Generation of Patent Abstracts: A Challenge for Automatic Text Summarization

Leo Wanner, ICREA and DTIC, UPF

It is well known that patents drive the modern economies. But they do even more: patents also serve as a valuable and unique source of up-to-date scientific and technological information. It is assumed that only 10% to 15% of the content presented in patents are described in other publications as well. The worldwide stock of patents thus comprises about 85% to 90% of scientific knowledge. Given that central parts of patents are authored in an idiosyncratic and complex language which is difficult to read and comprehend, and since author-written patent abstracts have the goal to obfuscate the precise nature and the real scope of the inventions rather than to clarify them, an efficient access to this knowledge, for instance, via concise and transparent summaries, appears crucial. However, partially due to the aforementioned language idiosyncrasy, which implies extremely long sentences with complex repetitive linguistic constructions, common extraction-oriented automatic text summarization techniques cannot be expected to show an acceptable performance when applied to patents. Other, more content-oriented (or abstractive) summarization techniques are needed. In my talk, I will present the recent and ongoing research on patent summarization carried out by the Natural Language Processing Group of the Department of Information and Communication Technologies, UPF as member European consortia. I will first describe the techniques for the summarization of patent claims developed in the scope of the PATExpert project and outline then how these techniques are about to be improved in the TOPAS project by considering information from other sections of a patent, notably the description of the invention. In the last part of my presentation, I will summarize the remaining challenges and suggest some lines of future research which are crucial if we want automatic patent summarization to be a real alternative to (semi-)manual abstracting, which still dominates the patent domain.