NLP-based Extraction of Knowledge from Patent Documents

Salvatore Sorce, Giovanni Garraffa, Nicole Dalia Cilia, Vincenzo Conti, Marco Siniscalchi
Università degli Studi di Enna - Kore, Italy
{salvatore.sorce, giovanni.garraffa, nicoledalia.cilia, vincenzo.conti, marco.siniscalchi}@unikore.it

Abstract

The Laboratory in Speech Technology and Machine Learning at the Università degli Studi di Enna - Kore (UKE), Italy addresses several research topics in those fields. In this document we present the activities based on NLP approaches to knowledge extraction from documents representing industrial patents, along with the results achieved so far and the future challenges we plan to deal with. The approach can be applied to a wide range of applications for different purposes.

1 Intelligent Patent Analysis

Analysis of patents can provide engineering design insight and identify potential infringement for promoting innovation. However, designers do not regularly engage with patents due to the intricate structure and legal terminologies used, especially in the early design stages. Methods for capturing patent knowledge from various patent sections and producing visualisations to facilitate understanding have been presented in the literature, including citation analysis and claims comparisons. In this context, we apply a standard Natural Language Processing (NLP)-based approach on different sections of the patent document (i.e. abstract, independent claim and all claim sections), to provide patent benchmarking for use in engineering design.

1.1 Achieved results

Figure 1 presents an overview of the NLP-based approach developed for performing analysis using different patent descriptive sections [Jiang *et al.*, 2021]. Word tokenisation in NLTK works on both single sentences and multiple sentences to split them into individual words, hence it is applied directly regardless of the section type. Part-of-Speech (POS) tagging is then applied to assign labels to each tokenised word. The next step is to lemmatise, i.e. consolidate noun phrases (e.g. POS tags start with N) only to eliminate duplicate expressions such as 'battery' and 'batteries'. This is accomplished by first converting NLTK tags to WordNet tags and then applying the WordNet Lemmatiser. The reason for lemmatising noun phrases only while keeping other phrases unchanged is to maintain the accuracy of parsing performed later. Before parsing, stopwords are removed from the text by referring to a

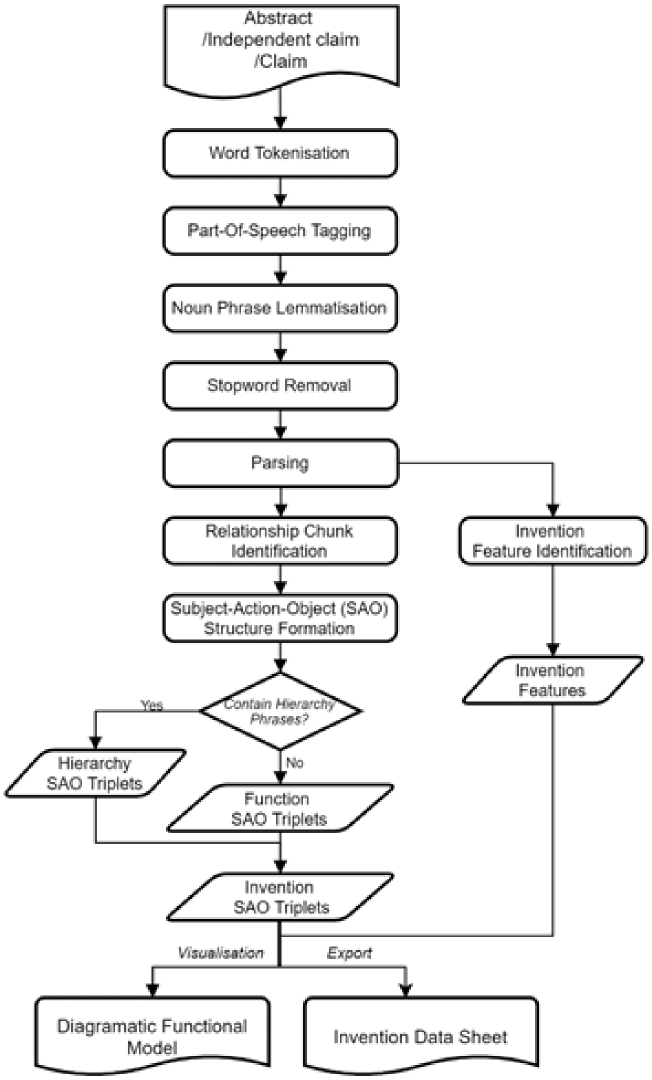

Figure 1: Methodology for patent functional modelling

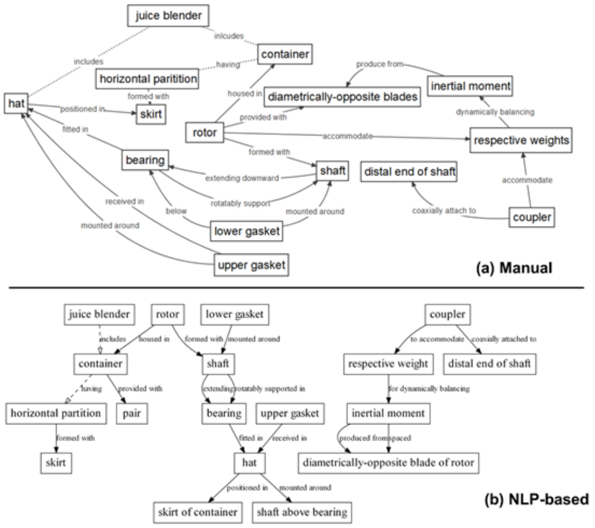

Figure 2: Manual vs. NLP comparison of functional modelling for Patent US6709150 Abstract

customised list. The use of a customised list is because stopwords that are built in the NLTK corpus contain too many meaningful preposition phrases such as 'in', 'of', 'from', and 'to', making it inaccurate for analysing patents if used. The next step is to perform parsing using NLTK Regular Expressions to identify different types of chunks including invention features and relationships.

The parsed results, in the form of an NLTK tree, are then post-processed for recognition of invention features, plus hierarchical and functional relationships. Subject-Action-Object (SAO) structure is used to transform the parsed results into a cluster of triplets centred on each relationship (Action) chunk identified. A recognised relationship chunk is in some cases a verb phrase such as 'for generating', or an adjective phrase such as 'smaller than'. For each relationship chunk, its Subject is considered to be the next invention feature chunk on its left and its Object is considered to be the next invention feature chunk on its right. When navigating the resultant NLTK tree, if the chuck next to a relationship (Action) is not an invention feature chunk then the algorithm will go further left until an invention feature chunk can be identified. The relationship chunks for all formed SAO triplets will go through a comparison with a list of hierarchy phrases, to distinguish hierarchical relationships from functional interactions. Finally, the invention SAO triplets will be plotted with hierarchy SAO and function SAO plotted differently to provide visualisation of a functional model.

The functional model produced using the NLP-based approach here starts by creating nodes, corresponding to each of the invention features identified. Then hierarchy and function SAO triplets are added as links between each pair of nodes. Figure 2 presents an example comparison between the functional model generated manually (a) and that from using the NLP-based approach (b) for a patent abstract, where a reasonable degree of similarity can be observed.

In order to evaluate the effectiveness of the approach, we developed a metric by considering the ratios of matched outcomes amongst the manual and NLP analysis (1). The effectiveness, *E*, is obtained by first multiplying two ratios which provides a value between 0 and 1, and then the value is multiplied by 100, providing a value ranges from 0 to 100. The first ratio indicates the usefulness of the NLP-based approach compared to manual analysis, obtained by dividing the number of matched results by the number of manual results. The second ratio reflects the precision of the approach, by calculating the proportion of matched results amongst all the results obtained by NLP. The manual analysis was carried out by the authors following the guideline introduced in [Atherton *et al.*, 2017].

$$E = \frac{Matched\,results}{Manual\,results} \times \frac{Matched\,results}{NLP\,results} \times 100 \quad (1)$$

In our experiments we noticed that *E* values for each patent and each descriptive section differ quite significantly, and average *E* alone is not sufficient to provide a reliable insight. We thus decided to consider standard deviation into the analysis to reflect the variation in the effectiveness values. A signal-to-noise ratio SNR concept was adopted, obtained by dividing the average effectiveness *E* by the standard deviation, expressed in Equation (2), where $\overline{E}$ stands for the average effectiveness of the samples and σ stands for the stand deviation among the samples. The best patent descriptive sections for analysis can be identified by looking for high SNR, indicating average *E* is strong in relation to its standard deviation. This definition of SNR potentially enables a more insightful analysis to be performed by looking at not only the effectiveness of the NLP-based approach but also its reliability when analysing different patents.

$$SNR = \frac{\overline{E}}{\sigma} \quad (2)$$

The results suggest that the patent claim is best suited to invention feature and hierarchy identification whilst the independent claim is best suited to invention design principle capture (see Figure 3). The results also reveal the limitation of the off-the-shelf NLP-based approach used, suggesting that better benchmarking could be achieved if more tailored techniques are developed and applied.

Compared to the manual approach which requires roughly 10 minutes to complete [Jiang *et al.*, 2018], the NLP-based approach only takes 3 seconds to produce the diagram and associated outcomes, suggesting that designers can save considerable time, especially when analysis of a large number of patents is required.

1.2 Challenges

From the results, it is obvious that the approach works best when identifying invention features for all patent descriptive sections, supported by larger average effectiveness and signal-to-noise ratio. This is within expectation because the invention feature exists in the simplest form in the context of NLP, e.g. compound nouns hence the highest effectiveness can be achieved. However, in some cases, the same feature with varied expressions is treated as a distinct feature in the automated analysis. For example, 'skirt' and 'skirt of container' are regarded as two features but in fact, they refer

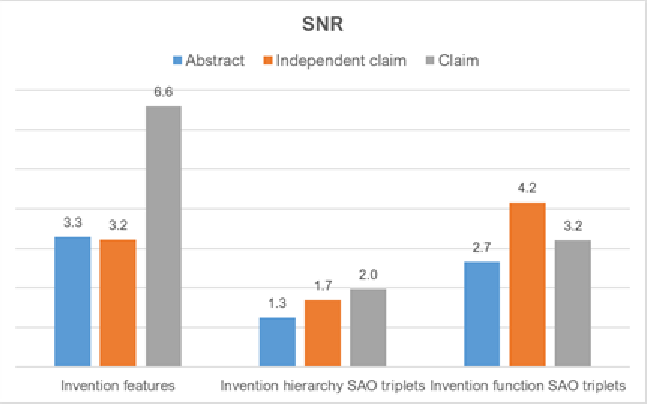

Figure 3: Comparison of average effectiveness

to the same one. This type of problem would normally be avoided by a designer. In some other cases, the invention features identified contain irrelevant elements, e.g. 'pair', 'shaft above bearing'. This is due to the POS taggers assigned to the tokenised words were not sufficiently accurate and this further leads to inaccurate identification of invention features and relationships. For example, 'juice blender comprising', in which 'juice blender' is supposed to be an invention feature and 'comprising' is supposed to be a hierarchical verb but in the analysis 'comprising' was tagged as a noun hence forming a compound noun with 'juice blender'.

The limitations of the approach developed so far can be summarised into three aspects:

1. Incapacity of consolidating variations of expressions that describe the same feature.
2. Incapacity of assigning highly accurate POS tags.
3. Incapacity of recognising inferred relationships in complex sentences

With respect to limitation 1, the implementation of an ontology could help. An ontology is a formal representation of concepts, data and relationships. It can be applied to create a conceptualised representation of the same feature expressed in various ways. However, as patent describes the state of the art in a narrow field it is nearly impossible to construct an ontology that works for all. Domain-specific ontology is a potential solution that can become powerful when dealing with patents around one specific topic. For example [Jiang *et al.*, 2018], a domain-specific ontology on beverage can patents was constructed to enable invention feature consolidation. With respect to limitation 2, specific POS taggers could be developed to improve the accuracy of tagged words. The default NLTK POS tagger used in this study enables rapid analysis of patents with a compromised accuracy. However, similarly, in order to train a POS tagger to provide more accurate results, a target domain will be necessary and machine learning is often needed (e.g. see [Mohammed, 2020]) to outperform standard toolkits. With respect to limitation 3, more sophisticated parsing grammar and text pre-processing techniques could be applied to recognise inferred relationships. This limitation mainly applies when Claims are being analysed whereas the Abstract tends to have simpler sentence structures.

1.3 Projects

This activity is part of an UK's ESPRC funded project, named Patent Knowledge Design Tool [PKDT, 2021], in which some from our staff has been involved as Co-Investigator within an ongoing partnership with the Brunel University London - Department of Mechanical and Aerospace Engineering, College of Engineering, Design and Physical Sciences.

1.4 Resources

- Sorce, S., Malizia, A., Gentile, V., Jiang, P., Atherton, M., Harrison, D., Evaluation of a Visual Tool for Early Patent Infringement Detection During Design, 7th International Symposium on End-User Development (IS-EUD 2019), Hertfordshire, 10-12 July, 2019.
- Jiang, P., Atherton, M., Sorce, S., Harrison, D., Malizia, A., Design for invention: a framework for identifying emerging design–prior art conflict, Journal of Engineering Design, Vol 29(10), Taylor & Francis, online 12 September 2018: 596-615. ISSN: 0954-4828 (print) ISSN: 1466-1837 (on-line). doi: 10.1080/09544828.2018.1520204.
- Sorce, S., Malizia, A., Jiang, P., Atherton, M., Harrison, D., A Novel Visual Interface to Foster Innovation in Mechanical Engineering and Protect from Patent Infringement, 2nd International Conference on Graphics, Images and Interactive Techniques (CGIIT), Hong Kong, China, 23-25 February, 2018. In Journal of Physics: Conference Series (Vol. 1004, No. 1, p. 012024). IOP Publishing. doi :10.1088/1742-6596/1004/1/012024.

References

[Atherton *et al.*, 2017] Mark Atherton, Pingfei Jiang, David Harrison, e Alessio Malizia. Design for invention: annotation of functional geometry interaction for representing novel working principles. *Research in Engineering Design*, 29(2):245–262, September 2017.

[Jiang *et al.*, 2018] Pingfei Jiang, Mark Atherton, Salvatore Sorce, David Harrison, e Alessio Malizia. Design for invention: a framework for identifying emerging design–prior art conflict. *Journal of Engineering Design*, 29(10):596–615, 2018.

[Jiang *et al.*, 2021] Pingfei Jiang, Mark Atherton, e Salvatore Sorce. AUTOMATED FUNCTIONAL ANALYSIS OF PATENTS FOR PRODUCING DESIGN INSIGHT. *Proceedings of the Design Society*, 1:541–550, July 2021.

[Mohammed, 2020] Siraj Mohammed. Using machine learning to build pos tagger for under-resourced language: the case of somali. *International Journal of Information Technology*, pages 1–13, 2020.

[PKDT, 2021] PKDT. On-line tool to intelligently interpret design information contained in patents. https://www.brunel.ac.uk/research/projects/on-line-tool-to-intelligently-interpret-design-information-contained-in-patents, 2021.