Big data for text: Next-generation text understanding and analysis

March 7 2016

News portals and social media are rich information sources, for example for predicting stock market trends. Today, numerous service providers allow for searching large text collections by feeding their search engines with descriptive keywords. Keywords tend to be highly ambiguous, though, and quickly show the limits of current search technologies. Computer scientists from Saarbruecken developed a novel text analysis technology that considerably improves searching very large text collections by means of artificial intelligence.

Beyond search, this technology also assists authors in researching and even in writing texts by automatically providing background information and suggesting links to relevant web sites. Living in the age of business smartphones and enterprise chatrooms, most information in companies is not distributed via spoken words but rather through e-mails, databases, and internal news portals. "According to a survey by the market analyst Gartner, a mere quarter of all companies are using automatic methods to analyze their textual information. By 2021, Gartner predicts 65 per cent will do so. This is because the amount of data inside companies is continuously growing and hence, it becomes more and more costly to have it structured and to search it successfully," says Johannes Hoffart, a researcher at the Max Planck Institute for Informatics and founder of Ambiverse. His team developed a novel text analysis technology for analyzing huge amounts of text where massive computing power and artificial intelligence (AI) are continuously "thinking along" in the background.

"For analyzing texts, we rely on extremely large knowledge graphs which are built upon freely available sources such as Wikipedia or large media portals on the web. These graphs can be augmented with domain- or company-specific knowledge, such as product catalogs or customer correspondences," says Hoffart. By applying complex algorithms, these texts are screened further and analyzed with linguistic tools. "Our software then assigns companies and areas of business to their corresponding categories, which allows us to gather valuable insights on how well one's own products are positioned in the market in comparison to those of the competitors," he explains. Particularly challenging hereby is the fact that product or company names are anything but unique and tend to have completely different meanings in different contexts, making them highly ambiguous.

"Our technology helps to map words and phrases to their correct objects of the real-world, that way resolving ambiguities automatically," explains the computer scientist. "Paris" for example stands for the city of light and the French capital, but also for a figure from Greek mythology or a millionfold-mentioned party girl with German ancestors - always depending on context. "Efficiently searching huge text collections is only possible if the different meanings of a name or a concept are correctly resolved," says Hoffart. The smart search engine developed by his team continuously learns and improves over time, and also automatically associates new text entries to matching categories. "These algorithms are hence attractive for companies that analyze online media or social networks to measure the degree of brand awareness for a product or the success of a marketing campaign," says Hoffart further.

At Cebit, Ambiverse will further present a smart authoring platform that assists authors in researching and writing texts. Users who enter texts are automatically provided with background information, for example company-internal guidelines and manuals or web links. "Relevant concepts are linked automatically and links for further research are

show", says the computer scientist.

Visitors to the Ambiverse Cebit booth (hall 6, booth 28) will also have the opportunity to compete with their novel AI technology by playing a question-answering game. Ambiverse is funded by the German Federal Ministry for Economic Affairs through an EXIST Transfer of Research grant.

Ambiverse, a spin-off company from the Max Planck Institute for Informatics in Saarbruecken, will be presenting this novel technology during Cebit 2016 in Hannover from 14 to 18 March at Saarland's research booth.

Provided by Saarland University