Artificial Researcher

Date
02/12/2020

Author
Linda Andersson, CTO/CEO Artificial Researcher GmbH
linda.andersson@artificialresearcher.com

Domain Knowledge makes Artificial Intelligence Smart

Now is the Time for Deep Learning and Natural Language Processing in Patent Text Mining

To develop patent text mining tools for scientists and patent experts, we need to understand their daily work tasks, as well as the linguistic characteristics of the text genre (i.e. patentese). In this talk, we will focus on research results that compare and combine supervised and unsupervised techniques for one real world patent text mining tool. To this day, many frequently used text mining methods still postulate that single words taken by themselves, e.g. bag-of-words, can capture the entire scope of a semantic concept. For many text genres and languages, this is a valid premise, however this is not true for text genres and languages characterized by frequent multi-word unit occurrences used to describe domain-specific concepts. Consequently, many of the state-of-the-art text mining techniques, as well as Natural Language Processing (NLP) tools have significant lower performance when applied on patent text.

In the patent domain, all types of issues, from very specific search requirements to the linguistic characteristics of the text domain, are accentuated. Patent text is a mixture of legal and domain-specific terms. In processing technical English texts, a multi-word unit method is often deployed as a word formation strategy in order to expand the working vocabulary, i.e. introducing a new concept without the invention of an entirely new word. This productive word formation is a well-known challenge for traditional NLP tools utilizing supervised machine learning algorithms due to the limited amount of domain-specific training data (labelled data). The out-of-domain data issue increases the unseen events and out-of-vocabulary term occurrences, which negatively affect the performance of the text mining tools. In comparison, deep learning algorithms do not require large amount of manually labelled training data since the algorithms derive knowledge out of unlabelled data (hence unsupervised methods). However, using an unsupervised method does not completely exclude labelled data since the deep learning algorithms still require (labelled) test data for performance evaluation. Furthermore, depending on the task, some labelled data seeds may be required to initiate the learning process.

Deep learning will help us to better design text mining tools, but will not remove the computational linguistic design process associated with text mining tools (Manning, 2015). There has been extensive work on applying deep learning algorithms to different text mining tools such as Information Retrieval (IR) and Information Extraction (IE) and, so far, they have improved on classic IE and IR tasks. However, when deploying the algorithms on more advanced tasks, such as semantic role labelling or domain-specific tasks, there is still more work to be done (Collobert et al., 2011), (Wang et al., 2016), (Rigouts Terryn et al., 2020).

Deep learning algorithms have several advantages compared to the supervised NLP methods. However, in this talk we will also address pitfalls associated with domain-specific text mining utilizing deep learning algorithms:

- The unsupervised algorithms need a significant amount of data in order to achieve implicit learning from it, while supervised algorithms do explicit learning but will only learn from the little data they are trained on.
- The unsupervised methods require a representative data set in order to reflect implicit learning that should take place. The notion "the more data the better will the performance become" is not entirely correct. If the data is unbalanced, the algorithms will still end up with issues regarding unseen events and out-of-vocabulary term due to the fact that implicit knowledge could not all be derived from the given data.
- Another topic which require more research attention is the risks of incorrect learning by the unsupervised algorithms. Leaving the algorithms to learn by itself with no guides of feature selection (labelled data), as well as, natural biases in the data, the learning outcome may be limited or even make the tool inoperative for usage.

We will present a multi-word term extraction tool, where we combine the Bidirectional Encoder Representations from Transformers (BERT) deep learning framework with domain knowledge utilizing the IPC taxonomy. This tool is an extension of the work presented in (Andersson et al 2017), (Andersson et al 2016), (Fink et al 2019). By combining the domain knowledge with supervised NLP and deep learning methods, we achieved the best performance. Our patent passage retrieval system, for example, is state-of-the-art since 2016. A showcase of our *Passage Retrieval Service* is available on https://artificialresearcher.com/.

References

Andersson L., Hanbury A., Rauber A. (2017) *The Portability of three type of Text Mining Techniques into the patent text genre*. In M. Lupu, K. Mayer, J. Tait, and A. J. Trippe, Second edition, Current Challenges in Patent Information Retrieval

Andersson, L., Lupu, M., Palotti, J., Hanbury, A., and Andreas, R. (2016) *When is the time ripe for natural language processing for passage patent retrieval monitoring of vocabulary shifts over time*. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM16.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011) *Natural language processing (almost) from scratch.* J. Mach. Learn. Res., 12:2493–2537, November.

Fink T., Andersson L., Hanbury A. (2019) *Detecting Multi Word Terms in Patents the same way as Named Entities.* In Proceeding 1st PatentSemTech Workshop, (Extended Abstract)

Manning, C. D. (2015) *Computational linguistics and deep learning.* Computational Linguistics, 41(4):701– 707.

Rigouts Terryn, A., Hoste, V., Drouin, P., & Lefever, E. (2020) *Termeval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (acter) dataset.* In 6th International Workshop on Computational Terminology (COMPUTERM 2020) (pp. 85-94). European Language Resources Association (ELRA).

Wang, R., Liu, W., & McDonald, C. (2016) *Featureless domain-specific term extraction with minimal labelled data.* In Proceedings of the Australasian Language Technology Association Workshop 2016 (pp. 103-112).