Capital Markets CRC Limited

Computer scientists develop new state-of-the-art method for quote attribution

Tim O'Keefe
CMCRC PhD candidate
U. of Sydney

A/Professor James Curran
CMCRC Research Associate
U. of Sydney

Politicians could soon be held accountable for all their quotes! A study demonstrates that sequence labelling techniques produce excellent results for quote attribution of the news.

News stories are often driven by the quotes made by politicians, sports stars, musicians, and celebrities. When these stories exit the news cycle, the quotes they contain are often forgotten by both readers and journalists. A system that automatically extracts quotes and attributes those quotes to the correct speaker would enable readers and journalists to place news in the context of all comments made by a person on a given topic.

A study by Tim O'Keefe, Associate Professor James Curran, Silvia Pareti, Associate Professor Irena Koprinska and Research Fellow Matthew Honnibal presents the first large-scale evaluation of a quote attribution system on newswire from the 1989 Wall Street Journal (WSJ) and the 2009 Sydney Morning Herald (SMH). Previous research into quote attribution had used small-scale or otherwise limited datasets and no two methods had ever been compared on the same data. This meant that researchers had no empirical evidence as to which method worked the best. Tim's research presents the first real comparison of several previous methods and it does this over three large-scale datasets.

Though quote attribution may appear to be a straightforward task, most current approaches proposed thus far use simple rules, which produce disappointing results. One previous approach used machine learning, which produced better results. However, O'Keefe's research showed that it relied on information that would not be available in practise and when this information was removed the accuracy of the system dropped dramatically. O'Keefe was able to reconstruct this information by treating quote attribution as a sequence labelling task. This has greatly improved results over prior research efforts.

A sequence labelling algorithm tries to predict the whole sequence of speakers that is most likely, rather than predicting each speaker in isolation. As such, it makes relatively fewer errors than other methods because it is more robust to errors early in the sequence. O'Keefe's newswire results are impressive with an accuracy result of 92.4% for the SMH and 84.1% for the WSJ, which demonstrates that it is clearly possible to develop an accurate and practical quote extraction and attribution system.

The research compares different methods of quote attribution across large multiple data sets and empirically demonstrates that O'Keefe's sequence labelling approach outperforms its rivals and allows facts and opinions to be extracted more easily by identifying the source in a more robust way.

The Capital Markets Cooperative Research Centre is a world-leading research organisation that provides thought leadership and break-through technology solutions for capital and insurance markets (www.cmcrc.com).