Through analysis of 'named entities', computers can extract more information from texts

May 23 2014, by Kim Bekmann

Mena B. Habib, a researcher at the University of Twente CTIT research institute, teaches computers to improve their reading comprehension. He developed a method by which computers can detect and interpret 'named entities' in a text. These are, for example, names of people, places and organizations, whose importance is dependent upon the context. Habib's

method allows computers to analyze the context and thus determine what is meant by the named entity.

Named entities

Maurice van Keulen, senior lecturer for Data Management Technology at the University of Twente, supervised Habib during his doctoral research. He explains: "An example of a named entity is rijksmuseum. The context determines which 'rijksmuseum' (national museum) is referred to. This may be related to the author, the subject of discussion, what was said before or after and sometimes even the location or the time. If the author lives in Enschede, then he or she is probably referring to the 'rijksmuseum' in Enschede. But he could also be referring to one of the many other national museums in the Netherlands. Another example is Paris Hilton: does this refer to the celebrity, the hotel in Paris, or something else?" With Habib's method, the computer detects which part of the text is a named entity and what is meant by the named entity.

Reading comprehension

There is considerable demand for new methods to extract information from texts. At present, computers can already retrieve quite a lot of information from texts, including the mood and even the age of the writer. Van Keulen: "These techniques are often based on a superficial analysis of plain words. As a result, most of the information remains 'hidden' and is only accessible to computers to a limited extent, unless they learn to read in an understanding manner. With greater understanding of the entities referred to and information available about these entities, computers are better able to extract a lot more information from texts for analysis purposes."

Application

Van Keulen: "We are involved in a number of projects in the scope of which we will apply the method. For the TEC4SE project, for example, we will use the software in the emergency rooms of the Twente fire brigade and police. At major events, the emergency services would like to be aware of what is happening. For example, if there is a disturbance, it is interesting to monitor a channel like Twitter. Our software can read all tweets with some understanding, and is thus able to better detect where and when something is wrong.

Van Keulen: "Habib made sure his method is as strong and robust as possible. The method also works well even if you do not have a lot of texts available to learn from. In addition, his approach is language independent: it doesn't only work for texts in Dutch; it works for texts in any language."

With this research, Habib won the Making Sense of Microposts challenge: #Microposts2013 and came second in 2014. This challenge is an international competition in which research groups perform a joint 'reading comprehension' task with their research prototypes.

The title of Mena Badieh Habib Morgan's PhD thesis is: 'Named Entity Extraction and Disambiguation for Informal Text - The Missing Link'. Habib will defend his PhD thesis on 9 May at the Databases department of the University of Twente CTIT research institute. He conducted his research under the supervision of dr. ir. Maurice van Keulen and prof. dr. Peter Apers.

Provided by University of Twente