Software maps ambiguous names in texts to the right person

February 26 2014

A new technique enables it to query documents by means of keywords, entities, and categories. Credit: Johannes Hoffart/Max Planck Institute

If a name is ambiguous and given without context, even humans struggle. When reading the last name "Merkel", people do not know if it refers to the Chancellor of Germany Angela Merkel or the famous soccer coach Max Merkel. It is a drawback for web search, too. Up to now, the programs can capture character strings like "Angela Merkel", but they do not pay attention to attributes like "German Chancellor" or "Germany's First Lady" at all. Even worse, after the word "Merkel" is entered, the search engines provide information about a lot of people with the same last name.

Researchers at the Max Planck Institute for Informatics have now

developed a program that enables accurate disambiguation of named entities by analyzing them with the help of the free Internet encyclopedia Wikipedia. Their software named AIDA establishes connections between the mentions in the text and potential persons or places. "The more references exist between a mention and a specific person in Wikipedia, the more words of the person's Wikipedia article can also be found in the input text, and the higher the score the mention-entity edge receives. AIDA checks this score and selects the mention-entity edge with the highest score as the accurate mapping," explains Johannes Hoffart, who co-developed AIDA at the Max Planck Institute for Informatics.

To demonstrate their novel technique, the researchers have implemented a search engine based on their approach. The search engine makes it possible not only to combine the search for strings with the search for specific objects like persons and locations, but also to search on categories. In this way, the search for "Angela Merkel + phone call + Ukrainian politicians" results in texts dealing with the German Chancellor within the context of Ukrainian politicians like "Yulia Tymoshenko" and the string "phone call". Currently the researchers use AIDA to analyze the text corpus of the German National Library to combine the search for keywords with the search for specific objects. "The search results are more precise this way", Hoffart points out.

"With our new technique we can not only build better search engines, but also make computers understand texts almost as a human does, in an efficient way," explains Gerhard Weikum, Scientific Director at the Max Planck Institute for Informatics in Saarbrücken. The approach also opens new possibilities for automatically generated recommendations and the analysis of datasets, says Weikum, who also does research at the Cluster of Excellence "Multimodal Computing and Interaction" in Saarbrücken. "Whoever is a fan of the soccer coach Merkel will receive recommendations for his books. Those more interested in the Chancellor

get referred to books dealing with her and her way of governing Germany," Weikum explains.

Provided by Saarland University