Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Using genre-specific features for patent summaries

Joan Codina-Filbà [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, Spain*
[b] *Institute for Natural Language Processing, University of Stuttgart, Germany*
[c] *Catalan Institute for Research and Advanced Studies (ICREA), Spain*

ARTICLE INFO

Article history:
Received 7 August 2014
Revised 19 July 2016
Accepted 20 July 2016
Available online xxx

Keywords:
Summarization
Patents
Lexical chains
Segmentation
Segment-based summarization
Sentence aggregation

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 vocabulary, 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 sub-sentential 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.

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 patent search is

[*] Corresponding author.
E-mail addresses: joan.codina@upf.edu (J. Codina-Filbà), nadjet.bouayad@upf.edu (N. Bouayad-Agha), alicia.burga@upf.edu (A. Burga), gerard.casamayor@upf.edu (G. Casamayor), simon.mille@upf.edu (S. Mille), horacio.saggion@upf.edu (H. Saggion), leo.wanner@upf.edu (L. Wanner).

http://dx.doi.org/10.1016/j.ipm.2016.07.002
0306-4573/© 2016 Elsevier Ltd. All rights reserved.

recall-driven (Lupu, Mayer, Tait, & Trippe, 2011). 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. (2008), 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., Bouayad-Agha et al. (2009); Shinmori, Okumura, Marukawa, and Iwayama (2003); Trappey, Trappey, and Wu (2009). Most often, these works either replicate the summarization metrics known from general discourse summarization (Trappey et al., 2009) 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 (Bouayad-Agha et al., 2009; Shinmori et al., 2003). Thus, Trappey et al. (2009) 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. (2003) prune the discourse structure of each claim represented in terms of the Rhetorical Structure Theory (Mann & Thompson, 1988) 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. (2009) 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.[2] 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 (Saggion & Poibeau, 2013). Extractive summarization is surface-oriented in that it applies relevance metrics usually based on distribution heuristics (e.g., *tf***idf* of individual tokens (Aone, Okurowski, Gorlinsky, & Larsen, 1999; Seki, 2003), lexical chains (Azzam, Humphreys, & Gaizauskas, 1999; Barzilay & Elhadad, 1999), position of a sentence in the text (Lin & Hovy, 1997), 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 (Khan & Salim, 2014). It can thus be considered to require the availability of the semantic analysis of the content of the text in question.

[1] http://thomsonreuters.com/derwent-world-patents-index/.
[2] A demo version of our summarizer is available at http://topas-engine.upf.edu/.

Please cite this article as: J. Codina-Filbà et al., Using genre-specific features for patent summaries, Information Processing and Management (2016), http://dx.doi.org/10.1016/j.ipm.2016.07.002