

Sheng Li

Sheng Li is a final-year PhD student in the English Department, and his thesis focuses on stock tweets. Using a corpus-driven approach, he analyses the linguistic features of stock tweets, and then attempts to apply these features to improve the accuracy of sentiment analysis, in order to implement stock prediction based on social media. Prior to his PhD, he was an MPhil student in corpus linguistics. While taking intensive courses on corpus theories and methods, he decided to upgrade to a PhD to continue pursuing his dream.

During the last three years, he has developed his research interests on corpus linguistics, sentiment analysis, and social media analysis. In May 2012, he was invited to give a talk on his research at the BAAL/CUP Seminar 'Discourse and Technology', which perfectly illustrated his research interests. In the seminar, the participants discussed ways to apply a more quantitative-and-qualitative-combined approach to linguistic analysis, and he believes it is one of the approaches with the greatest potential in linguistics, and he adopts it in his studies with confidence.

Corpus linguistics have always been his favourite area of study. When he was doing his MPhil, he used Google Ngram Viewer data to study the political correctness phenomenon in written English in the 20th century. With a focus on the singular third person pronouns, this research used super-big data to gain an overview of how language changed during that long period.

As an emerging area, social media data has captured his attention. In his thesis, he has conducted a detailed analysis to discuss what stock tweets are. Although there are many studies on tweets and the stock market, few of them present a sound argument about the linguistic features of tweets discussing the stock. He manually annotated sample tweets about the General Electrics ticker, and analysed them quantitatively and qualitatively. The result shows that stock tweets differ from other general tweets in a number of aspects.

Sentiment analysis is a recently developed sub-branch of natural language processing, and he uses this approach to investigate the influence of tweets on the stock market. In his dissertation, he argues that filtering irrelevant data will improve the overall accuracy of sentiment analysis. Conventionally, many studies have classified tweet data into three categories, i.e. positive, negative and neutral, and have put all unclassified data into the neutral category. In his opinion, this is inappropriate as noisy data frequently occur in tweet data, and they might have some indirect influence on stock movements. Therefore, he has designed a hierarchic classification criterion to solve this problem.

Apart from his main research interests as aforementioned, he is also fascinated by other innovative research aspects, such as collaborative research and tutoring tools for linguistic analysis.

He collaborated with Dr. Helen Liu and Ruodan Zhang from the University of Hong Kong to investigate the lenders' online profiles on Kiva.org, which is one of the most successful micro-finance online platforms. He applied corpus analysis to investigate 100,000 lender profiles in order to understand their donation motivation. Presenting this project at two social science conferences has made him aware of the need for developing linguistic approaches to social science research.

He is keen on using his programming skills to facilitate linguistic analyses. He participated in the University of Birmingham PhD corpus project led by Dr. Paul Thompson, where he was in charge of crawling and manipulating data for further analysis. He has delivered a number of workshops on programming skills, for instance, at the Linguistics in the Midlands at the University, student seminars in the Department, and the R Meetup group in Birmingham.

He believes that Birmingham is one of the best places to study corpus linguistics, because the Collins COBUILD dictionary inspired him a lot when he was doing his bachelor's degree. He has never regretted choosing to study at Birmingham, as he has received enormous help from the Department.

