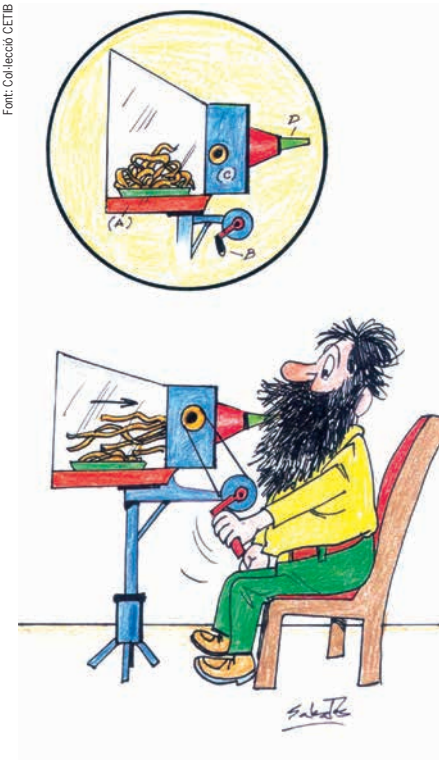
Terminologia per als invents i les patents

dossier

Automatic processing of patents: an exercise in computational terminology

Leo Wanner
ICREA and Natural Language Processing Group (TALN)
Pompeu Fabra University



Font: Col·lecció CETIB

With the increasing importance of Intellectual Property Protection (IPR), patents have become one of the most prominent genres of specialized discourse:[1] they are the only legal means available to document in detail an invention and to certify its intellectual author- and ownership. The downside of this hype is that accurate and timely examination of patent applications and patent infringement monitoring grew to an extraordinary challenge. To address this challenge, patent search engines have traditionally been used. However, in order not to miss any relevant material, these search engines tend to be recall-oriented, which implies that patent examiners and patent monitoring specialists have to go over large quantities of patent material on a daily basis. The only possible instrument to alleviate their workload is the use of natural language processing (NLP) applications, adapted to the patent genre. The central applications involve, at least: (*a*) automatic patent summarization; (*b*) (semantic) patent analysis; (*c*) term chain detection; and (*d*) term co-reference resolution. Thus, patent summaries and semantic patent analysis outcomes, which can be casted into compositional and functional diagrams, facilitate a quick overview of the content of a patent without the need to read the whole document, and term chain detection and co-reference resolution highlight the most

Terminàlia 16 (2017): 54-56 · DOI: 10.2436/20.2503.01.114
ISSN: 2013-6692 (impresa); 2013-6706 (electrònica) · *http://terminalia.iec.cat*

important elements in a patent, which, again, speeds up the inspection procedure. In our work, we address all four of these tasks.

Most of the research on automatic patent summarization focused so far on the summarization of patent claims (cf., e.g., Shinmori et al., 2003; Bouayad-Agha et al., 2009; Trappey et al., 2009), which arguably constitute the central section of a patent (application). However, the other sections (such as the description) elaborate on, e.g., the preferred embodiment and possible applications of the invention, such that it is also important to summarize them together with the claims in order to obtain a coherent summary of the whole document. We present a proposal in this respect in (Codina et al., 2017).

The majority of the works that address the problem of patent analysis do it at the lexical level (Cascini et al., 2009; Choi et al., 2012; Xu et al., 2015). However, it is important to note that the level of abstraction of the vocabulary varies significantly between the claims and the description. Thus, in a claim we might read *apparatus* and *recording device*, while in the description, the same invention may be referred to as *tape recorder*. In order to obtain an accurate analysis of the composition and the mode of operation of the patented invention, we need to be able to identify the components (or concepts) referred to by the individual terms. In other words, we need to be able to derive from a patent the conceptual representation of the invention. In order to ensure that this representation is flexible enough and formally verifiable, we draw upon Web Ontology Language (OWL) representations; see (Dasiopoulou et al., 2015).

As a rule, the components of the invention are mentioned in the course of the patent several times. The mention frequency and the positions of the mentions indicate the relevance of the corresponding component. Therefore, it is important to capture the "chains" of the terms that refer to the same component. As already pointed out above, the same component is referred to by different terms (cf., *apparatus* and *tape recorder*, *device* and *projector*, *light source* and *diode*, etc.), which makes the task more complex. The results of our work in this area are documented, e.g., in (Bouayad-Agha et al., 2014; Brügmann et al., 2015).

Directly related to the problems of patent analysis and term chain identification is term co-reference resolution in patents. In general discourse, single antecedent co-reference (SAC) prevails, as, e.g., in

> (1) [*An excavator*]$_i$ *in* [*which*]$_i$ [*a lower traveling body*]$_j$ *is equipped with an upper rotating body* [*thereon*]$_j$, *and* ...

In patents, multiple antecedent co-reference (MAC) is also very common:

> (2) *The electric circuit wherein each of the DC-to-AC converters comprises* [*a first switch*]$_i$ [...] *and* [*a second switch*]$_j$ [...]. *The electric circuit wherein* [*the first and second switches*]$_{i+j}$...

Both SAC and MAC need to be resolved in order to obtain an objective picture of the significance of a component. To address this problem, we adapt the sieve strategy already suggested by Raghunathan et al. (2010); see (Bouayad-Agha et al., 2014; Burga et al., 2016).

It is evident that in all four of the above tasks, terminology is central: Thus, for patent summarization, the relevance metrics that identify the relevant segments of the patent document that are to be included into the summary are based on proper term identification and linking; semantic patent analysis, term chain identification and term co-reference resolution draw on terms by their very nature. Again, we have to keep in mind that terms in patents possess some idiosyncrasies we need to deal with. In particular: (*a*) abstract terms (such as *apparatus*, *device*, *means*, etc.) act as place holders of more specific signifiers of the same concept; (*b*) terms in patents are very often multiword terms (cf., e.g., *renewable energy*, *wind mill*, *rechargeable electronic device*, etc.); (*c*) as a rule, we need to capture not only concrete terms (i.e., terms that denote concrete objects), but also predicative terms (San Martin and L'Homme, 2014) that denote actions and procedures.

References

Bouayad-Agha, N.; Burga, A.; Casamayor, G.; Codina, J.; Nazar, R.; Wanner, L. (2014). Bouayad-Agha, BN., A. Burga, G. Casamayor, J. Codina, R. Nazar, and L. Wanner. 2014. "An exercise in reuse of resources: adapting general discourse coreference resolution for detecting lexical chains in patent documentation". In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Bouayad-Agha, N.; Casamayor, G.; Ferraro, G.; Mille, S.; Vidal, V.; Wanner, L. (2009). "Improving the comprehension of legal documentation: the case of patent claims". In: *Proceedings of the International Conference on Artificial Intelligence in Law*.

Brügmann, S.; Bouayad-Agha, N.; Burga, A.; Carrascosa, S.; Ciaramella, A.; Ciaramella, M.; Codina-Filbà, J.; Escorsa, E.; Judea, A.; Mille, S.; Müller, A.; Saggion, H.; Ziering, P.; Schütze, H.; Wanner, L. (2015). "Towards content-oriented patent document processing: intelligent patent analysis and summarization". *World Patent Information Journal*, 40, pp. 30–42.

Burga, A.; Cajal, S.; Codina-Filbà, J.; Wanner, L. (2016). "Towards multiple antecedent coreference resolution in specialized discourse". In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Cascini, G.; Cugini, U.; Frillici, F. S.; Rotini, F. (2009). "Computer-aided conceptual design through TRIZ-based manipulation of topological optimizations". In: *Proceedings of the 19th CIRP Design Conference–Competitive Design*.

Choi, S.; Park, H.; Kang, D.; Lee, J. L.; Kim, K. (2012). "An SAO-based text mining approach to building a technology tree for technology planning". *Expert Systems with Applications*, 39 (13), pp.11443–11455.

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

Dasiopoulou, S.; Lohmann, S.; Codina, J.; Wanner, L. (2015). "Representing and visualizing text as ontologies: a case from the patent domain". In: *Proceedings of the Visualizations and User Interfaces for Ontologies and Linked Data (VOILA!) Workshop, co-located with ISWC 2015*, Bethlehem, PA.

Raghunathan, K.; Lee, H.; Rangarajan, S.; Chambers, N.; Surdeanu, M.; Jurafsky, D.; Manning, C. D. (2010). "A multi-pass sieve for coreference resolution". In: *Proceedings of the 2010 EMNLP*, pp. 492–501.

San Martin, A.; L'Homme, M. C. (2014). "Definition patterns for predicative terms in specialized dictionaries". In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Shinmori, A.; Okumura, M.; Marukawa, Y.; Iwayama, M. (2003). "Patent claim processing for readability: structure analysis and term explanation". In: *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing*, pp. 56–65.

Trappey, A. J. C.; Trappey, C. V.; Wu, C.-Y. (2009). "Automatic patent document summarization for collaborative knowledge systems and services". *Journal of Systems Science and Systems Engineering*, 18 (1), pp. 71-94.

Xu, H.; Gui, J.; Qu, P.; Zhu, X.; Wang, L. (2015). "Exploring similarity between academic paper and patent based on Latent Semantic Analysis and Vector Space Model". In: *Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*.

Note

1. Only in 2015, the European Patent Office received about 280,000 patent applications; cf. *https://www.epo.org/about-us/annual-reports-statistics/statistics.html*..

El refrany

Font: *Paremiologia catalana. www.dites.cat*