Project

Nowadays scientific knowledge can be published digitally within many different forms and sources, such as encyclopedias, scientific papers, regulatory documents, but also structured knowledge sources like ontologies or knowledge bases. Beside that also news articles, blog posts, forums or social media can contain relevant information or can be used for research. All this is published everyday in a large number of different languages. The volume and speed of production of digital content has become too fast however in some domains for humans to be able to keep up with them and maintain an up-to-date view of current scientific evidence. In MEDLINE for instance every year close to one million new articles are included.

The present project aims to design Artificial Intelligence (AI) methods that **automatically digest these different types of text sources and jointly extract such knowledge and observations in order to populate existing knowledge bases**. Our project showcases these methods in the domain of **pharmacovigilance**, which endeavors to maintain up-to-date knowledge on adverse drug reactions (ADRs) for the benefit of public health. In this domain, authoritative sources include scientific journals and drug labels while elementary observations are reported in patient records and social media.

Current mainstream information extraction methods use self-supervised extraction of word representations from large text corpora and tend to neglect existing knowledge on the target domain. In contrast, the present project aims to **integrate existing knowledge** into the word representation acquisition and information extraction processes to improve the extraction of new information and knowledge. This is all the more needed to address less formal sources and hence more challenging sources such as social media. Additionally, it will take advantage of the existence of similar information published in **multiple languages to pool knowledge across countries**.

Literature mining can boost the collection of both current knowledge and additional elementary observations, resulting in automatically maintained digital encyclopedias in the form of knowledge graphs usable for both machine inference and human display. We believe this may further apply to various scientific fields such as global warming that need to collect and integrate elementary observations into current knowledge. Language barriers hamper the free flow of knowledge and thought across languages. Relevant findings need to be articulated across these barriers, which requires time and effort to collect and translate into the respective languages. In the not too distant future, tools will assist researchers and other citizens in finding and linking information distributed across sources and languages. In this project, we will help to improve such technologies and will demonstrate them for adverse drug reactions.

This cross-language dimension obtains a clear benefit from the proposed trilateral collaboration. To strengthen our collaboration and mutual knowledge, we plan internships for **early career researchers** at each of the other two partner teams under joint supervision, as well as plenary, jointly taught training actions, to provide them with a **shared international exposure and training** and build the ambassadors of tomorrow's partnerships.

The consortium is composed of three internationally recognized teams specialized in natural language processing. NAIST (JP) has created the de-facto natural language processing tools for Japanese, and produced a number of document and text analysis tools for extracting knowledge from scholarly documents. DFKI (DE) has a strong background in corpus generation, general information extraction and biomedical text processing. LISN (FR) has a long and strong experience in corpus annotation, hybrid information extraction and biomedical language processing, including for pharmacovigilance

from patient forums.