A Platform for the Analysis of Document Content with Machine Learning and NLP: The Kairntech "Sherpa"

Stefan Geißler
Kairntech – www.kairntech.com

kairntech
From Documents to Business Impact

Introduction

The Sherpa – Powerful Machine Learning – Intuitive Usability – Diverse Use Cases

Machine learning methods (ML) are setting the standards in many areas of information processing today, This is also the case with natural language processing (NLP): thesaurus-based indexing, document categorization, extraction of persons, locations or key terms can often be performed by ML procedures with otherwise unrivaled quality.

At the same time, ML procedures often place high demands on the **availability of sample data** for training of corresponding models as well as regarding the technical competence required in order to be able to optimize and productively use these models.

The Kairntech Sherpa offers powerful and versatile learning algorithms under an **easy to understand** user interface and allows domain experts to easily and quickly create, evaluate and productively use NLP analysis models. The analysis of scientific papers, contracts or invoices, news texts, patents or web content are only some of the use cases that can be addressed with Sherpa.

The Sherpa is based on methods such as "**Deep Learning**", which achieve first-class results on publicly available evaluation data such as Conll2003. Some of the procedures implemented by members of the Kairntech team are available as open source in the public domain - at the same time they are accessible as part of the Sherpa without programming knowledge and embedded in a modern web interface.

The Sherpa can be installed "on premise" in the client's IT environment, or it can be installed in the cloud and integrated into applications via a detailed REST API (programming interface).

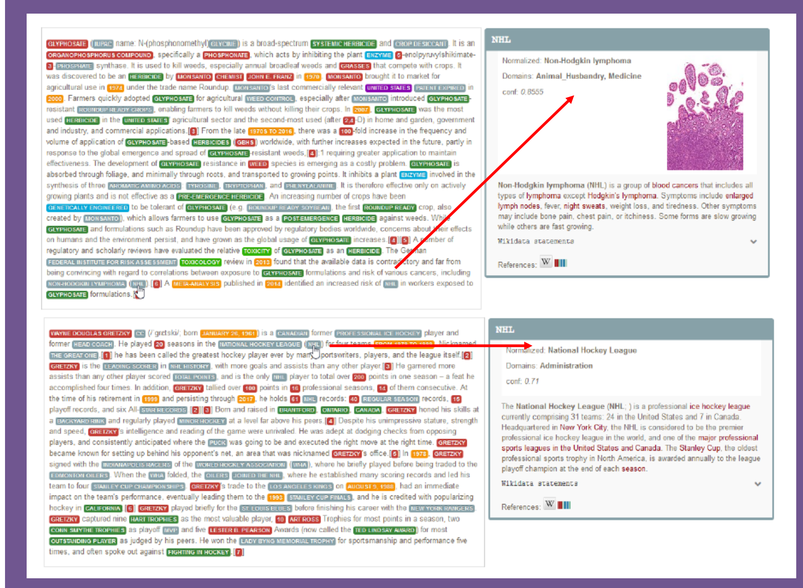
The Sherpa is multi-user capable and supports project-based work.

We present important components and application scenarios of the Sherpa in the following sections.

Thesaurus-based Indexing

Extraction, Normalization, Desambiguation and Linking of Key Concepts

Annotating content with thesauri requires much more than just searching for character strings in the text: Thesaurus terms can appear as variants (motor bike / motor cycle) or inflected in the text (mouse / mice), one and the same term can have different meanings (charge: the process of loading a battery or a fee or a n indictment, etc) and finally the indexing of content benefits when the recognized terms are linked to background information (e.g. image and CV of a prominent person, geolocation of a place, web page of a company, structural formula of a chemical substance).



The Sherpa allows to access data with large thesauri - here Wikidata with almost 80 million concepts in several languages.

In the example on the left, the system recognizes that the string "NHL" stands for "Non-Hodgkin Lymphoma" in the first case, but for the "National Hockey League" in the second. For this desambiguation **no manual rule writing** is necessary. The learning procedure recognizes the correct meaning based on the **context**.

Annotating Data: Easy and fast

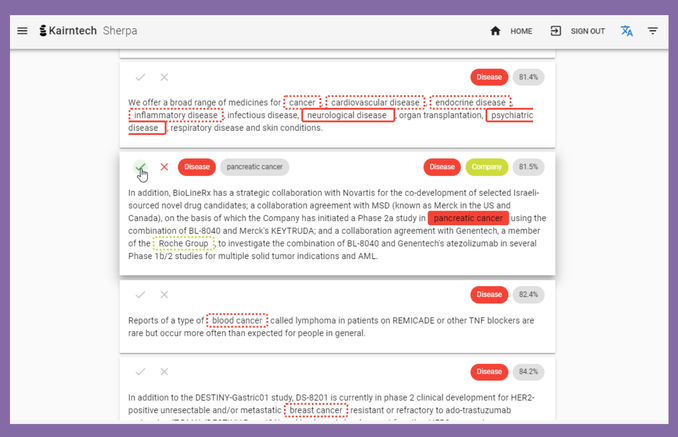
ML approaches require Training Data: The Sherpa supports you in their Creation.

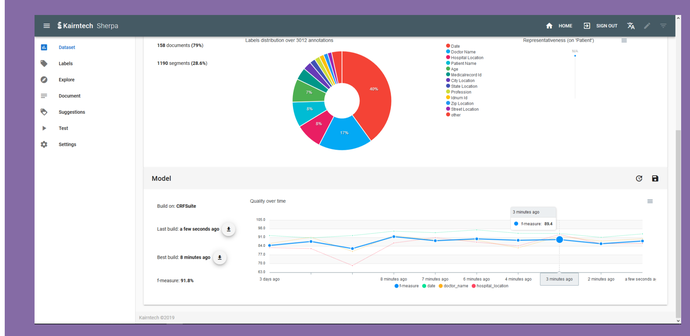
Modern learning methods are often comparatively data-hungry: In order to use their full potential, they often need to be trained on large numbers of examples. The **manual creation of such training data can be extremely time- and cost-intensive**.

The Sherpa user interface allows you to **quickly annotate large amounts** of data with simple mouse and keyboard actions.

A model **constantly learns in the background** and generates new data on demand. The users thus always receive feedback on the currently achieved quality.

By using "**Active Learning**", the required amount of data is reduced to a minimum.



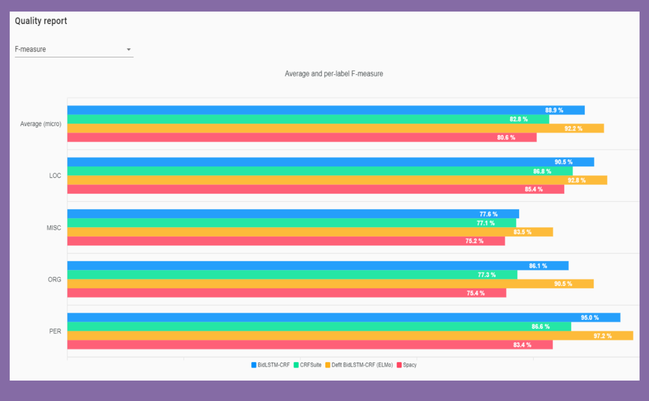


The Sherpa provides **detailed feedback** on the corpus data, the distribution of the different information types and the quality of the trained models.

Users can thus reliably see whether the most recently added data have further improved the quality or whether the quality targets they set themselves have already been achieved.

Different ML algorithms often show different behaviour: One approach may be able to be trained very quickly, but ultimately does not achieve the desired quality. Another approach, on the other hand, may require much more training time, but then delivers the best results.

The Sherpa allows the application of different learning methods to the data of a project and the selection of the best result for the respective requirements.

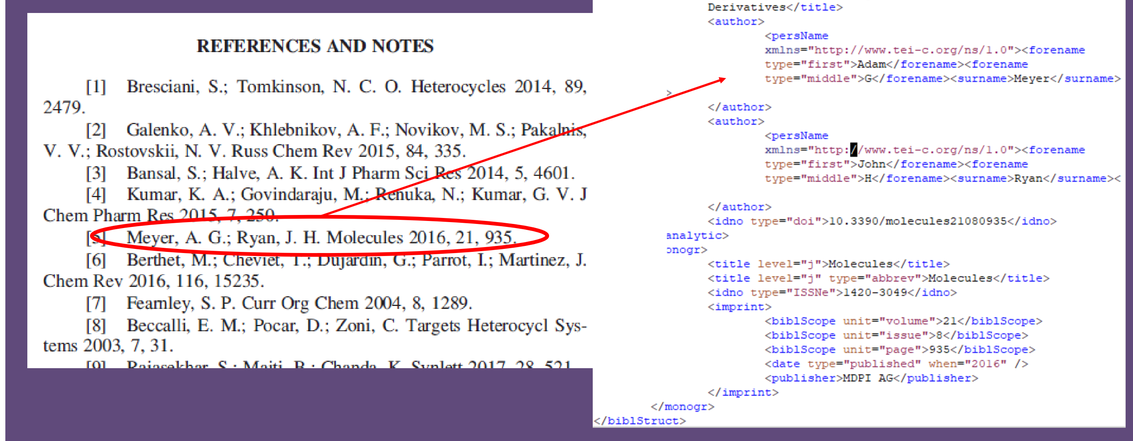


Our key ML library by the way is available as **open source software** at https://github.com/kermitt2/delft

The Sherpa can either import pre-notated data or raw data that is turned into a training corpus through manual processing. The manual annotation is often a necessary evil - time-consuming but unfortunately necessary wherever training data is not already available. The implementation of the annotation process in Sherpa therefore places great emphasis on intuitive and simple user guidance. Extensive tests with users confirm the **ease of use**: users tell us: "You really feel like annotating more data", "The immediate feedback from the system and the constantly improving results are especially motivating", "It's really addictive!

Document Structure Recognition: PDF → XML

PDF is the de facto standard for the exchange and storage of documents. A disadvantage of the format is, however, that **important structural information** about the document, that was still available when it was created, is **no longer explicit in PDF**, but must instead be elaborately inferred again: What is the title of the document? Who are the authors? The publication date? Is there an abstract? Which other papers are cited and what were their authors?



REFERENCES AND NOTES

[1] Bresciani, S.; Tomkinson, N. C. O. Heterocycles 2014, 89, 2479.
[2] Galenko, A. V.; Khlebnikov, A. F.; Novikov, M. S.; Pakalnis, V. V.; Rostovskii, N. V. Russ Chem Rev 2015, 84, 335.
[3] Bansal, S.; Halve, A. K. Int J Pharm Sci Res 2014, 5, 4601.
[4] Kumar, K. A.; Govindaraju, M.; Renuka, N.; Kumar, G. V. J Chem Pharm Res 2015, 7, 259.
[5] Meyer, A. G.; Ryan, J. H. Molecules 2016, 21, 935.
[6] Berthet, M.; Cheviet, T.; Dujardin, G.; Parrot, I.; Martinez, J. Chem Rev 2016, 116, 15235.
[7] Fearnley, S. P. Curr Org Chem 2004, 8, 1289.
[8] Beccalli, E. M.; Pocar, D.; Zoni, C. Targets Heterocycl Systems 2003, 7, 31.

<biblStruct xml:id="b4">
<analytic>
<title level="a" type="main">1,3-Dipolar Cycloaddition Reactions of Azomethine Ylides with Carbonyl Dipolarophiles Yielding Oxazolidine Derivatives</title>
<author>
<persName xmlns="http://www.tei-c.org/ns/1.0"><forename type="first">Adam</forename><forename type="middle">G</forename><surname>Meyer</surname></persName>
</author>
<author>
<persName xmlns="http://www.tei-c.org/ns/1.0"><forename type="first">John</forename><forename type="middle">H</forename><surname>Ryan</surname></persName>
</author>
<idno type="doi">10.3390/molecules21080935</idno>
</analytic>
<monogr>
<title level="j">Molecules</title>
<title level="j" type="abbrev">Molecules</title>
<idno type="ISSNe">1420-3049</idno>
<imprint>
<biblScope unit="volume">21</biblScope>
<biblScope unit="issue">8</biblScope>
<biblScope unit="page">935</biblScope>
<date type="published" when="2016" />
<publisher>MDPI AG</publisher>
</imprint>
</monogr>
</biblStruct>

Powerful ML methods use a variety of information (position in the document, fonts, layout, ...) to determine the structure of the analyzed documents and to provide the document as TEI XML for further evaluations (e.g. citation networks). In the example above, the bibliography of a paper is analyzed: The system recognizes the authors, the title and the journal and can automatically expand the author "Meyer, A. G.", which is only mentioned in an abbreviated way in the PDF, as "Adam G. Meyer".

The system used is available as open source (https://github.com/kermitt2/grobid) and pre-trained for use on **scientific documents**. Processing other document types is possible, but requires the training of the corresponding models by Kairntech.

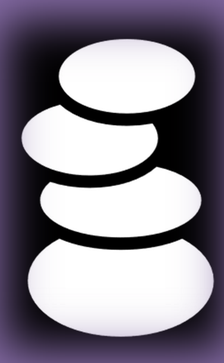
Cascades of ML models are used here: a first process determines the rough structure, then details within the sections recognized in step 1 are computed.

Kairntech – Who we are.

Kairntech is a specialized provider of **software and professional services** in the field of **artificial intelligence (AI) and NLP**. Kairntech was founded in December 2018 by a team of experts who had previously worked together for years on similar topics: At Xerox, IBM, INRIA and TEMIS.

Different components of the Sherpa are in productive use at **CERN, NASA, ResearchGate, EPO, Mendeley**, ... The Sherpa is used for example at the German pharmaceutical company **Boehringer Ingelheim.**

Our logo - the cairn - alludes to the mission we have set ourselves: Just as a cairn shows mountaineers the way through rough terrain, we are ready to accompany our customers through the sometimes confusing world of AI and NLP. Furthermore, our headquarters are located in the mountains in Grenoble, within sight of the Mont Blanc.



Contact: **www.kairntech.com**, email: **info@kairntech.com**