Boronat, 138, 08018 Barcelona, Spain
Email addresses: joan.codina@upf.edu (Joan Codina), nadjet.bouayad@upf.edu (Nadjet Bouayad-Agha), alicia.burga@upf.edu (Alicia Burga), gerard.casamayor@upf.edu (Gerard Casamayor), simon.mille@upf.edu (Simon Mille), horacio.saggion@upf.edu (Horacio Saggion), leo.wanner@upf.edu (Leo Wanner)

patent search is recall-driven [1]. This goes hand in hand with at least a partial sacrifice of precision. As a consequence, patent analysts have to regularly view and examine large amounts of patents, which implies a very high workload. Interactive analysis aids to reduce this workload are thus of high demand. Still, these aids do not reduce the volume of the material to be inspected, they only facilitate its inspection. The reduction of the volume can be achieved working with patent summaries instead of full patent documents. So far, the only source of high quality patent summaries is Thomson Reuters's Derwent World Patents Index (WPI).[1] The summaries in the WPI are written by specialists of the domain in question, and as any product that requires manual labor of specialists, they constitute an important cost factor for their consumers. Furthermore, with the rapidly growing patent markets in Northeast Asia, especially in China, but also in Japan and South Korea, the supply of manually-written high quality summaries is in danger to become a bottleneck. As already argued by Wanner et al. [2], automatic summarization of patents offers itself as a solution. However, only a few research works address the problem of the summarization of patents; cf., e.g., [3, 4, 5]. Most often, these works either replicate the summarization metrics known from general discourse summarization [5] or focus on the *Claims* of a patent that outline the scope and the nature of the invention and that are organized in a hierarchical structure, such that subordinated claims draw upon the content of their superordinated claims [3, 4]. Thus, Trappey et al. [5] rely upon the relevance of keywords determined using distribution- and ontology-based metrics to select paragraphs across the entire patent document for inclusion into the summary. Shinmori et al. [3] prune the discourse structure of each claim represented in terms of the Rhetorical Structure Theory [6] to obtain a summary. The pruning procedure is guided by the nature of the individual discourse relations in the structure and discourse tree depth: a branch of a discourse tree is cut off (and thus not included in the summary) if its origin is labelled by a "less relevant" discourse relation or if it is beyond the threshold depth of the tree. Bouayad-Agha et al. [4] prune the claim structure as well as the discourse and syntactic dependency structures of each claim to obtain a summary.

However, it can be observed that neither of the strategies (i.e., use of general discourse summarization techniques or focus on claims, respectively) is adequate. General discourse state-of-the-art summarization techniques are of limited use due to the idiosyncrasies of the patent genre such as high frequency of very abstract terms of the kind *apparatus*, *means*, *device*, etc. and excessive length of claim sentences. Since the techniques tend to select for inclusion into summaries sentences with high frequency terms, the summaries risk to be composed of few very long abstract sentences. Techniques that focus on claims only, without considering other sections of a patent, are of limited use because they will by definition not contain any embodiment information, which is also of primary relevance to readers.

In this paper, we present a patent summarization model that takes the idiosyncrasies of the patent genre into account and considers not only the Claims but also the other sections (and, in particular, the Description) of a patent during summarization. The central characteristics of the model are that it (i) is based on the notion of a subsentential *segment* as basic unit of summarization; (ii) aligns the segments in the Claims with thematically-related segments in the Description in order to capture the entire information on a content element in a patent; (iii) uses *lexical chains*, i.e., sequences of semantically-related entities, and their length to capture the distribution of the information on a content element in a patent; and (iv) draws upon segment- and lexical chain-oriented features to calculate the relevance of a given segment to the summary.

The remainder of the paper is structured as follows. Section 2 analyzes the idiosyncrasies of the patent genre and outlines our proposal. Section 3 describes how we identify lexical chains and segments. Section 4 discusses the features that we use in our summarization metric to determine the relevance of a segment to the summary and presents the metric itself. In Section 5, we show how the segments selected in terms of relevance for inclusion into the summary are aggregated into a coherent and cohesive summary, and in Section 6, we present an evaluation of the proposed summarization model. Section 7 briefly reviews the related work in the field of patent summarization, before Section 8 recaputulates the central aspects of our proposal, and sketches our future research in this area.

2. The Problem of Patent Summarization

In text summarization of general discourse, extractive and abstractive summarization techniques are often contrasted [7]. Extractive summarization is surface-oriented in that it applies relevance metrics usually based on distribution heuristics (e.g., *tf*idf* of individual tokens [8, 9], lexical chains [10, 11], position of a sentence in the text

[1] http://thomsonreuters.com/derwent-world-patents-index/

[12], etc.) to select entire sentences of a given text for inclusion into the summary. Extractive summarization can be thus assumed to presuppose sentences of a "reasonable" length, the same expressiveness of all open class tokens, and a certain locality of the content. Abstractive summarization selects from the semantic representation of a text summary-relevant content elements and uses natural language text generation techniques to assemble them and generate a coherent summary [13]. It can thus be considered to require the availability of the semantic analysis of the content of the text in question.

Let us, in what follows, analyze the idiosyncrasies of patent material, assess to what extent patent material can be expected to fulfill the prerequisites of either of the general discourse summarization techniques and if it does not, make a proposal for a model of patent summarization.

2.1. Idiosyncrasies of the Patent Genre from the Viewpoint of Summarization

Patent material reveals a number of idiosyncrasies that make it appear different from generic discourse from the viewpoint of summarization; see, e.g., [14] for a patent writing style manual. These idiosyncrasies concern (i) the general structure of the patent document (and, related to the structure, the distribution of the content across the patent), (ii) the vocabulary across the patent, and (iii) sentence length and structure complexity in a patent.

2.1.1. General structure of a patent document

Depending on the regulations of each patent office, the overall structure of a patent (application) may somewhat vary. However, in general, we can assume that it contains the following sections:

- Title
- Field of the invention
- Prior art
- Description of the drawings
- Description
- Claims
- Abstract
- Drawings

The most central of them for summarization are Claims and Description.[2] Abstract is written by the author, following the same intention as the Claims: to obfuscate the limitations of the invention, such that its utility as summary that is supposed to provide a concise, but clear description of the invention is limited. Therefore, we concentrate in what follows on Claims and Description. For the definition of the other sections, see, e.g., [14].

The Claims contain the primary information on the invention in that they outline the invention and its components in terms of a hierarchical structure of individual *claims*: *independent claims*, which present in very general and broad terms the overall invention, and *dependent claims*, which add more specific features to the presentation in superior independent or dependent claims.

Consider, for illustration, the first independent claim of US 5142421 in (1)[3] and the overall claim dependency structure of this patent in Figure 1. As we can see, the structure is rather flat. It contains one independent claim (1.) and a rather high number of first level dependent claims (claims 2.–11.). Five of these first level dependent claims are further detailed by second level dependent claims (3., 7., 10., 12., and 14.), two of which (13. and 14.) are in continuation spelled out by third level dependent claims. It is not uncommon for patents to have a considerably deeper claim dependency structure.

[2] In what follows, the names of patent document sections are written in capitals. When referring to individual parts of a patent document (as, e.g., *claim*), small letters are used.

[3] The parts of (1) in italics will be commented upon later on.

(1) *A device for recording a digital information signal in an information track on a magnetic record carrier, and for converting, prior to recording, a n-bit information words in the presented digital information signal into (n+m)-bit channel words, where n and m are integers such that* $m \geq 1$ *and* $n \geq m$, comprising: an input terminal for receiving the n-bit information words, an encoding device having an input coupled to the input terminal and having an output, which encoding device comprises an aT precoder, a being an integer greater than or equal to two and T being the bit period, which encoding device is arranged for converting the n-bit information words into the (n+m)-bit channel words and for presenting the channel words at the output; and a recording device having an input coupled to the output of the encoding device, for recording the (n+m)-bit) channel words in the information track on the magnetic record carrier; characterized in that: *the encoding device comprises signal affixing means for affixing an m-bit digital word, where m is equal to 1, to each consecutive n-bit information word to obtain an (n+1)-bit information word*; the aT precoder is arranged for converting the (n+l)-bit information words into (n+l)-bit channel words; the encoding device further comprises control signal generating means for receiving the (n+1)-bit channel words from the aT precoder and deriving a control signal therefrom; and the signal affixing means is arranged for affixing a l-bit digital word to an n-bit information word in response to said control signal, such that the running digital sum value of the output signal of the precoder has a desired pattern as a function of time.

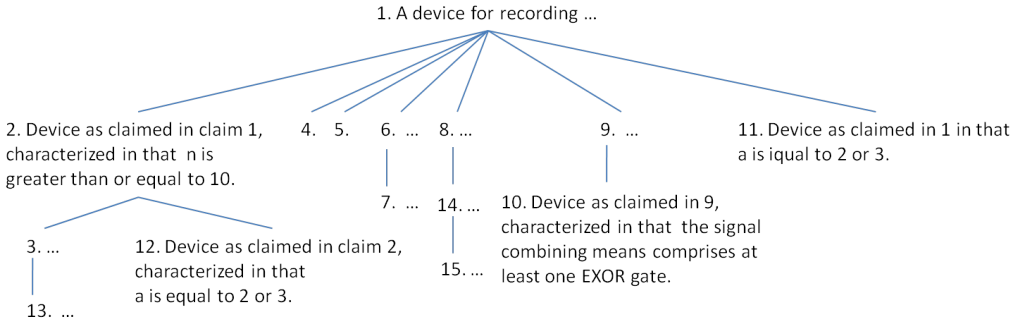

Figure 1. Claim dependency structure of the patent # US 5142421

In the Description, each claim is further detailed and information on the preferred embodiment of the claimed invention is provided, possible applications, etc.; cf. in (2) a paragraph from the description of US 5142421, which elaborates on the two elements of the invention marked in the claim in italics. That is, Claims and individual statements in the Description are related, although this relation is not symmetric.

(2) ... Encoding the n-bit information words into (n+l)-bit channel words in the device according to the invention is thus realised in a very simple manner by providing a 1-bit digital word preceding the n-bit information words. From the (n+1)-bit information word thus obtained the n most significant bits are those of the original n-bit information word. Thus, no look-up Table as in the prior-art device is necessary any more. ...

From the perspective of summarization, the analysis of the structure of a patent thus teaches us that the information that is to be summarized is distributed across Claims and Description. The hierarchical claim structure and the relations between claims and statements in the Description establish a relevance gradient that needs to be taken into account. In other words, claims and, in particular, independent claims are of primary relevance. This is why some works in patent summarization see it justified to focus on Claims and claim structure. However, the examples (1)+(2) and also (3)+(4) below show that the content in the Description needs also to be taken into account, although not "linearly". The content of the Description provides namely often the details needed for the comprehension of the invention.

2.1.2. Patent vocabulary

The most significant particularity of patent vocabulary is its abstract nature in the Claims section, with the dominance of terms like *device*, *apparatus*, *means*, *carrier*, *medium*, *part*, *portion*, etc.; see (1) and also (3)[4]. In contrast to general discourse, where we can assume with Grice [15] that the vocabulary aims to reflect the content in accurate terms and wordings that are more generic and more vague than necessary are avoided, the claim vocabulary is on purpose as generic as possible. For instance, in (1) we read *magnetic record carrier* instead of *tape*, *device for recording a digital information signal in an information track on a magnetic record carrier* instead of *tape recorder* and in (3) *automatic focusing apparatus for optical instruments* instead of *camera*.

The dominance of abstract terms in the Claims is further increased by the common practice to minimize the pronominalization of the introduced terms in order to avoid ambiguity. The consequence is a repetition (either with the demonstrative *said*; cf. *said dvisions*, *said electrical signal*, or *said objective lens* in (3) below or as such).

> (3) In an automatic focusing apparatus for optical instruments, comprising an objective lens for forming an image of the object; a photoconductive light receiving means disposed to receive the light passing through the objective lens from the object; the photoconductive light receiving means being divided into a number of divisions by a number of electrodes, each of said divisions thereby constituting discrete, photosensitive means which generate an electrical signal in response to the intensity of the light of the image which is formed thereon, and means for axially displacing the objective lens in response to said electrical signal to make a sharp image of the object on the photoconductive light receiving means; the improvement comprising means for dividing and deflecting the light beam passing through said objective lens in order to form plural images on the photoconductive light receiving means, the light receiving means being disposed between the objective lens and the photoconductive light receiving means.

In contrast, in the Description, often (although not always) "standard" specialized discourse vocabulary is used. Consider (2) and (4), which is a fragment of the description that elaborates on the claim in (3):

> (4) ... an example of the camera having an automatic focusing light receiving portion equipped with an image multiplier plate I. The camera includes an objective lens L, a focal plane F, a half-mirror I-I interposed between the lens L and focal plane F, and a photo-conductive element S of the same type as that described with respect to FIGS. 1 and 2. The image multiplier plate I is an optical image multiplier member comprising, as shown more clearly in FIGS. 8 and 9, a number of tiny pyramid-shaped prisms, which sides each have predetermined declinations. The optical image multiplier member is made of glass, transparent acrylic resin or the like. Thus, as shown in FIG. 10, a light beam directed from an object toward the photoresponsive surface S via the half-mirror H is divided and deflected by the ridges of the image multiplier plate I disposed in the light path, so that multiple images 23' and 23" are formed at predetermined points on the photo-responsive surface S.

That is, there is, on the one hand, clash of abstraction between the vocabulary in the Claims and Description sections; cf. *optical instrument* vs. *camera*, *photoconductive light receiving means* vs. *image multiplier plate*, *number of divisions* vs. *number of tiny pyramid-shaped prisms*, etc., and, on the other hand, a high frequency of abstract terms that obfuscate the exact nature of the signified.

2.1.3. Sentence length and sentence structure

As already with respect to vocabulary, Claims and Description also differ with respect to sentence length and sentence structure. Due to legal regulations, each claim is rendered as a single sentence—which means that the claim sentences are, in general, very long. Claims of 500 words are not seldom and can count as many as 900 words. Consider, for illustration, (1) and (3), which display claims of an average length. Due to their length, but also again due to legal regulations, the claim sentences furthermore show a significant complexity, with clause coordinations and multiply embedded subordinations. Consider Figure 2, which captures the syntactic dependency structure of a first

[4]The first claim of US 3688673 A

level nominal phrase (NP) in (1).[5] As the figure shows, already this NP shows considerable complexity, while further subordinations and coordinations follow.

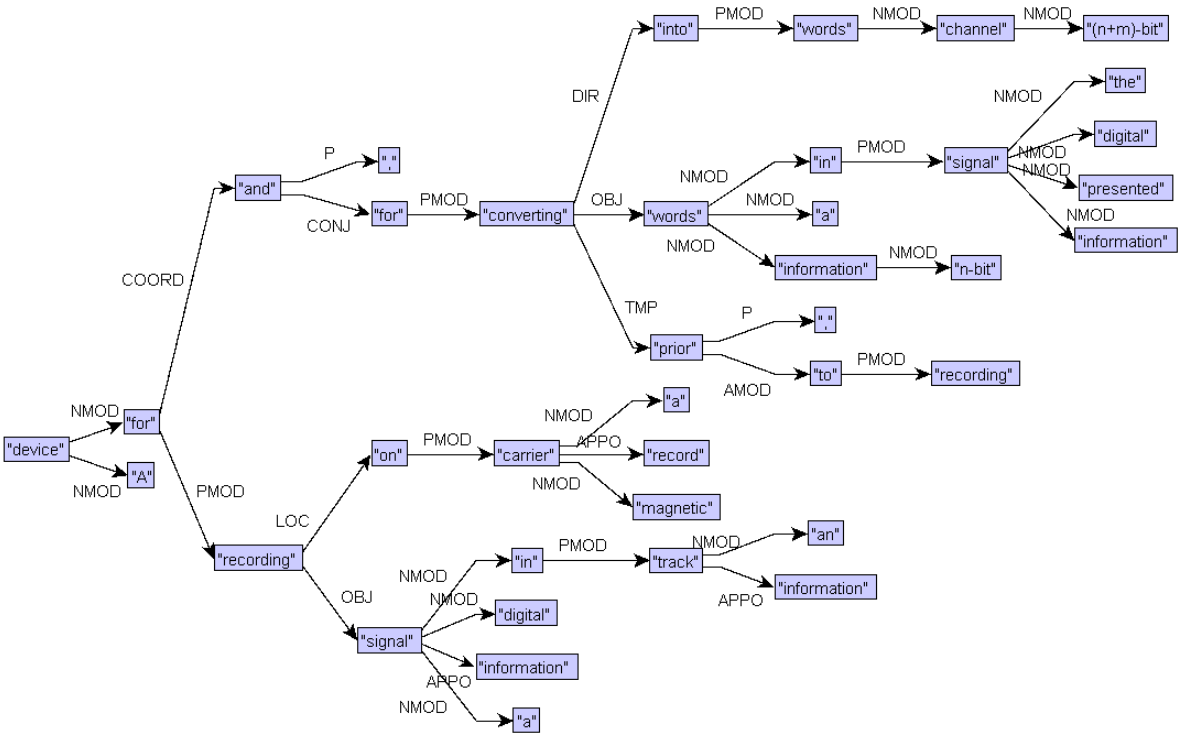

Figure 2. Syntactic dependency structure of the NP *A device for recording a digital information signal in an information track on a magnetic record carrier, and for converting , prior to recording, a n-bit information words in the presented digital information signal into (n+m)-bit channel words* from (1)

The length and structure of description sentences is, on the other side, comparable to general discourse; the description fragments in (2) and (4) illustrate this. That is, there is, again, significant unbalance between Claims and Description.

2.2. The Proposal: Towards patent genre adapted summarization

The idiosyncrasies of patents discussed above signal that the applicability of the general discourse summarization techniques to patent material can be only rather limited. As for extractive summarization, already the length of the sentences in the Claims is prohibitive for their inclusion as they are into the summary. But it will be, in particular, the claim sentences that will be suggested (to a certain extent correctly since the Claims contain the most important information) by the conventional term distribution-oriented metrics for inclusion due to the elevated frequency of the abstract terms and the concentration of these terms in the Claims; cf., e.g., *means* in (3). Restrictions on the length of the summary will furthermore favor the claims in the upper part of the claim dependency structure—as already observed, e.g., in [3, 4], and the length of the summary as such will depend on the length of the claims: the inclusion of another sentence into the summary might imply the duplication of the length of the summary.

The use of positional criteria to force the inclusion of content from the Description would help to enrich the summary by distinctive features of the invention, as expected to be found in a good summary. However, purely

[5] Such a structure specifies the grammatical function of each dependent token with respect to its governing token: OBJ (object), NMOD (nominal modifier), COORD (coordination), CONJ (conjunction), LOC (locative object), PMOD (prepositional modifier), P (punctuation), etc. We do not show the whole syntactic structure due to the space it would occupy.

positional criteria will not do justice to the fact that in the Description, different (more concrete) terms are used to denote the same signified as the abstract terms in the Claims—such that the relevance of the terms for the summary will be distorted. Lexical chain-based extractive summarization [10] will help to amend the relevance, but, at the same time, it will further lead to very long summaries with unbalanced vocabulary.

Abstractive summarization is hampered by the lack of a robust semantic analysis. The prospects of the availability of such an analysis within the next years are very remote even for general discourse, let alone for the structurally considerably more challenging patent material.

That is, a new summarization technique that adapts to the specifics of the patent genre is needed. Given that truly abstractive summarization is not feasible, this technique must rely on central extraction-oriented features, but be still linguistically as versatile as abstractive summarization. In particular, this technique must:

- operate at a subsentential level (in particular, as far as Claims are concerned) to select content elements relevant to the summary;
- recognize multiple word terms: most often abstract terms are part of complex terms that denote different elements; for instance, in (3), *means* is head of different five complex terms: *photoconductive light receiving means*, *photosensitive means*, *means for axially displacing the objective lens*, *improvement comprising means*, and *light receiving means*—all of them denoting different elements;
- relate the complementary content elements in the Description to the corresponding content elements in the Claims, taking the former also into account for summarization;
- apply content-oriented relevance metrics for the selection of the summary-relevant elements to the entire patent document, but be driven by the content of the Claims;
- have the natural language processing means to aggregate the selected elements into a coherent summary.

Before the actual summarization takes place, we must thus first "drill down" to the elements of the invention and relate those elements in that we

(i) identify lexical chains between entities within the Claims, within the Description and across Claims and Description; and

(ii) segment claim sentences and description sentences, such that each segment contains an aspect of the invention (or, in other words, a propositional statement on an element of the invention), and align the claim and description segments.

Relevance metrics then assign to each segment in the patent document (be it in the Claims or in the Description) a relevance score based on a variety of features, including lexical chain length, position, frequency, etc. Most relevant segments are selected and aggregated into a coherent summary using text generation techniques that start from the syntactic structures of the segments obtained in a preprocessing stage by a dependency parser.

To facilitate a flexible design and execution of summarization, GATE [16] is used as the infrastructure for patent document representation and module integration. Any patent chosen for summarization is preprocessed and then stored in GATE-format. The preprocessing pipeline incorporates: part-of-speech tagger, lemmatizer, and dependency parser from Bohnet's MATE tools environment [17], GATE's tokenizer, the Sentence Splitter from OpenNLP, and a proprietary patent-tuned NP chunker.[6] Figure 3 shows the architecture of our summarizer.

The individual components of the summarization module have also been developed either using GATE's Java Annotation Patterns Engine (JAPE) or integrated into GATE by a "wrapping" mechanism: the material in GATE-format is transformed into the component's proprietary format to be processed by the component, and its output is transformed from the proprietary format again into GATE-format.

[6]MATE tools were chosen because of their ability to handle very long sentences of up to 900 words found in the Claims section of patents, and the possibility to retrain them in order to adapt to the patent domain; cf. [18] for such an adaptation of the parser.

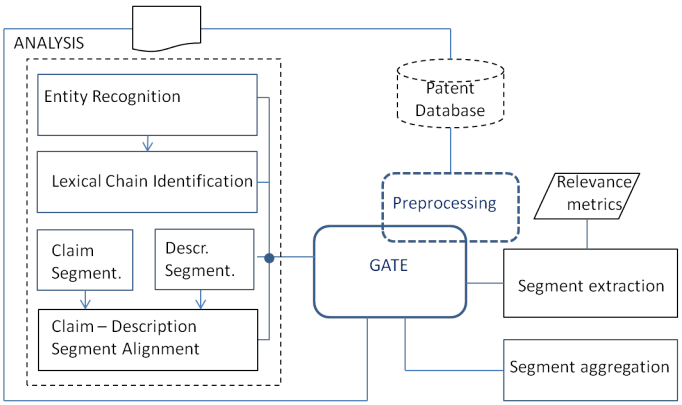

Figure 3. Patent summarizer architecture

3. Drilling down to the elements of the invention and their relations

Our summarization strategy is thus based on two types of linguistic units that allow us to reach out to the content elements of the invention, penetrating the sentence barrier that results to be a hindrance for patent summarization: individual linguistic entities and their distribution across the patent and segments and their similarity-driven alignment.

In a patent, any term can be considered to denote an entity: a functional object (or a component thereof), a substance, a process, etc. One way to obtain a picture on the distribution of the entities could be thus to first identify terms (as, done, e.g., in [19]) and then look for the repetition of the terms and relations between similar terms to obtain figures about their distribution. However, in coreference resolution research, a more straightforward (and "cheaper") methodology proved to be successful: any nominal phrase (NP), potentially with some restrictions, is *a priori* considered a *mention* of an entity [20]. If several identical or related (according to some clearly defined semantic relations) mentions are identified in the document, we can assume that the sequence of these mentions captures the distribution of the entity they denote.

In the same vein as in general discourse summarization [10], a sequence of mentions of an entity across a patent, i.e., its lexical chain, gives us a measure of its relevance for the invention and thus also of the importance of a segment it appears in for the summary. A segment is considered to be given by an entity and its relation to the invention described in the patent. The relation can denote composition, as in *the photoconductive light receiving means being divided into a number of divisions by a number of electrodes*, purpose, as in *an objective lens for forming an image of the object*, or any other meaningful relation.

If we align related segments in the Claims and in the Description, we can take the complete information on an entity in the patent into account and also order this information coherently such that if a claim segment and a description segment related to it are selected for inclusion into the summary, the two segments can be placed in sequence. Let us thus describe, in what follows, in more detail the realization of lexical chain detection, segmentation and segment alignment in our summarizer.

3.1. Lexical Chain Detection

Let us first have a closer look at lexical chains in patents and then proceed with the presentation of the framework of lexical chain detection.

3.1.1. Lexical chains in patents

As already mentioned above, lexical chains are sequences of semantically related lexical entities that ensure cohesion in discourse [21, 22]. Some of the involved semantic relations are generic and well-known (as, e.g., identity of reference, synonym/antonymy, meronymy/holonymy, hyponymy/hypernymy, paronymy, metonymy), while others are *ad hoc* and context- or domain-specific. Each discourse genre articulates lexical cohesion by relying on specific relations more than on others. To capture the most important types of relations in the patent genre and tailor

our lexical chain detection techniques to them, we carried out an empirical study on a selection of patents from the Green Repository collection of the World Intellectual Property Organization (WIPO).[7] The results of our study revealed the dominance of the following three types of relations: (i) identity of reference, (ii) meronymy/holonymy, (iii) hyponymy/hypernymy.

The 'identity of reference' (or 'coreference') relation connects elements that denote the same entity. This relation is common in patents. However, in contrast to other domains, coreference in patents is expressed mostly through reiteration of NPs rather than through pronouns or synonyms. This is because patents, and particularly the claims, are written to avoid ambiguity. One of the most common instantiations of coreference is the reference to an entity that undergoes a process. We refer to this type of reference as "entity-in-process"; cf. (5), where *a current temperature of the battery* and *the temperature detected by said temperature detection device* corefer to the same entity that undergoes measuring:

(5) [. . .] *a temperature detection device for detecting* [*a current temperature of the battery*]$_{\text{entity-in-process1}}$; *a temperature rise output device for obtaining the temperature rise from* [*the temperature detected by said temperature detection device*]$_{\text{entity-in-process2}}$ [. . .]

The 'meronymy'/'holonymy' (i.e., 'part–whole' and 'set–member') relations are used to indicate components or steps that belong to the patented object or method. Lexical chains formed by these relations often contain relational pronouns (such as *each*, *every*, etc.); cf. (6), where the indices 'part1' and 'part2' indicate components of *each head* indexed as 'whole'.

(6) [*each head*]$_{\text{whole}}$ *having* [*a DC-side arranged to connect across positive and negative terminals of cells received by the circuit*]$_{\text{part1}}$ *and* [*an AC-side for carrying an AC voltage converted from the DC-side*]$_{\text{part2}}$ . . .

The 'hyperonymy'/'hypernymy' relations are used to relate an object or a component of it with their corresponding type or class; cf. (7), where the coreferring mentions indexed by 'object' denote an entity of the type *electric circuit* indexed by 'class' (i.e., both denote the same specific instantiation of an electric circuit):

(7) [*An electric circuit for receiving a battery of cells in series*]$_{\text{object}}$, *comprising: a plurality of DC-to-AC converters (24A-24H)* [. . .] [*The electric circuit*]$_{\text{object}}$ *wherein the inductive coupling comprises transformer windings (26A-26H). An electrically powered device including* [*an electric circuit*]$_{\text{class}}$ [. . .]

3.1.2. Lexical chain detection in patents

Our lexical chain recognition is based on the rule-based *Stanford Deterministic Coreference Resolution System* (henceforth, StCR) for English; see [23, 24, 20]. The decision to use StCR was based, on the one hand, on its high performance on general discourse (it ranked first in the 2011 competition on coreference resolution at the Computational Natural Language Learning Conference), and, on the other hand, on the lack of annotated corpora in the patent domain that could be used to implement a supervised machine learning-based lexical chain detection. To integrate StGR into the summarizer's GATE infrastructure, a GATE plug-in was created as a wrapper around the StCR code. The plug-in integrates the StCR into our patent processing pipeline by converting our annotations into a format accepted by StCR, and the lexical chains delivered by StCR into GATE annotations.

The central idea behind the StCR approach to coreference resolution is the application of successive independent models (*sieves*) of decreasing confidence, so that coreference matches for which the system has a higher confidence are coped with first, and further matches are detected on the basis of the earlier matches. Sieves are based on features extracted from entity mentions in the text and their context, including surface forms, shallow linguistic traits, deep linguistic analysis and semantic features obtained from Wikipedia info boxes, Freebase and WordNet. Features are not evaluated together. Instead, they are separated into different sieves according to their contribution to a secure identification of coreferences.

StCR consists of two main stages:

[7] http://www.wipo.int/classifications/ipc/en/est/.

(i) Entity candidate detection: A high-recall algorithm is used to collect a large number of mentions of entity candidates based on nominal, pronominal and named mentions in the text. This stage includes a filtering stage to exclude undesirable mentions such as partitives, numericals, bare NPs, etc.

(ii) Coreference resolution and lexical chain detection: Sieves are applied from highest to lowest confidence to the set of mentions. For each sequence of matched mentions, only the first mention in the sequence is left as a candidate, while the rest is removed from the candidate pool.

In a third post-processing stage, singleton sequences are eliminated, thus leaving only chains with at least two mentions.

In order to adapt StCR to the patent domain, a number of changes had to be carried out; see [25] for more details. The first change concerned the replacement of the Stanford's original processing pipeline, which is aimed at general discourse and performed poorly on patent material. Our replacement pipeline is largely based on Bohnet's [17] dependency parser, which is better suited to handle the very long sentences found in patent texts [18]. Furthermore, we used a modified version of GATE's Annie tokenizer [16], followed by the PoS tagger, lemmatizer and parser taken from Bohnet's parsing environment. Other changes concerned the patent-specific mention detection, the adaptation of the sieves and creation of new sieves. In the remainder of this subsection, we focus on the most significant of these changes.

Entity candidate detection. As pointed out above, in the case of patent material, entity detection is equivalent to NP mention detection. NPs in patents can be very long, with multiple NPs embedded in a head NP. Consider, e.g., *an encoding device having an input coupled to the input terminal and having an output, which encoding device comprises* ... in (1) or *the transmission of data between a USB host and a USB peripheral device*, where *a USB host* and *USB peripheral device* are subordinated to *between*, which is governed by *transmission*, which is in its turn modified by *of data*.

Most of the common strategies to detect NPs are based on part-of-speech (PoS) and lemma information and produce flat NPs (rather than embedded structures); see, among others, [26, 27, 28, 29]. This is not appropriate for NPs in patent discourse, where NPs are often of hierarchical nature. Thus, in the example above, we need to be able to identify that *coupled to the input terminal* modifies *input* and *having an input coupled to the input terminal* modifies *encoding device*. Furthermore, the common strategies tend to use personal pronouns and named entities. Personal pronouns *de facto* do not appear and named entities are not important in patents; their presence is limited to some references to patents or companies in the State of the Art section of the patent.[8] Therefore, we developed a hierarchical NP detector specifically designed for the patent discourse and the peculiar syntactic structures of NPs in patent sentences. The detector is based on a set of rules that use PoS and dependency relations of the tokens to detect the head candidates and then expand them by their dependants.

To detect the heads, nouns and pronouns that are not modifiers of other nouns are considered; cf. a sample rule:

IF (AND token τ's PoS is NNP or NN or NNS
τ is governed by a relation different from NMOD or APPOS)
THEN τ is head of an NP

The heads are then expanded inspecting their dependants. Depending on the type of dependency, position (before of after) and PoS, the dependants are considered to form part of the core NP or of a sub-NP that is embedded in the core NP. At the same time, some metadata are added to the NP—for instance, that it includes a definite or indefinite determiner; cf.:

IF token τ_2 modifies a noun τ_1 by the relation NMOD
THEN include τ_2 into the NP headed by τ_1

---

[8] This is also why we could not reuse in our application the candidate detection algorithm provided with StCR.

Prepositional modifiers, as well as participles that are posterior to the head, form a longer NP that creates a new level of NP embeddedness. Thus, as the next example shows, an NP can include several embedded NPs:

(8) [[[*the type*]$_{\text{NP1}}$ *of* [*heat source*]$_{\text{NP1}}$]$_{\text{NP2}}$ *used to heat* [[*the surface*]$_{\text{NP1}}$ *of* [*the food*]$_{\text{NP1}}$]$_{\text{NP2}}$]$_{\text{NP3}}$

The NP detector output contains multiple levels of embedded annotations. Since for the next stage StCR requires partial consitutency-based analyses of sentences, the NP detector output is converted into a phrase-based structure consisting of a sequence of hierarchally embedded NPs.

Co-reference resolution and lexical chain detection. The original StCR system applies eleven main sieves for coreference resolution and one for lexical chain derivation in the following order:[9] 1. Discourse Processing, 2. Exact String Match, 3. Relaxed String Match, 4. Precise Constructs Match, 5.–7. Strict Head Match, 8. Proper Head Word Match, 9. Alias, 10. Relaxed Head Match, 11. Lexical Chain derivation and 12. Pronoun Match.

From the twelve sieves that compose the StCR, except the sieve for detecting the antecedent of the deictic *I*/*you*, all sieves were included in the same order, albeit with some modifications. In addition, a new sieve was created that detects mentions related via copula. Furthermore, the following specific adaptations have been implemented (see also [25], in particular for the evaluation of the performance of the adapted StCR):

- Sieve 1 (Discourse Processing Sieve): This sieve was removed, as patents do not include the conversational text that is the target of this sieve.
- Sieve 2 (Exact String Match): This sieve showed a high precision and was left as it is in the original configuration; cf.:

  (9) *The second arrangement is based on the first one, except that instead of pillars working under a compression load, chains are used that work under* **[*a tensile stress*]**$_i$. [...] *When the platform rises, the chain pulls to make the hydraulic cylinder work under* **[*a tensile stress*]**$_i$, *while when the platform descends again the hydraulic cylinder is compressed by a counterweight placed on the inner end.*

- Sieve 3 (Relaxed String Match): Bare NPs are linked when they occur at the beginning of sentences; cf.:

  (10) *Figure 1, although it shows* **[*the first preferred embodiment*]**$_i$, *includes the various essential elements of the'invention that are common to all the embodiments.* [...]
  The form of arranging and operating said hydraulic cylinders define four possible embodiments of the same invention.
  [*First embodiment*]$_i$*: In this embodiment the platform is anchored* [...].

- Sieve 4 (Precise Constructs): The detection of appositives has been disabled; in addition, relative pronouns are assumed to refer back to the nearest antecedent mention, whilst relational pronouns such as *one another* or *each other* are assumed to refer back to the nearest plural antecedent mention.

  (11) [...] *and that is driven from* **[*an intermediate point*]**$_i$ *in* **[*which*]**$_i$ *is placed the jointed union of the end of a pillar (4) meant to anchor the platform (1) to the sea bed (5);*

- Sieve 5 (Strict Head Match): This sieve is kept as it was, but the list of stop words was adapted to the patent domain.

  (12) [...] *so that this chain pulls on* **[*the hydraulic cylinder*]**$_i$ *(3) as the platform (1) rises and the counterweight (14) compresses and restores* **[*said hydraulic cylinder*]**$_i$.

- Sieves 6 and 7 (Variants of Strict Head Match): The mention and its antecedent are not required to match in number so as to allow set-member relations.

[9] See [23] and [24] for a detailed presentation.

(13) *The rocking motion is transmitted to* **[*the hydraulic cylinder*]**$_i$ *(3) at a point (7.2) that is internal* [. . .]
Third embodiment: In this third embodiment of the invention, shown in figure 3, the connecting rods are eliminated and rigid pillars (4) are connected directly to **[*the hydraulic cylinders*]**$_i$ *(3).*

- Sieve 8 and 9 (Proper Head Word Match and Alias Match): These sieves were disabled, given that neither proper nouns nor aliases appear in patents.
- Sieve 10 (Relaxed Head Matching): Is kept as it was.

(14) **[*The height oscillations of the platform*]**$_i$ *are transformed into displacements in either sense of the hydraulic cylinder.* [. . .]
The last alternative for the invention alters the position about which pivots the connecting rod that transmits **[*the platform oscillations*]**$_i$ *to the impulsion hydraulic cylinder.*

- Sieve 11 (Lexical Chain): The lexical chain sieve was modified so as to use elaborate patterns for detection of part-whole relations and instance-of relations between mentions within the same sentence.

(15) **[*A wind turbine generator*]**$_{whole}$ *comprising:* **[*a rotor*]**$_{part}$ *comprising a hub,* **[*at least one rotor blade*]**$_{part}$ *coupled to said hub, and* **[*a rotor shaft*]**$_{part}$ *coupled to said hub for rotation therewith;*

- Sieve 12 (Pronominal Coreference resolution): This sieve was modified according to the occurrence of the pronouns in patents. Most personal pronouns were excluded and adverbial pronouns were included.

(16) **[*A control system*]**$_i$ *(40) in accordance with Claim 1* **[*wherein*]**$_i$ *said at least one selection mechanism comprises at least one of: at least one discrete switch* [. . .]

Unlike in general discourse, where copulative constructions <NP$_1$> *is* <NP$_2$> are detected in the precise pattern-based construct, in patents, this construct needed its own sieve, applied towards the end of the sieve sequence. This is because both mentions NP$_1$ and NP$_2$ can be involved in separate lexical chain relations, such that an early merge into a single chain would make the subject NP$_1$ the antecedent of the chain, while NP$_2$ would become unavailable for further linking. The new copula sieve distinguishes between attributive relations where the copulative construction indicates a property of the subject (see (17a) below for illustration) and identity relations (see (17b) for illustration):

(17a) [*the present invention*] *is* [*a wind turbine generator*]$_{\text{attribute}}$

(17b) [*the difference of the wind direction deviation*]$_{\text{i}}$ *is* [*the error of the anemoscope 6 due to drift wind*]$_{\text{i}}$

The post-processing stage was adapted to post-process the obtained clusters. First, clusters that contain a mixture of reference relations have been divided into different chains. Second, if compatible mentions (i.e., same head, same attributes, and one smaller mention included in the other larger one) are used to establish different cohesion relations, their clusters are merged into one.

The lexical chain recognition has been evaluated in qualitative terms with respect to the satisfaction of the user; correctness, completeness, transparency, etc. were assessed. In the average, it was rated 3.6 on the 1 (bad). . . 5 (excellent) scale.

3.2. Segmentation and segment alignment

As already argued above, in the patent domain, subsentential segments that capture statements on the invidiual entities of the invention are more appropriate as summarization units than entire sentences. First, they allow for a considerably more fine-grained selection of content to be included in the summary (recall the length of the sentences in Claims). Second, when aligned based on their similarity, they allow for grouping all of the information distributed across the entire patent on a given entity and thus for a more targeted relevance assessment and content selection.

3.2.1. Segmentation

Segmentation is done by means of a set of linguistic rules, these rules are based on the characteristics of the language used in patents. When writing these rules we took into consideration the style guidelines provided by the different patent offices on how a Claim must be written see [14] or [30]. As the objective of the segmentation is to split the long sentences into smaller parts eligible for the summary the rules should not break the statements but is not problematic if they miss some splits. For this reason when designing the rules we took a conservative approach.

Due to the different writing styles of Claims and the Description, the strategies for their segmentation must be also different. Below, we thus first discuss the segmentation of Claims and then the segmentation of the Description.

Claim segmentation. Given that claims are structurally divided into parts that describe elements of an invention, the detection of claim segments can be done using surface features. The most dominant surface features are clearly identifiable delimiter keywords (such as, e.g., *comprising*, *consists of*, *contains*) and the punctuation symbols ":", ";" and ",". Consider, for illustration the claim in (18):

(18) *A wind turbine generator* ***comprising:*** *a plurality of wind-turbine rotor blades for receiving wind power* ***;*** *a rotor head to which the plurality of wind-turbine rotor blades are attached, the rotor head being rotated and driven by the wind power received by the plurality of wind-turbine rotor blades; a head capsule for covering the rotor head****;*** *a blade-side disc portion that extends radially outward of each wind-turbine rotor blade from the wind-turbine rotor blade and is inclined toward the head capsule; a cylindrical portion that is disposed substantially coaxial with each wind-turbine rotor blade and extends from the head capsule toward the tip of the wind-turbine rotor blade; a capsule-side disc portion extending from each cylindrical portion radially outward of the wind-turbine rotor blade****; and*** *a bent portion that extends radially outward from the outer circumferential edge of each capsule-side disc portion and is inclined toward the tip of the wind-turbine rotor blade.*

As can be observed, (18) can be naturally divided into segments by ":", ";", "comprising" and "and": [*a plurality of wind-turbine rotor blades for receiving wind power*], [*a rotor head to which the plurality of wind-turbine rotor blades are attached*], [*the rotor head being rotated and driven by the wind power received by the plurality of wind-turbine rotor blades*], etc. Punctuation marks such as ":", ";" and keywords such as "comprising" can be considered "accurate" or "reliable" segments delimiters because they always delimit a claim segment. Opposed to them are "unreliable" delimiters such as "," or "and", which are used not only to delimit segments of the invention, but also to connect simpler linguistic structures; cf., e.g., *the first handle* ***and*** *the second handle* or *first handle****,*** *second handle****,*** *third handle****,***. . . .

The segmentation furthermore depends on the internal pattern structures of the claim. We identified three predominant patterns:

1. <Element> <primary keyword >":" $[\text{segment}]_1$";" $[\text{segment}]_2$";" $[\text{segment}]_3$; . . . ,

with 'Element' as an element of the invention, 'primary keyword' as "comprising", "contains" or another accurate delimiter and 'segment_i' as a segment separated by ";". Consider, for instance:

(19) *<A sail for a sailboard> <comprising>*: [*a sail body having a luff and a leach*]$_1$; [*a luff pocket attached to the sail body for receiving a sailboard mast*]$_2$; [*a batten carried on the sail body between the luff and leach and having a batten tip extending forward of the luff within the luff pocket*]$_3$; [*a batten tensioner located on the sail for applying tension to the batten*]$_4$; [*a cam body located in the luff pocket for rotateably engaging the batten with a sailboard mast*]$_5$,

where *a sail for a sailboard* is the element, *comprising* is the keyword and the five segments follow the ":".

2. <Element> <primary keyword >":" $[\text{segment}]_1$ <primary keyword> $[\text{segment}]_{1.1}$";" $[\text{segment}]_{1.2}$";" . . . ,

with 'Element' and 'primary keyword' defined as above and '$[\text{segment}]_{i.j}$' as segment j embedded into segment i. Consider, for illustration:

(20) *<A hydroelectric generator system having fault isolation> <comprising>*: [*at least one off-shore hydroelectric turbine generator*]$_1$ *comprising* [*a direct-drive shaftless permanent magnet generator having a plurality of conductive coils and a plurality of magnets*]$_{1.1}$; [. . .],

where *a hydroelectric generator system having fault isolation* is the element, *comprising* is the keyword, *at least one off-shore hydroelectric turbine generator* is the first level segment, and *a direct-drive shaftless permanent magnet generator* ... is the embedded segment.

The third pattern contains either none or just one of the symbols ";" and ":", but never both at the same time. This makes automatic detection of segments of this kind more difficult because it has to rely on such keywords as "comprising" to find segment boundaries. However, most cases are covered by a relatively small variety of keywords, which is why even in the absence of ";" and/or ":" the alignment is possible. The pattern is as follows:

3. <Element> < primary keyword> [segment]$_1$ <primary keyword>|",” [segment]$_2$ <primary keyword>|",” ...

The following example illustrates this type of pattern:

(21) *<An excavator in which a lower traveling body is equipped with an upper rotating body thereon, and an excavating attachment is provided on the upper rotating body>*, *<comprising>* [*an engine as a power source*], [*a generator driven by the engine*], ...

An evaluation of the division of Claims into segments and the recognition of important elements consisted in comparing the algorithm results with a manual gold-standard, composed of 146 claims and 446 segments. It yielded a result of 0.95 precision, 0.65 recall, and F1-score = 0.77. This can be considered a satisfactory outcome such that the division of claims into segments and the recognition of important elements of the invention can be used for the automatic alignment of claim segments with their corresponding segments in the Description.

Description segmentation. The main purpose of the segmentation of the Description is to identify statements that elaborate on elements introduced in claims. We assume that these statements are finite clauses, i.e., contain a finite verb and are delimited by a punctuation symbol (".", ";", ",", ":"), a conjunction "and", or a disjunction "or". For the sake of simplicity, we divide the segmentation procedure into three stages; cf. Algorithm 1.

Algorithm 1: Algorithm alg:segm: Description segmentation

1. Collect all sequences of tokens s separated by a segment delimiter
   $\delta \in$ {".", ";, ",, ":", "and", "or"}, from all sentences of the Description, into the list of segments Σ.
2. **for** $\forall s \in \Sigma$: **if** s contains a finite verb
   tag s as "independent",
   else tag s as "dependent"
3. **for** i = 1 to $|\Sigma| - 1$ **do**
   if $s_i \in \Sigma$ is tagged as "dependent"
   merge s_i with s_{i+1}
   keep the tag of s_{i+1} for the merged element
   endif

Consider, for illustration, the resulting segmentation of a paragraph of the Description of patent chosen randomly:

(22) [*The following method is normally adopted in the earlier technology to detect the charge state of a battery*]. [*Namely, the battery charge state is ascertained by measuring the voltage at the battery*], [*the battery charge state thus ascertained is set as an initial value*] *and* [*the battery charge state is subsequently calculated as necessary by totalizing the charge / discharge quantity as the vehicle travels*].

3.2.2. Claim–Description segment alignment

A claim segment and a description segment are aligned if they are sufficiently similar, i.e., if their similarity is above a given threshold.[10] To obtain this similarity, we need to compute:

[10]The purpose of this task is similar to that of the task of similarity-based paragraph clustering in [5, 31].

(i) the similarity between the entities of the invention in the claim segment and in the description segment,

(ii) the lexical overlap between the claim segment and the description segment.

The computation of the similarity between two linguistic structures which both denote entities of the invention but are lexically not identical is not a trivial task. Due to the significant difference between the linguistic structures of segments of object and method claims, we use two different similarity functions, one for object segments and one for method segments. The similarity function for object claims is based on the cosine overlap function:

$$Sim(S_C, S_D) = \begin{cases} 1.0 & | \text{ if `(' N`)'} \in S_C \wedge \text{`(' N`)'} \in S_D \\ \frac{\sum_{j=1}^{n} sc_j \bullet sd_j}{\sqrt{\sum_{j=1}^{n} sc_j^2} \sqrt{\sum_{j=1}^{n} sd_j^2}} & |\text{otherwise} \end{cases} \quad (1)$$

where 'S_C' is the claim segment, 'S_D' the description segment, $sc_j \in S_C$, $sd_j \in S_D$ (with 'sc_j' and 'sd_j' as individual tokens), '(' N')' a token sequence, and N a figure reference.

Thus, [*separating means (21)*] and [*main switch (21)*] have the similarity 1.0 since they share the same figure reference. In other words, we assume that *separating means* and *main switch* refer to the same component of the invention, even if they do not share any lexical element. [*fuel vapor discharge apparatus*] as claim segment and [*fuel vapor recirculation apparatus*] as description segment have a cosine similarity of about ~ 0,75: both segments mention an apparatus that does something with vapor.

The similarity function for method claims takes the linguistic structure of the segments into account: the method segments are composed by a gerund verb that expresses an action and at least its first argument. Because the same or a related action can be expressed by different verbal lexemes, it cannot be presupposed that both the claim segment and the description segment contain the same verbal lexeme, even if they are semantically similar. However, what can be expected is that if their corresponding first arguments are complex, they will overlap lexically. For this purpose, the linguistic composition of the first arguments is drawn upon to compute the similarity:

$$Sim(S_C,S_D) = \begin{cases} HM(S_C, S_D)\times 0.25 + VM(S_C, S_D)\times 0.5 + AO(S_C, S_D)\times 0.25 & |\text{if } HM(S_C, S_D) \neq 1 \\ 0 & |\text{otherwise} \end{cases} \quad (2)$$

where 'HM' is a head matching function that returns '1' if the stems of the head of the argument of S_C and S_D match, and '0' otherwise, 'AO' is an "argument overlap" function which computes the Jaccard coefficient between the tokens in the first arguments of S_C and S_D. The multiplier '0.25' assures that the contribution of the lexical overlap can be at most 0.25. If the stems of the heads of the structures match, the score is increased by 0.25. 'VM' is the "verb match" function returning '1' if the stems of the verbs of S_C and S_D match, and '0' otherwise. That is, if the stems of the heads of the segments match, the score is increased by 0.5.

Two segments can only have a perfect similarity score if their verb heads match, the heads of their arguments match and their first arguments contain exactly the same tokens. For instance, the segments [*determining a voltage change*] and [*determines voltage change*] have a similarity score of 1 because the stem of the verb is in both *determin*, the stem of the argument head is in both *voltage* and the modifier of the head is also the same. The segments [*measuring a temperature difference*] and [*measuring a voltage difference*] have the similarity of 0,833. The similarity is high because in both segments the verbal head and the argument head are the same although the measured dimension is different.

4. Segment relevance assessment and segment selection

In order to assess the relevance of the individual segments for their inclusion into the summary, we follow an extractive summarization strategy in that we compute the relevance of a segment based on a number of distribution- and position-oriented features.

4.1. Segment relevance features

Three types of features are used for the computation of the relevance of a given segment for the summary: (i) mention and lexical chain-oriented features, (ii) segment-oriented features, and (iii) classical similarity features. All features are summarized in Table 1. Let us discuss each of them in turn[11].

Type	#	feature	description
Mention/ lex.chain features	1.	mention frequency	relative frequency of m in a section m (f_m/N)
	2.	coreference chain length score	relative length of m's coreference chain Ch ($coefRel(m) = \|Ch\|/N$
	3.	meronym/holonym and hyponym/hyperonym chain score	m that expresses a holonym/hyperonym →1.0; m that expresses meronym/hyponym → 0.5
	4.	claim structure relevance	the depth L of the claim in the claim dependency structure in which m occurs (1/L)
Segment-oriented features	5.	best segment alignment similarity	highest cosine similarity between s and any claim segment s_C
	6.	second best segment alignment similarity	second highest cosine similarity between s and any claim segment s_C
	7.	segment length relevance	number N of mentions in s ($1/N$)
	8.	segment position relevance in claims	s is a claim segment → 1.0; s is not a claim segment (i.e., is a background, drawings, ... segment) → 0.5
	9–13.	invention segment	s is a segment in which the invented method/apparatus is introduced → 1.0; s is not a segment in which the invented method/apparatus is introduced → 0.0
Classical features	14.	similarity to the summary	similarity of s with the summary, measured as cosine
	15.	similarity to title	similarity of s to the title, measured as cosine
	16.	similarity to claims	similarity of s to the claims, measured as cosine
	17.	mention distribution in claims	sum of $tf * idfs$ of m in the claims
	18.	mention distribution in abstract	sum of $tf * idfs$ of m in the author abstract
	19.	mention distribution in description	sum of $tf * idfs$ of m in the description

Table 1. Features used for the assessment of segment relevance (m: mention under consideration, s: segment under consideration); the number of the feature (2nd column) refers to the number in Table 2.

4.1.1. Mention and lexical chain-oriented features

Mention and lexical chain-oriented features cover: 1. mention frequency, 2. distribution of mentions in coreference (i.e., identity) chains, 3. distribution of mentions in meronymy/holonymy and hyponymy/hyperonymy chains, and 4. mention position in the patent structure.

1. *Mention frequency feature:* Term frequency is traditionally considered to be a key feature for the creation of automatic summaries [32, 33]. We compute the relative frequency of each mention in a section of a patent (Abstract, Claims, Description, etc.) as the frequency f of the mention m divided by the total number of mentions N in the section: f_m/N.
2. *Coreference chain features:* There are several ways in which an entity of the invention can be referred to in discourse. For instance, both *optical device* and *camera* refer to the same entity, although they are different terms (see also Section 2.1.2 above). Lexical chains that associate different mentions of the same entity via coreference relations provide a means to measure the relevance of an entity (and thus of the mentions that denote it) in the patent. Coreference chain features such as length and spread have been exploited in [11] in the information extraction context. In our experiments, length proved to be the most relevant one. Given a mention

[11] Note that summation features and others which could produce values above 1 are normalised usin formula (6)

m and the coreference chain *Ch* it occurs in, *m* is assigned the coreference relevance feature *corefRel*(*m*), whose value is the ratio of the length of *Ch*, i.e., the number of mentions in *Ch*, divided by the total number *N* of mentions in the patent under consideration:

$$corefRel(m) = |Ch|/N \tag{3}$$

We also experimented with the position of *m* in *Ch* (i.e., 1, 2, ...) and with a location relevance feature, whose value is inversely proportional to the position of *m* in the chain (that is, as in clasical summarization approaches, mentions appearing earlier in the chain have been considered more relevant): $1/\sqrt{P(m, Ch)}$ (with $P(m, Ch)$ as the function that gives us the position of *m* in the coreference chain *Ch*). However, neither of them led to an improvement of our overall relevance metric.

3. *Meronymy/holonymy and hyponymy/hyperonymy chain features:* Another source of information that we take into account in order to assess the relevance of a mention are the 'part–whole' and 'is-a' relations between mentions, captured in the meronymy/holonymy and hyponymy/hyperonymy chains. For illustration, consider the 'part – whole' relation between [*the conventional differential device*] and [*a housing*] in *The conventional differential device comprises a housing that ...*, and the 'is-a' relation relation between [*image multiplier plate*] and [*optical image multiplier member*] in *The image multiplier plate I is an optical image multiplier member ...*.
   We consider these two relations of particularly high importance because of the distinctive content they introduce. The corresponding features of the mention *m* that expresses the holonym/hyperonym receive the weight of 1.0, while *m* that expresses the meronym (part or component)/hyponym receives the weight of 0.5.
4. *Claim structure relevance feature:* Finally, we draw upon the structure of the claims in order to measure mention relevance: (i) mentions in the Claims are considered of higher relevance than those that appear in other sections of the patent, and (ii) the relevance of a mention in a claim is inversely proportional to the level in the claim structure tree at which the entity is situated. More precisely, the value of the claim structure-related feature of the mention *m* is 1/L, where 'L' is the level of the claim in which *m* occurs in the claim structure tree (with '1' being the level of independent claims).
   The value of the feature is first computed for the mentions within the Claims.[12] If a mention also occurs in other sections of the patent, its claim structure-related feature receives outside the Claims the maximal value obtained for all of its occurrences in the Claims. To decide whether two mentions stand for the same entity, their surface forms are compared.

As our objective is to score the relevance of the segments of an invention based on mention relevance and given that each segment may involve more than one mentions (especially in the Description), an aggregated score A_F is calculated for each individual feature f:

$$A_F = \sum_{i=1}^{n} \frac{f_i}{\sqrt[K]{n}} \tag{4}$$

where n is the number of mentions in the segment, f_i is the feature value of the mention i, and K is a smooth factor that balances the weight of segments that contain many mentions compared to those that contain only few mentions. In our experiments, K has been set to 2.

4.1.2. Segment-oriented relevance features

In addition to scoring the relevance of segments based on the mentions they contain and the lexical chains of the mentions, we also assess their relevance drawing upon a series of other segment-oriented features:

1. *Segment alignment similarity features:* The strength of the association of segments in one of the descriptive sections of the patent with claim segments is measured in terms of their similarity to claim segments. Two features are used for this purpose. The value of the first is the maximal cosine similarity of a given segment with any claim segment; the value of the second is the second highest cosine similarity of this segment with any claim segment.

[12] Note that mentions that not appear in Claims are not considered here. This is in accordance with our assumption that the content selection for the summaries is claim-driven.

2. *Segment length relevance feature:* In order to avoid that long segments with many mentions are given too much (unjustified) relevance when compared to segments composed of few mentions, we use a feature that reflects the length of the segments. Its value is 1/N (with N as the number of mentions in the segment). Similar features have been used in summarization in the past [34].
3. *Segment position relevance feature:* It is well known that in specific textual genres the position of the information within the text (at the beginning, in the middle, at the end) is very telling with respect to the relevance of this information for the summary. Therefore, we assume that it is relevant to capture whether a segment appears in the Claims or in any of the other sections of the patent. The segment position relevance feature of a segment is assigned the value 1.0 if this segment is a claim segment and 0.5 otherwise.
4. *Invention segment feature:* Entity recognition allows us to identify the segment that introduces the method or the apparatus that is patented. The 'invention feature' of this segment is assigned the value 1.0; the 'invention feature' of all other segments receives the value 0.0.

4.1.3. Classical Summarization Features

In addition to patent-specific features discussed above, we compute for each patent a number of relevance measures using the summarization library SUMMA [35]. Many of SUMMA's features are based on the computation of similarities between different units (sentences, clauses, sections, etc.) in a document. In order to compute the similarities, the units are represented as vectors of term stems and weights (with stop words and other meaningless units removed using domain-specific tables). The weight of each term (stem) in the vector representation is calculated using *tdf*idf*, with *idf*, i.e., the inverted document frequencies, obtained from external patent collections. The similarity between units is calculated as the cosine of the angle between two vector representations. In our application, the computed similarity features capture the similarity of a given segment with the author Summary of the patent, the Title of the patent, and the Claims section of the patent.

Further classical features concern the mention distribution, which are the sum of *tf*idfs* of the mentions in the author Abstract, Claims, and Description, respectively.

4.2. Segment selection metric

To use features introduced above for summarization, a summarization metric assigns to each segment of a patent document a cumulative score of the weighted values of its features using the following formula [35]:

$$score(segm) = \sum_{i=0}^{n} w_i * f_i \tag{5}$$

with *segm* as the segment in question, f_i as the value of its feature i, w_i as the weight assigned to i, and n as the total number of features. Note that in our approach all features which could produce values above 1 are normalised so as to produce values between 0 and 1. The normalisation is carried out per each patent using the following formula:

$$fv = \frac{fv - min}{max - min} \tag{6}$$

where fv is the value to normalise, *max* is the maximum possible value for the feature in the patent, and *min* is the minimum possible value for the feature in the patent (when *max* and *min* are identical, then fv is defined as zero).

To obtain the weights for the individual features, we use training data and a linear regression procedure [36]. The training data consist of the cosine similarity figures between segments of the patents from a training set and the gold standard summaries of these patents. The linear regression procedure suggests for the individual features a weight distribution w_i that approaches the cosine similarities of the segments of the patents in the training set with their respective gold standard summaries. Table 2 summarizes the weights of the features as used in the final application of our summarization module.

The linear regression model computed with 26,498 datapoints (i.e. sentences extracted from a set of patents containing gold standard human created abstracts) has a coheficient of determination of 0.68 and is, according with an ANOVA analysis, statistical significant with 99% cofidence level which is good from the predictive standpoint. Many of the variables of the model are also significant wih a 99% confidence level, notably, variables related to classical summarization features such as the similarity of a segment to the summary (feature 14), the similarity of a segment

		#	description	weight
Mention/ lex.chain features		1.	mention frequency	0.1842
		2.	coreference chain score	0.0665
		3.	meronym/holonym and hyponym/hyperonym chain score	0.2270
		4.	claim structure relevance	0.0202
	Segment-oriented features	5.	best segment alignment similarity	-0.0068
		6.	second best segment alignment similarity	0.0250
		7.	segment length relevance	0.0143
		8.	segment position relevance in claims	0.0000
		9.	segment position relevance in background	0.0498
		10.	segment position relevance in drawings	-0.0318
		11.	segment position relevance in embodiment	-0.0214
		12.	segment position relevance in summary	-0.0265
		13.	invention segment	0.2830
	Classical features	14.	similarity to the summary	0.6025
		15.	similarity to title	0.1597
		16.	similarity to claims	0.0000
		17.	mention distribution in claims	-0.3397
		18.	mention distribution in abstract	-0.2544
		19.	mention distribution in description	0.5101

Table 2. Weights of the features used in summarization as obtained by linear regression

to the title (feature 15), and the distribution of mentions in claims (feature 17) and description (feature 19). Also mention-based features which are novel to this work are relevant such as mention frequency (feature 1), coreference (feature 2), and lexical relations (feature 3). From the set of position-based features only significant resulted the position of the segment in the background section. The similarity of a description segment to a claim (feature 5) is also significant as is the length of the segment itself (feature 7). All other features have less explanatory power in the regression model according to our observations.

Segments with the highest score according to equation (5) across the whole patent are included in the summary. The length of the summary can be controlled by the inclusion of more or less top-ranked segments. Consider, for illustration, Figure 4, which displays the segments selected from the patent EP2700814 A1 for inclusion into the summary, based on (5).

5. Summary generation

In the summary generation stage, the selected segments are fused into a coherent summary. As summary generation module, we use the graph transducer-based MATE generator [37]. Apart from the list of the selected segments, their relevance score, origin (Claims, Description, ...) and linear order of their appearance, summarization generation draws upon the context of the selected segments, namely: (a) the sentences in which the individual segments appear, their syntactic dependency structures, linear order of their tokens as well as the morphosyntactic features (PoS, number, gender, lemma) of the tokens; (b) the type of the invention described in the patent (apparatus or method); (c) all detected mentions and the lexical chains these mentions are involved in.

This information is used for

1. completion and grammatical adjustment of individual segments to grammatical sentences;
2. targeted removal of parts of segments or of whole segments from the summary in case they cannot be completed to grammatical sentences;
3. increase of the cohesion between the completed sentences.

[*Glider for electric power production from wind*]

[*transferred to the ground via the tether into electric power*]

[*Therefore , a larger amount of the total lift force generated by the airfoil is available for electric power production*]

[*control device also allow for automated optimization of the flight , in particular in order to maximize the lift force during the energy production phase and in order to minimize the pull on the tether during the recovery phase*]

[*The underlying problem of the invention is to provide for electric power production from wind using an airborne airfoil , wherein*]

[*the electrical machine constructed for converting a lift force generated upon exposure of the airfoil to wind and transferred to the ground via the tether into electrical power*]

[*which produce lift forces upon exposure to wind*]

[*One of the challenges of airborne wind energy production is the transferal of energy extracted from the wind at high altitudes to the ground*]

[*which is continuously or repeatedly taken during the flight and*]

Figure 4. Segments selected for inclusion into the summary from patent EP2700814 A1

The three types of operations, which are discussed below, are implemented as MATE-grammars, which draw upon application-oriented syntactic and morphological dictionaries. The output of the summary generation module is a coherent and maximally cohesive summary, as shown in Figure 5.

What is claimed is a glider for electric power production from wind. A larger amount of the total lift force generated by the airfoil is available for electric power production. The underlying problem of the invention is to provide for electric power production from wind using an airborne airfoil. The electrical machine is constructed for converting a lift force generated upon exposure of the airfoil to wind and transferred to the ground via the tether into electrical power. One of the challenges of airborne wind energy production is the transferal of energy extracted from the wind at high altitudes to the ground.

Figure 5. Summary generated from the segments selected for the inclusion into the summary from patent EP2700814 A1

5.1. Sentence completion

Sentence completion consists of the following operations 1–4:

1. Introduction of the missing nominal (group) head of a segment. If a segment starts with a gerund or a bare infinitive with no syntactic governor, the corresponding head (governor) is recovered from the sentential context of the segment and the verb is inflected. If the segment starts with a relative clause, the relative pronoun is replaced by the recovered head. Thus, any of the three segments below is realized as indicated after the right arrow:

(23) [*contain a signal processing unit*]
[*containing a signal processing unit*]
[*which contains a signal processing unit*] } ⇒ *The device contains a signal processing unit*

If the segment starts with a past participle, the copula *be* is furthermore introduced in order to create a passive sentence; for instance:

(24) [*contained in a rectangular device*] ⇒ *The unit is contained in a rectangular device*

2. Completion of the invention segment. The first segment that carries an 'Invention' feature (see previous section) is completed to a sentence by the introductory *What is claimed is*. The following segments that carry the 'Invention' feature are introduced by *The invention covers*. Segments that are identified as a component of a device or a step of a method are completed by *The device contains* or *The method consists in*, respectively. Consider, for illustration, the following three consecutive segments extracted from a patent with their relevant features, and the corresponding generated text:

(25) [*a sail for a sailboard*]$_{invention=initial}$
[*a device for coupling a sail batten to a mast in a board sail*]$_{invention=\neg initial}$
[*a first end for rotateably bearing against a mast*]$_{component=yes}$
$\Rightarrow$ *What is claimed is a sail for a sailboard. The invention covers a device for coupling a sail batten to a mast in a board sail. The device contains a first end for rotateably bearing against a mast.*

3. Syntactic completion. If a segment is composed of a complete nominal syntagm with a verb dependant in gerund, infinitive or participle, it is transformed into a sentence in that the gerund and the infinitive are inflected and the past participles are connected with *be*. Consider (26) and (27):

(26) [*a device* (*for*) *containing a signal processing unit*]
[*a device to contain a signal processing unit*]
[*a device which contains a signal processing unit*]
$\Rightarrow$ *A device contains a signal processing unit*

(27) [*a unit contained in a rectangular device*] $\Rightarrow$ *A unit is contained in a rectangular device*

In the case of the first segment of the summary, an introductory *There is* is added; cf. [*a unit for the shading of the light*] vs. *There is a unit for the shading of the light*.

4. Fallback completion. If a segment's syntactic structure consists of several unconnected fragments because of the failure of the parser to come up with a connected parse, operations 1–3 cannot be applied. In this case, we introduce the segment with a generic sentence *One feature of the* followed by the nominal head of the segment:

(28) [*a water current power generation system*]$_{invention=initial}$
[*an induction type power generation unit disposed within a housing associated with the flotation chamber*]$_{component=yes|disconnected=yes}$
$\Rightarrow$ *What is claimed is a water current power generation system.* ***One feature of the invention****: an induction type power generation unit disposed within a housing associated with the flotation chamber.*

5.2. Segment filtering

If a segment cannot be completed to a grammatical sentence, for instance, due to its erroneous syntactic parse, two different filtering mechanisms are applied: 1. word-oriented filtering that removes individual words from the segment, and 2. segment-oriented filtering that removes the segment as a whole. This is in order to take into account the outcome of an empirical study that users do not trust summaries with ungrammatical sentences.

Word-oriented filtering removes specific words used as delimiters during segmentation and left at the beginning or end of selected segments. Among others, the following words and groups of words are filtered: *when*, *wherein*, *such that*, *characterized in that*, gerunds, incomplete relative clauses and the coordinating conjunction *and*. Consider, as illustration, the two segments in (29) and (30) and their corresponding form in the final summary:

(29) [*Such damage is most likely to result from debris ingress, in particular between the stator and the rotor* ***where the coils***]
$\Rightarrow$
Such damage is most likely to result from debris ingress, in particular between the stator and the rotor.

(30) [*the shell is a structural support member for frictionally securing the levers* ***such that***]
$\Rightarrow$
The shell is a structural support member for frictionally securing the levers.

Some initial connectors are also removed because they could refer to a sentence which has not been selected for inclusion into the summary:

(31) [***Therefore*** *, a larger amount of the total lift force generated by the airfoil is available for electric power production*]
⇒
A larger amount of the total lift force generated by the airfoil is available for electric power production.

Segment-oriented filtering removes entire segments which have been identified as being ungrammatical, mainly because they are irreparably incomplete. A segment is considered irreparably incomplete if it begins with: (1) a relative pronoun without the noun it modifies, (2) a finite verb with no subject, (3) a coordinating conjunction without the first conjunct, or (4) a non-finite verb (infinitive, gerund or past participle) which is at the same time the syntactic root of the segment and has no first argument in it. Consider, for instance, the following segments which have been removed from the summary in one of our experiment runs:

(32) [***which*** *produce lift forces upon exposure to wind*]
[***which*** *is continuously or repeatedly taken during the flight* ***and***]
[***and permits*** *ease of operation of the rather complex linkage structure while imparting to the scissors linkage a rigidity which permits safety seat belts to be affixed to the seat structure*]
[***includes*** *a light emitting element, a light receiving element and an arcuate shaped optical fiber*]
[***characterised*** *by debris management means for preventing the ingress of debris into the gap and / or actively removing debris from within the gap*]

Segments that are not completed or filtered are left as they are in the final summary, which is why in spite of the good coverage that the grammars provide, some sentences can look awkward or incomplete.

5.3. Increase of sentence cohesion

The increase of the cohesion between the sentences obtained after sentence completion is achieved by generation of anaphora and introduction of discourse markers. Thus, if two consecutive sentences have syntactic subjects which are part of the same coreference chain, the second subject is pronominalized; cf.:

(33) [*The device$_i$ contains a signal processing unit*] + [*The device$_i$ is located next to the shading unit*]
⇒
The device contains a signal processing unit. It is located next to the shading unit.

If in the same kind of structure the verb is also repeated, instead of creating a (possibly heavy) coordinated structure, a discourse marking adverb is added to the sentence:

(34) [*The device$_i$ contains a signal processing unit*] + [*The device$_i$ contains a shading unit*]
⇒
The device contains a signal processing unit. It also contains a shading unit.

6. Evaluation

In order to assess the performance of the proposed patent summarization approach, we carried out two types of evaluation: (i) a quantitative evaluation that draws upon gold standard summaries, and (ii) a qualitative evaluation of the produced summaries from the viewpoint of the user. To assess the competitivity of our approach, we contrast in both evaluations the quality of TOPAS summaries with the quality of the summaries produced by two state-of-the-art techniques, the *Centroid-based* summarizer [38] and *LexRank* [39].

Centroid-based summarization is usually applied in a multi-document summarization context. Therefore, we have adapted the algorithm described in [38] to the single-document summarization setting of TOPAS. We consider a document as a set of sentences (segments), each of which is represented as a normalized vector. A centroid of the sentence vectors is computed using the SUMMA platform [35]. Each sentence (segment) vector is then compared wih the centroid using cosine. The obtained value is used as the centroid-based relevance feature of the sentence/segment. Finally, the sentences with highest cosine similiarity value are selected for inclusion into the summary.

LexRank has been implemented following the algorithmic description in [39]. The only customizations in our context have been the computation of the sentence (segment) connectivity matrix, for which we SUMMA's normalized

vectors [40] and SUMMA's cosine similarity metric. The score obtained running the customized LexRank is used as sentence (segment) relevance feature for selecting sentences for inclusion into the summary.

In both the Centroid-based and LexRank implementations, redundancy is controlled by blocking any sentente from appearing in the summary if its vector is too close to the vectors of sentences already in the summary. The redundancy threshold is set to 0.05. Figure 6 shows sentences selected by LexRank and Centroid (in this specific case both methods selected exactly the same content) and segments selected by our TOPAS summarizer.

Sentences selected by LexRank and Cetroid

[*A full seat adjustable suspension comprising, in combination, a base structure including substantially horizontal rails and an upwardly extending spring bracket having an upper region, a scissors linkage having lower pivots attached to said rails, lower guides guided by said rails and upper guides, a seat frame having both seat bottom and seat back structure and having pockets slidably receiving said linkage's upper guides, lever structure having a pivot end pivotally mounted on said base structure and having a free end engaging said seat frame at a location spaced from said pockets, a slide movably mounted on said lever between said pivot and free ends, extension spring means interposed between said spring bracket upper end and said slide, and adjustment means mounted upon said seat frame operatively connected to said slide selectively positioning said slide upon said lever structure.*]

[*Such cushioning usually utilizes extension or compression springs which support the driver's weight.*]

[*Fig. 5 is a detail enlarged sectional view as taken along Section 5-5 of Fig. 2.*]

[*The lower ends of the inner links 32 is provided with a follower, each of which is mounted within a base guide 24 for close sliding movements therein.*]

[*The similarity of this adjusting structure to that disclosed in Patent 5,601,338 will be appreciated by one skilled in the art.*]

[*The particular construction of the cushions constitutes no part of the present invention.*]

Segments selected by TOPAS

[*both seat bottom and seat back structure mounted upon said linkage for vertical movement therewith*]

[*However, because of variations in the driver's weight, often in excess of 100 pounds, it is difficult to "tune" a spring supported seat suspension to provide optimum comfort*]

[*A seat frame includes linear pockets affixed to the seat frame front edge receiving followers affixed to the upper ends of the scissor linkage levers pivotally mounted to the base, and*]

[*This relationship between a scissors linkage lever and the base permits ease of operation of the rather complex linkage structure while imparting to the scissors linkage a rigidity which permits safety seat belts to be affixed to the seat structure*]

[*The hereinafter described embodiment of the invention further provides a full seat adjustable suspension utilising scissor or other linkages wherein*]

[*However, because of the interrelationship of the studs 88 and slots 90, a high strength connection between the scissors linkage 28 and the base plate 14 is produced which permits safety seat belts to be directly attached to the seat structure*]

[*The support of the seat frame 12 upon the scissors linkage 28 permits the seat frame to move up and down in a substantially linear vertical movement*]

Figure 6. Sentences selected by LexRank, Centroid, and fragments selected by TOPAS for patent 1050428

6.1. Quantitative summary evaluation

For the quantitative evaluation, we assessed the overlap of the nominal and verbal chunks between a baseline, the summaries generated by TOPAS, the Centroid-based summarizer and LexRank and the manually crafted gold standard summaries for 26 patents chosen arbitrarily from the Green Inventory IPC of the WIPO (see also footnote 7). Table 3 shows the average figures for precision, recall and F-score. The table also includes the difference between TOPAS and the other techniques. The figures show that the quality of the selection of the content by TOPAS is slightly above the quality obtaind by the state-of-the-art techniques.

	Baseline			LexRank			Centroid			TOPAS		
	p	r	$F1$	p	r	$F1$	p	r	$F1$	p	r	F1
	0·34	0·36	0·34	0·49	0·45	0·45	0·45	0·43	0·44	0.49	0.46	0.47
Δ	0·156	0·105	0·131	0·05	0·019	0·015	0·036	0·031	0·029	–	–	–

Table 3. Quantitative evaluation of the quality of the summaries (p = 'precision', r = 'recall', $f1$ = 'F1-measure). Δ row shows the increment of precision, recall and F-Measure of our summarizer (TOPAS) with respect to the baseline and the other two summarizers

6.2. Qualitative summary evaluation

Two qualitative evaluations of the produced TOPAS patent summaries have been carried out. In the first evaluation, the TOPAS summaries have been evaluated on their own with respect to their content and linguistic quality. In the second evaluation, the quality of the TOPAS summaries has been contasted to the quality of the summaries produced by the Centroid-based and LexRank.

	Statement	Average rating
1	Given the length of the summary (about 150 words), the most important parts of the patent are in the summary (more is better)	3.38
2	There are segments missing that are more relevant than the ones selected (less is better)	2.46
3	There are irrelevant segments in the summary(less is better)	2.58
4	The summary is too short (1 means that should be shorter, 3 means is ok, 5 that should be longer) (less is better)	3.17
5	There are too similar segments in the summary (less is better)	1.83

Table 4. Evaluation of the quality of the content of the TOPAS summaries

The first evaluation was carried out by three employees from three companies specialized in intellectual property management and patent technologies. Each of the evaluators assessed TOPAS summaries of a distinct subset of eight patents, selected from the set of the summaries of 26 patents.[13] The evaluators assigned to each statement in a precompiled questionnaire a grade on the Likert scale from 1 to 5: '1' = "Strongly disagree"; '2' = "Disagree"; '3' = "Neither agree nor disagree"; '4' = "Agree"; '5' = "Strongly agree". Table 4 displays the questionnaire and summarizes the outcome of the evaluation of the content of the summaries. The table shows that, in general, the users were inclined to consider that in TOPAS summaries, the most important parts of the patent are reflected in the summary and that no segments that would be more important than those included in the summary are missing. The users also tend to disagree that the summary contains irrelevant material, or, in other words, they tend to agree that the summary contains only relevant material. With the rating of 3.17, there is a very slight tendency towards the opinion that the summaries should be longer ('3' would mean that the length is appropriate). Finally, users definitely disagree that the summaries include repeated information. Overall, the judgements of the users with respect to the content of the summaries is thus rather positive.

In Table 5, the outcome of the evaluation of the linguistic quality of the summaries (and thus mainly of the summary generation component) is displayed. The general impression of the users can be considered positive: all aspects of the linguistic quality (comprehensibility, grammaticality, readability, ordering, and faithfulness) have been rated in the average between 3.67 (which can be interpreted as "tend to be satisfied" and 4.04 ("Satisfied").

For the second evaluation, summaries of each of the 26 patents from the first evaluation were generated using LexRank, the Centroid-based summarizer and our technique and presented in random order, together with the corresponding patents, to three evaluators (different from those who participated in the first evaluation). Two of the evaluators evaluated the same subset of patents. The evaluators were asked to rate the same statements as already

[13]The evaluation of patent summaries with respect to their adequacy and quality is a time consuming task since it also requires a thorough study of the corresponding patents. This has consequences for the size of the test dataset.

	Statement in all of them more is better	Average rating
1	It can be well understood what the summary is about	3.83
2	The text fluency and grammatical correctness are good	3.63
3	The readability of the summary is good	3.67
4	The ordering of the sentences is correct	4.04
5	The generation does not change the sense of the patent	3.67

Table 5. Comparative evaluation of the language quality of the TOPAS summaries

used in the first evaluation. In addition, they had to answer the question "Which of the three summaries do you think is better?" Table 6 shows the outcome of the content assessment of the summaries.[14] It can be observed that TOPAS summaries are consistently rated better than those produced by LexRank and the Centroid-based summarizer, respectively. It should be also noted that the ratings of TOPAS summaries are similar to the ratings in the first evaluation, when only TOPAS summaries were shown to the evaluators—except for the statement on the missing segments, where the rating is lower. But in this case too TOPAS outperforms the other techniques.

	Statement	Average Ratings		
		LexRank	Centroid	TOPAS
1	Given the length of the summary, the most important parts of the patent are in the summary (more is better)	2.7	2.6	**3.4**
2	There are segments missing that are more relevant than the ones selected (less is better)	3.9	3.8	**3.3**
3	There are irrelevant segments in the summary (less is better)	3.8	4.0	**2.3**
4	There are too similar segments in the summary (less is better)	2.5	2.5	**1.7**

Table 6. Evaluation of the quality of the content of the summaries

In Table 7, the outcome of the contrastive evaluation of the linguistic quality of the summaries is displayed. TOPAS summaries again clearly outperform the summaries of the other two state-of-the-art techniques. Recall that the TOPAS technique extracts segments (rather than full sentences) and then fuses them during the generation stage, while the other two techniques extract sentences. The fact that the TOPAS summarization outperforms sentence-based extractive summarization with respect to readability and text fluency indicates that segment-based extraction and posterior linguistic fusion of the selected segments into grammatical sentences is the right approach for the patent genre.

	Statement in all of them more is better	Average Ratings LexRank	Centroid	TOPAS
1	It can be well understood what the summary is about	2.5	2.6	**3.6**
2	The text fluency and grammatical correctness are good	2.7	2.7	**3.6**
3	The readability of the summary is good	2.6	2.3	**3.8**
4	The ordering of the sentences is correct	3.0	2.6	**3.6**
5	The generation does not change the sense of the patent	3.4	3.3	**3.67**

Table 7. Evaluation of the language quality of the summaries

[14]The average deviation in the ratings of the two evaluators who assessed the same subset of patents is 0.52 ponts (on the 1 – 5 scale).

7. Related work

An increasing number of NLP-related works deals with the patent genre; see, for instance, [41] or, more recently, [42] for general reviews. The majority of the works targets patent analysis using either text mining or visualization techniques. Text mining is based on keywords or on semantic relations such as synonymy, hyponymy or hypernymy, which can be obtained from a domain crafted ontology or using distributional semantics based techniques such as Latent Semantic Analysis (LSA)[43] [44, 45] or SAO (Subject-Action-Object) triples [43, 46, 47] to determine relevant elements and their distribution across a collection of patents. Visualization techniques rely on metadata, citation networks and information extracted using, e.g., text mining [44, 48].

Both text mining and visualization thus help improve and accelerate patent search in order to identify competitors, trends and new research directions. They narrow down the subset of patents to be examined (either automatically or by offering the interactive techniques for this purpose) and provide an overview of the content of patents. However, although bags of words, keywords or triples of SAOs serve to know what a patent is about, they are not sufficient to assess the definite relevance of a patent to the user. The user still has to read all preselected patents—a task which is considerably alleviated when summaries of these patents are available. As already pointed out in the Introduction, only a small number of existing works addresses patent summarization. The works that focus on patent summarization tend to either take only the Claims into account or draw upon summarization features and units known from general discourse summarization.[15] Shinmori et al. [3] and Bouayad-Agha et al. [4] are examples of proposals that focus on the summarization of Claims. Both derive in a first stage a Discourse Tree in the sense of the Rhetorical Structure Theory [6] of the statements in the Claims in order to prune it then in a second stage. The pruning procedure is guided by the nature of the discourse relations in the tree, the tree depth and the desired length of the summary. Bouayad-Agha et al. furthermore suggest, as an alternative, a further summarization strategy that is based on the syntactic dependency structures of the claim sentences. Again, the dependency structures are pruned to obtain a summary, guided by such criteria as the nature of the relations, depth and summary length. The pruned discourse respectively dependency trees are converted into a coherent summary using full-fledged text generation techniques [49]. In this respect, our strategy is similar to that of Bouayad-Agha et al.since we also use text generation techniques to obtain a coherent and cohesive summary.

Tseng et al. [50] select for inclusion into the summary sentences from each section of the patent, based on such features as the number of keywords, title words, and clue words they contain and the position they are at.[19] Trappey & Trappey [31] include whole paragraphs into a predefined summarization template. The relevance of the paragraphs is assessed in terms of their *information density* that is computed using the relevance of key phrases, title phrases, topic sentences, etc. that occur in them. The central criterion for the weight (and thus relevance) of a given keyword phrase is its *tf*idf* score. Trappey et al. [5] improve on [31] in that they also use manually compiled domain ontologies to compute the relevance of key terms / key phrases. We also draw upon, for instance, term (in our case, mention) distribution and position features. However, as became clear from Section 4, our use of these features is oriented towards segment selection. Equally, our use of lexical chains is different from past uses of lexical chains in general discourse summarization discussed in [10, 51]. Thus, while in general discourse lexical chains are used for identification of sentences to be included in the summary in that sentences in which mentions of the longest lexical chains occur are selected for inclusion, in our proposal, the length of coreference lexical chains is just one feature that contributes to the computation of the relevance of a segment. For other types of lexical chains, we do not consider the length at all (but rather the type of the relation between the mentions).

Since patents are technical documents, related work relevant to TOPAS also includes summarization of technical or scientific articles. Traditional approaches in this area have concentrated on developing patterns in close or open scientific domains [52] or on proposing discourse classification algorithms [53] to select key sentences. In recent years, a new generation of scientific summarization techniques has emerged that takes the significant growth of online scientific publishing into account. These techniques take advantage of the citations that a research paper has in order

[15] This seems to be also the case in commercial products that deal first of all with patent search, but offer as an additional function automatic patent summarization, as, e.g., PatBase[16], Questel Orbit[17], and Intellexer[18], etc. The lack of publicly available details prevents us from their description here.

[19] Note that in [50] the authors use the term "segment" for what we call "section" of the patent, i.e., Claims, Description, Abstract, etc.; cf. Subsection 2.1.1.

to extract and summarize its main contributions [54]. Thus, Qazvinian and Radev [55] and Elkiss et al. [56] analyze the information provided by citation contexts of all citations to evaluate their utility for the generation of a summary. It is shown that citation contexts provide information which can be seen as complementary to what is reported in author abstracts. Methods to improve the coherence of the generated citation summaries use sentence classification to decide what type of information a sentence is convening [57].

8. Conclusions and future work

In this paper, we presented an advanced technique for summarization of patents. The technique draws upon a *segment* as basic working unit. A segment captures an individual statement on the invention or on one of its aspects and is thus detached from the surface notion of a sentence. This allows for a more fine-grained selection of content to be included into the summary and surpasses the problem of the inclusion of long claim sentences, as seen in some of the previous works that apply extractive summarization to the patent genre. Furthermore, this allows for the coverage of all sections of the patent.

The proposed metric for the selection of segments for their inclusion into the summary uses three types of features: (i) mention and lexical chain-oriented features, (ii) segment-oriented features and (iii) a number of classical similarity and position features. An average F1-score of 0.47 of the generated summaries when compared to manually written gold standard summaries shows that our summarization technique performs well and considerably better than the baseline and also above two other state-of-the-art techniques. An additional qualitative evaluation furthermore shows the high satisfaction of users with our summaries. The use of a graph transduction-based text generation environment to aggregate the syntactic structures of the segments selected for the summary and to generate a coherent summary out of the aggregated structures further contributes to the quality of our summaries. Extractive summarizers that draw upon sentences instead of segments and that do not use text generation techniques are not able to achieve the same grade of user satisfaction.

However, the tests also show that the individual components of the TOPAS summarization module can be further improved. This concerns both some preprocessing tasks such as dependency parsing, which provides input to segmentation, and core tasks such as segmentation and lexical chain identification. Furthermore, recent advances in general discourse summarization underline the importance of the discourse structure for summarization [58]. As already pointed out above, some previous works on patent claim summarization already use discourse structure [3, 4], but their focus on the Claims section let them miss summary-relevant information from the other sections of the patent. The extension of the summarization features to aligned discourse structures of all sections of the patent seems promising and will also be considered in our future work.

Acknowledgements

The work reported on in this paper has been carried out in the framework of the TOPAS (*Tool Platform for Intelligent Patent Analysis and Summarization*) project, which has been partially funded by the European Commission within its FP7 Programme under the contract number FP7-SME-286639. The TOPAS Consortium was composed of Brügmann Software, Papenburg; IALE, Barcelona; IntelliSemantic s.a., Torino; Pompeu Fabra University, Barcelona; and the University of Stuttgart.

References

[1] M. Lupu, K. Mayer, J. Tait, A. Trippe, Current challenges in patent information retrieval, Springer, Heidelberg, Berlin, New York, 2011.

[2] L. Wanner, R. Baeza-Yates, S. Brügmann, J. Codina, B. Diallo, E. Escorsa, M. Giereth, Y. Kompatsiaris, S. Papadopoulos, E. Pianta, G. Piella, I. Puhlmann, G. Rao, M. Rotard, P. Schoester, L. Serafini, V. Zervaki, Towards content-oriented patent document processing, World Patent Information Journal 30 (1) (2008) 21–33.

[3] A. Shinmori, M. Okumura, Y. Marukawa, M. Iwayama, Patent processing for readability. structure analysis and term explanation, in: Proceedings of the Workshop on Patent Corpus Processing held at the ACL Meeting, ACL, Morristown, PA, USA, 2003, pp. 56–65.

[4] N. Bouayad-Agha, G. Casamayor, G. Ferraro, S. Mille, V. Vidal, L.Wanner, Improving the comprehension of legal documentation: The case of patent claims, in: Proceedings of the International Conference on Artificial Intelligence in Law., Barcelona, 2009.

[5] A. Trappey, C. Trappey, C.-Y. Wu, Automatic patent document summarization for collaborative knowledge systems and services, Journal of Systems Science and Systems Engineering 18 (1) (2009) 71–94.

[6] W. C. Mann, S. Thompson, Rhetorical structure theory: Toward a functional theory of text organization, Text 8 (3) (1988) 243–281.
[7] H. Saggion, T. Poibeau, Automatic text summarization: Past, present andfuture, in: T. Poibeau, H. Saggion, J. Piskorski, R. Yangarber (Eds.), Multi-source, Multilingual Information Extraction and Summarization, Springer Verlag, Berlin, 2013.
[8] C. Aone, M. E. Okurowski, J. Gorlinsky, B. Larsen, A trainable summarizer with knowledge acquired from robust nlp techniques, in: I. Mani, M. T. Maybury (Eds.), Advances in Automatic Text Summarization, MIT Press, Cambridge, MA, 1999, pp. 71–80.
[9] Y. Seki, Sentence extraction by tf/idf and position weighting from newspaper articles, in: Proceedings of the Third NTCIR Workshop, 2003.
[10] R. Barzilay, M. Elhadad, Text summarizations with lexical chains, in: I. Mani, M. Maybury (Eds.), Advances in Automatic Text Summarization, MIT Press, Cambridge, MA, 1999.
[11] S. Azzam, K. Humphreys, R. Gaizauskas, Using coreference chains for text summarization, in: Proceedings of the ACL'99 Workshop on Coreference and its Applications, Baltimore, 1999.
[12] C.-Y. Lin, E. Hovy, Identifying topics by position, in: Proceedings of the Fifth conference on Applied Natural Language Processing, ACL, Morristown, PA, USA, 1997, pp. 283–290.
[13] A. Khan, N. Salim, A review on abstractive summarization methods, Journal of Theoretical and Applied Information Technology 59 (1) (2014) 64–72.
[14] D. Pressman, Patent it yourself, 12th Edition, Nolo, Berkeley, CA, USA, 2006.
[15] H. Grice, Logic and Conversation, in: P. Cole, J. Morgan (Eds.), Syntax and Semantics 3: Speech Acts, Academic Press, New York, 1975, pp. 41–58.
[16] H. Cunningham, Text processing with gate (version 6) isbn 0956599311, Tech. rep., University of Sheffield, Department of Computer Science (2011).
[17] B. Bohnet, Top accuracy and fast dependency parsing is not a contradiction, in: Proceedings of the International Conference on Computational Linguistics (COLING), Beijing, 2010, pp. 89–97.
[18] A. Burga, J. Codina, G. Ferraro, H. Saggion, L. Wanner, The challenge of syntactic dependency parsing adaptation for the patent domain, in: Proceedings of the ESSLIWorkshop on Extrinsic Parse Improvement (EPI), Düsseldorf, Germany, 2013.
[19] A. Judea, H. Schütze, Unsupervised training set generation for automatic acquisition of technical terminology in patents, in: Proceedings of the International Conference on Computational Linguistics (COLING), Dublin, 2014.
[20] H. Lee, A. Chang, Y. Peirsman, N. Chambers, M. Surdeanu, D. Jurafsky, Deterministic coreference resolution based on entity-centric, precision-ranked rules, Compututational Linguistics 39 (4) (2013) 885–916.
[21] M. . R. H. Halliday, Cohesion in English, Longman, London, 1976.
[22] J. Morris, G. Hirst, Lexical cohesion computed by thesaural relations as an indicator of the structure of text, Computational Linguistics 17 (1) (1991) 21–48.
[23] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, C. Manning, A multi-pass sieve for coreference resolution, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, Morristown, PA, USA, 2010, pp. 492–501.
[24] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, D. Jurafsky, Stanford's Multi-pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task, in: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, 2011, pp. 28–34.
[25] N. Bouayad-Agha, A. Burga, G. Casamayor, J. Codina, R. Nazar, L. Wanner, An exercise in reuse of resources: Adapting general discourse coreference resolution for detecting lexical chains in patent documentation, in: Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 1999.
[26] L. Taylor, C. Grover, T. Briscoe, The syntactic regularity of english noun phrases, in: Proceedings of the 4th Biannual Meeting of the European Chapter of the Association for Computational Linguistics (EACL), 1989, pp. 256–263.
[27] D. Bourigault, Surface grammatical analysis for the extraction of terminological noun phrases, in: Proceedings of the 14th International Conference on Computational Linguistics (COLING), 1992, pp. 977–981.
[28] C. Cardie, D. Pierce, Error-driven pruning of treebank grammars for base noun phrase identification, in: Proceedings of the 17th International Conference on Computational Linguistics (COLING) and the 36th Annual Meeting of the Association for Computational Linguistics (ACL), 1998, pp. 218–224.
[29] R. Subhashini, V. Kumar, Shallow NLP Techniques for Noun Phrase Extraction, in: Proceedings of Trendz in Information Sciences & Computing, 2010.
[30] USPTO, Manual of Patent Examining Procedure, Revision 07.2015, 2015.
[31] A. J. C. Trappey, C. V. Trappey, An R&D knowledge management method for patent document summarization, Industrial Management and Data Systems 108 (2) (2008) 245–257.
[32] H. P. Luhn, The automatic creation of literature abstracts, IBM J. Res. Dev. 2 (2) (1958) 159–165.
[33] A. Nenkova, L. Vanderwende, The impact of frequency on summarization, Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101.
[34] J. Kupiec, J. Pedersen, F. Chen, A trainable document summarizer, in: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '95, ACM, New York, NY, USA, 1995, pp. 68–73.
[35] H. Saggion, SUMMA. A Robust and Adaptable Summarization Tool, TAL 49 (2) (2008) 103–125.
[36] I. H. Witten, E. Frank, M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.
[37] B. Bohnet, L. Wanner, Open source graph transducer interpreter and grammar development environment, in: Proceedings of theInternational Conference on Linguistic Resources and Evaluation (LREC), Valletta, Malta, 2010.
[38] H. Saggion, R. Gaizauskas, Multi-document summarization by cluster/profile relevance and redundancy removal, Proceedings of the Document Understanding Conference (2004) 6–7.
[39] G. Erkan, D. R. Radev, Lexrank: Graph-based lexical centrality as salience in text summarization, J. Artif. Int. Res. 22 (1) (2004) 457–479.
[40] H. Saggion, Creating summarization systems with SUMMA, in: Proceedings of the Ninth International Conference on Language Resources

and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014., 2014, pp. 4157–4163.
[41] D. Bonino, A. Ciaramella, F. Corno, Review of the state of the art and forthcoming evolutions in intelligent patent informatics, World Patent Information 32 (1) (2014) 30–38.
[42] A. Abbas, L. Zhang, S. Khan, A literature review on the state-of-the-art in patent analysis, World Patent Information 37 (2) (2014) 3–13.
[43] H. Park, J. Yoon, K. Kim, Identifying patent infringement using SAO based semantic technological similarities, Scientometrics 90 (2) (2011) 515–529.
URL http://www.akademiai.com/doi/abs/10.1007/s11192-011-0522-7
[44] J. Tang, B. Wang, Y. Yang, P. Hu, Y. Zhao, X. Yan, B. Gao, M. Huang, P. Xu, W. Li, et al., Patentminer: topic-driven patent analysis and mining, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2012, pp. 1366–1374.
[45] B. Van Looy, B. Baesens, T. Magerman, K. Debackere, Assessment of latent semantic analysis (lsa) text mining algorithms for large scale mapping of patent and scientific publication documents, Available at SSRN 2096159.
[46] S. Choi, H. Kim, J. Yoon, K. Kim, J. Y. Lee, An SAO-based text-mining approach for technology roadmapping using patent information, R&D Management 43 (1) (2013) 52–74.
URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9310.2012.00702.x/pdf
[47] C. Lee, B. Song, Y. Park, How to assess patent infringement risks: a semantic patent claim analysis using dependency relationships, Technology Analysis & Strategic Management 25 (1) (2013) 23–38. doi:10.1080/09537325.2012.748893.
URL http://dx.doi.org/10.1080/09537325.2012.748893
[48] P. Érdi, K. Makovi, Z. Somogyvári, K. Strandburg, J. Tobochnik, P. Volf, L. Zalányi, Prediction of emerging technologies based on analysis of the us patent citation network, Scientometrics 95 (1) (2013) 225–242.
[49] S. Mille, L. Wanner, Multilingual summarization in practice: The case of patent claims, in: Proceedings of the European Machine Translation Summit (EAMT 2008), 2008.
[50] Y.-H. Tseng, C.-J. Lin, Y.-I. Lin, Text mining techniques for patent analysis, Information Processing and Management 43 (5) (2007) 1216–1247.
[51] H. G. Silber, K. F. McCoy, Efficiently computed lexical chains as an intermediate representation for automatic text summarization, Compututational Linguistics 28 (4) (2002) 487–496.
[52] H. Saggion, G. Lapalme, Generating indicative-informative summaries with sumum, Comput. Linguist. 28 (4) (2002) 497–526.
[53] S. Teufel, M. Moens, Summarizing scientific articles: Experiments with relevance and rhetorical status, Comput. Linguist. 28 (4) (2002) 409–445.
[54] V. Qazvinian, D. R. Radev, Scientific paper summarization using citation summary networks, in: Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1, COLING '08, Association for Computational Linguistics, Stroudsburg, PA, USA, 2008, pp. 689–696.
[55] V. Qazvinian, D. R. Radev, Identifying non-explicit citing sentences for citation-based summarization, in: ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, 2010, pp. 555–564.
[56] A. Elkiss, S. Shen, A. Fader, G. Erkan, D. States, D. Radev, Blind men and elephants: What do citation summaries tell us about a research article?, J. Am. Soc. Inf. Sci. Technol. 59 (1) (2008) 51–62.
[57] A. Abu-Jbara, D. Radev, Coherent citation-based summarization of scientific papers, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 500–509.
[58] T. Hirao, Y. Yoshida, M. Nishino, N. Yasuda, M. Nagata, Single-document summarization as a tree knapsack problem, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, 2013, pp. 1515–1520.