World Patent Information 40 (2015) 30–42

Contents lists available at ScienceDirect

World Patent Information

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

Towards content-oriented patent document processing: Intelligent patent analysis and summarization

Sören Brügmann [a], Nadjet Bouayad-Agha [b], Alicia Burga [b], Serguei Carrascosa [c], Alberto Ciaramella [d], Marco Ciaramella [d], Joan Codina-Filba [b], Enric Escorsa [c], Alex Judea [e], Simon Mille [b], Andreas Müller [e], Horacio Saggion [b], Patrick Ziering [e], Hinrich Schütze [f], Leo Wanner [g, b, *]

[a] Brügmann Software GmbH, Germany
[b] Pompeu Fabra University, Spain
[c] IALE, Spain
[d] IntelliSemantic S.r.l, Italy
[e] University of Stuttgart, Germany
[f] University of Munich, Germany
[g] Catalan Institute for Research and Advanced Studies (ICREA), Spain

ARTICLE INFO

Article history:
Available online 15 December 2014

Keywords:
Entity recognition
Segmentation
Lexical chain identification
Claim description alignment
Summarization
TOPAS
Patent analysis
Document processing

ABSTRACT

In this article, we present an operational prototype of a workbench for intelligent patent document analysis and summarization that has been developed in the context of the R&D project TOPAS, partially funded by the European Commission. The workbench uses the GATE environment as infrastructure for document representation and algorithm integration. It contains, apart from several preprocessing tools, five modules for the individual aspects of patent analysis (entity recognition, lexical chain identification, invention composition derivation, segmentation, and claim – description alignment) and a module for patent summarization. The workbench, which has been tested in different application settings, can be used as a standalone engine or as component within a more global patent processing line. Most of its modules can be also used separately.

1. Introduction

With the increasingly prominent role of patents for the economy and progress in advanced information and communication technologies, patent processing aids are of increasing demand. The most standard among them continue to be, as already in the earlier days of computer-supported patent management, patent search engines and machine translation programs: search engines provide patent material to work with and translation programs ensure access to patent markets that would otherwise be inaccessible to many patent users. Aids that support the user, for instance, in the tasks of patent content analysis or contrastive evaluation, are much more seldom. Patent specialists (among them, examiners, experts in patent departments, and patent attorneys) are still left on their own with the task to inspect, assess and compare the content of this material – although it is generally acknowledged that parts of patent material documentation (such as the claims) are difficult to read and comprehend due to their complex (linguistic) style and writing conventions [27]. Available reading aids as offered, e.g., by Questel Orbit[1] and Minesoft,[2] provide access to multilingual dictionaries/thesauri that contain the translation, definition and synonyms of the terms. Advanced options relate terms to topics these terms refer to (the "topics" can be mere hyperonyms of the terms in question or more abstract), highlight the occurrences of the terms in the document, indicate the frequency of the occurrence of the terms, and so on. Especially the recognition and highlighting of domain terms (including multiword terms) the user is interested in is helpful during the examination. However, this is not sufficient.

* Corresponding author. Catalan Institute for Research and Advanced Studies (ICREA), Spain.
E-mail address: leo.wanner@upf.edu (L. Wanner).

[1] https://www.piug.org/Resources/Documents/PIUG_2010_QUESTEL_workshop.pdf.
[2] http://www.patbase.com/Manual.pdf.

http://dx.doi.org/10.1016/j.wpi.2014.10.003
0172-2190/

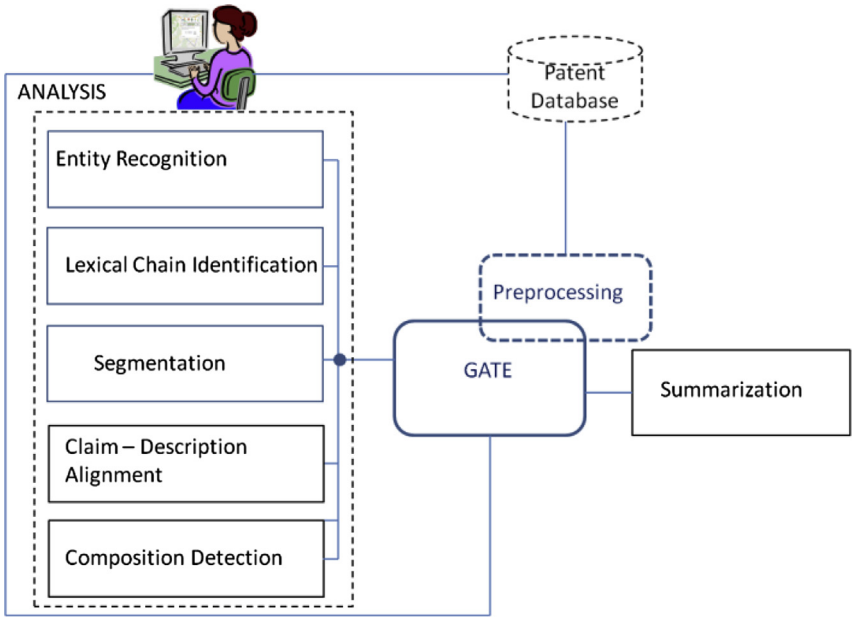

Fig. 1. TOPAS Workbench architecture.

Thus, it is important to detect and highlight not only the occurrence of a given term, but also the occurrence of all terms that are related to it across the patent, it is convenient to see which part of the description elaborates on which claim, etc.

In what follows, we present the TOPAS[3] workbench, which has been designed based on the findings of the previous PATExpert project by Wanner et al. [30] to support interactive analysis and summarization of patent material. In the next section, an overview of the workbench is given. Section 3 presents the individual modules of patent analysis integrated in the workbench, while Section 4 focuses on the presentation of patent summarization in TOPAS. Section 5 summarizes the functionality of TOPAS, illustrates its application and outlines its possible extension.

2. The TOPAS workbench

The TOPAS Workbench prototype integrates the technologies for entity recognition, lexical chain identification, patent segmentation (or zoning), claim – description alignment, invention composition derivation and patent summarization. These tasks have been identified as central in collaboration with patent processing practitioners. Each of the TOPAS technologies is realized as a separate module. Most of them can be used independently from each other. Only alignment and composition detection presuppose a prior segmentation.

To facilitate linguistically preprocessed input for the individual modules, in addition, a linguistic preprocessing module has been incorporated. The open source software GATE by Cunningham et al., [7] is used as the infrastructure for document representation and algorithm integration. See Fig. 1 for the sketch of the TOPAS architecture. TOPAS processes patents in all three European patent languages: English, French and German, although the performance of most modules on English patents is higher than on French and German patents.

Any patent chosen by the user for closer inspection is stored, after being preprocessed, in GATE-format. The modules 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 module's proprietary format to be processed by the module, and its output is transformed from the proprietary format again into GATE-format. The use of GATE as "blackboard" ensures thus total flexibility with respect to the design and execution of the working system in that new modules can be plugged-in and modules can be arranged in a unique pipeline on a single computer or in several pipelines, each running on a different server.

The preprocessing module incorporates a series of linguistic processing tools: part-of-speech tagger, lemmatizer, and dependency parser from B. Bohnet's [2] MATE parsing environment,[4] GATE's tokenizer, the Sentence Splitter from OpenNLP, and a patent genre chunker developed in the scope of TOPAS. 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. [4] for such an adaptation of the parser.

The other modules of the Workbench are described in the following sections.

3. Patent analysis technologies

As already mentioned above and as illustrated in Fig. 1, five different analysis techniques are offered in the TOPAS Workbench. In what follows, these techniques are presented in more detail.

3.1. Entity recognition

Entity recognition is a subtask of term recognition; it captures terms that denote entities of the invention. We consider a sequence of words $w_1,w_2,\ldots,w_k$ an entity term of domain D if

[3] TOPAS stands for "Tool Platform for Intelligent Patent Analysis and Summarization".

[4] To the best of our knowledge, no off-the-shelf NLP techniques are available for the patent domain.