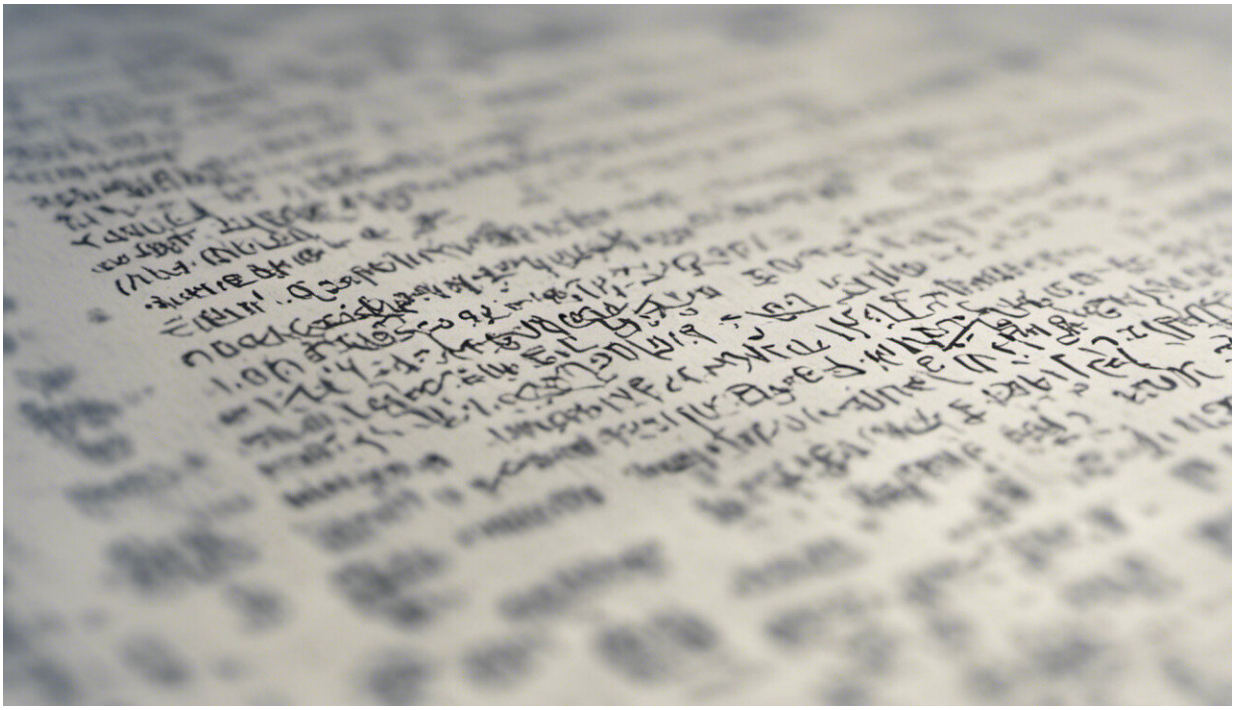
Text mining for chemists

June 6 2016



Credit: AI-generated image (disclaimer)

A collaboration between two companies in Hungary and the UK has resulted in the inception of the first ever interactive text mining platform for chemists, overcoming difficulties with extracting information about chemicals from scientific literature.

Scientists and researchers today have exponential volumes of information at their fingertips – but it can be very difficult to extract and

analyse relevant data from the scientific literature. Traditional keyword searches are often limited as they are only able to retrieve documents containing the pre-specified keyword, leaving it up to the individual to go through the documents and analyse them for relevance. Text mining software has been hailed as a welcome solution to this problem. It uses sophisticated algorithms to derive high-quality information from documents – and it recognises concepts rather than words, highlighting patterns and correlations within and between vast datasets.

Although text mining is pretty well established in some fields – like drug development, for instance – it has not been of much use for discovering novel compounds. This is because it cannot understand every single part of the long chemical names under scrutiny, preventing it from performing deep chemical analyses. This knowledge gap was identified some years ago by researchers at Linguamatics in Cambridge and ChemAxon in Budapest. The two companies had already integrated their software products to enable text mining for known compounds – and, supported by the EUREKA Network, they embarked on a new project that went a step further than this.

An integrated solution

The ChiKEL project – which stood for Chemically Informed Knowledge Extraction from Literature –centred on the development of an interactive text mining platform aimed at the unique needs of chemists that would have the capacity to understand, analyse and deal systematically with long chemical names. It extended the existing software integration between Linguamatics and ChemAxon by enabling the recognition of novel chemical compounds expressed via words or images. "Our aim was to combine chemical search and text mining so that users can find relevant chemicals and their properties or relationships from unstructured texts such as patents or scientific articles," outlines Dr David Milward, Chief Technology Officer at

Linguamatics.

The project was a success because we knew from the beginning what we wanted to achieve and we had a very well defined plan that described achievable deliverables.

This fully automated approach enables users to extract information on biological and chemical entities in a range of documents, paving the way for deeper analyses. Moreover, the software presents the search results in an enhanced way so that users can view chemical structures as well as browse through the clusters of structures found within the documents. "The project was a success because we knew from the beginning what we wanted to achieve and we had a very well defined plan that described achievable deliverables," states Dr Krisztian Niesz, a business analyst at ChemAxon who participated in the project. "We communicated with Linguamatics in a biweekly basis and worked together in an agile framework."

The project has helped ChemAxon become a global leader in naming technology, while Linguamatics has benefited by achieving plans to provide its customers with fully chemically enabled text mining software. "Sales of chemically enabled text mining have primarily been based on upsell within existing accounts – but also to new groups within these accounts such as patent searchers," remarks Milward. Going forward, as use of this novel text mining software becomes more widespread, it is anticipated that it will provide the apparatus for new drug discoveries and fuel the growth of the pharmaceutical industry.

Provided by Eurostars