Improving automatic patent classification system with language processing approach: Case of fuel cell vehicle

Samira Ranaei[*], Matti Karvonen[*], Arho Suominen[**] and Tuomo Kassi[*]

samira.ranaei@lut.fi ; Matti.Karvonen@lut.fi;Tuomo.kassi@lut.fi

[*]Industrial Engineering and Management, Lappeenranta University of technology, Skinnarilankatu 34,
Lappeenranta, 53850 (Finland)

[**]arho.suominen@vtt.fi
[**]VTT Technical Research Centre of Finland, Tekniikantie 2, Espoo, FI-02044 VTT (Finland)

Abstract

One of the main concerns in retrieving patent data of emerging technologies is how to be assured of the accuracy of retrieved patents. The relevance, reliability and accuracy of data is the main issue in creating information for effective decision making. In this paper an automated patent classification (APC) system is used that utilizes both structured and unstructured patent data. What distinguish APC from other tools is its capability of defining and classifying patents based on their content rather than bibliometric features by text mining techniques.

It has been suggested that the application of linguistic processing (such as N-gram) specifically the use of compound words, instead of single terms in APC, can help the constitution of richer training sets. Previous research results, by using experimental corpora regardless of context, show that addition of phrases (specifically bigrams) to unigrams can significantly improve the classification accuracy. We have investigated to what extent bigrams would increase APC accuracy level in context of fuel cell vehicle (FCV) technologies, and if higher levels of aggregation yield even better results

The findings show that, trigram in addition to bigrams and single words caused three percent improvement of accuracy rate. Although bigram shows very good rate of classification, trigrams worked better. However, the challenging task of forming training set can be considered as a limitation in this research. For future works, one promising approach for improving APC accuracy is to specialize them to a various patent fields, which probably leads to richer training sets.

Keywords: Automatic patent classification, linguistic processing, Textmining, trigram, fuel cell vehicle technology.