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

Artificial Researcher a Technical Summary

Domain Knowledge makes Artificial Intelligence Smart

To develop 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 genres. 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-art text mining techniques, as well as Natural Language Processing (NLP) tools have significant lower performance when applied on domain-specific text genres.

Methodology

At the core of our R&D we focus on developing software which make use of both supervised and unsupervised techniques for scientific and patent text mining tools. We derive our text mining solution from the linguistic characteristic of a text genre as well as the language, both in the indexing and in the search process. For example, re-ranking technologies based upon clicks or user history, a common practice among commercial search engines, have limitation in the scope to retrieve and explore new information. Therefore, instead of customize the search results based upon user history, and clicks, our re-ranking algorithms measure the confident in the query terms themselves using deep learning to estimate levels of domain terms as well as novel terms to obtain balance between relevancy and redundant information. To obtain a wider span of scientific fields we alternate the collections and test different classification technologies on the showcase page (https://passageretrieval.artificialresearcher.com/) to receive more feedback from users. In our showcase page we do collect search information to optimize our algorithms, but we do not use cookies or collect information about the users. With our collaboration with universities we work to refinement of search and index technology to meet the need of different scientific fields and text genres.

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 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, there are several 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.

In our R&D process we compare and combine unsupervised and supervised methods in order to optimize of algorithms’ learning curves but at the same time diminish effect of the limitation of each technology.

Artificial Researcher Services

The Artificial Researcher services and software, aim to provide a holistic approach to the up-and-coming AI solutions required by industry as well as academy by offering a palette of services and software such as text segment similarity, ontology population, automatic term recognition as well as different type of search applications.

In our Artificial Researcher Passage Retrieval Service:

- To improve cross-genre retrieval and optimize the search functions our system now includes identification of technical terms based on the WIPO catchword Index for both scientific publications as well as patents. The paragraphs have been assigned technical terms using an automatic term recognition tool which combines SciBERT and the IPC information.
- We have processed the EP full-text data for text analytics collection.
- Established a process to design special focus indexes, making it possible for our clients to specify the data to be indexed.
- We have curated Open Access Data together with our provider CORE.uk, we have access to over 200 million scientific open access publications.

The Artificial Researcher Ontology Service, which we have developed a long side the AR-Passage Retrieval Service, we use the automatic term recognition tool to extract single words as well as phrases and with a combination of NLP and Distributional Semantic connect the terms to related concepts. As a first test project, for our AR-Ontology Service, we used the COVID-19 Open Access data set and out of 180,000 publications we could generate 1.3 million candidates. This was a co-creation project supported by the European Open Science Cloud to explore more NLP using a combination of Deep Learning and NLP to extract hypernym and hyponym. We are currently test processing the EP full-text data for text analytics collection through the system in order to generate candidate terms and related concepts extracted from patent documents.

Our Artificial Researcher NLP-toolkit Services are used as modules part of our Passage Retrieval as our Ontology Services but are also standalone services to provide the text mining industry and users with a set of NLP tools from domain-specific trained Distributional Semantic models, domain entity extractions, to cross-genre classifications between patent, scientific fields and medical publications.

References

Linda Andersson, Allan Hanbury, Andreas Rauber (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. 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.

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).