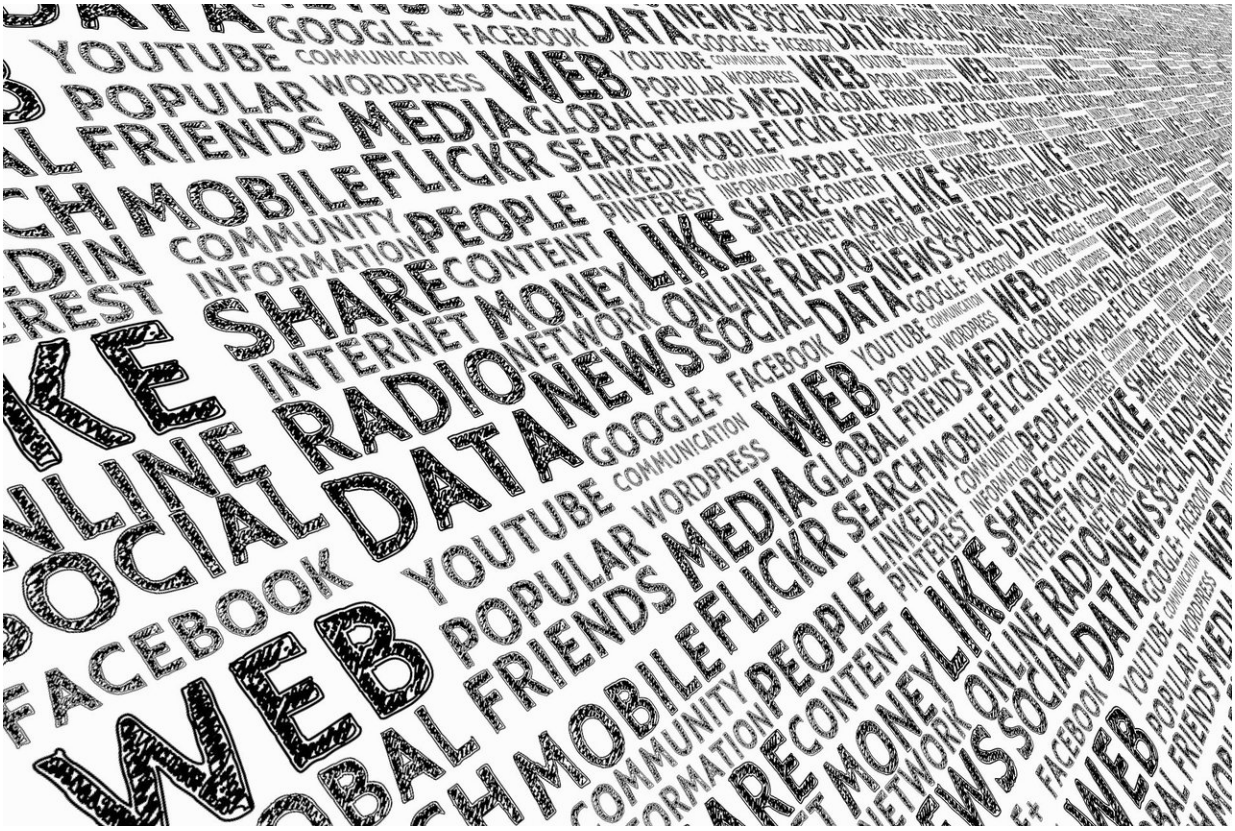
Data mining hyphenated headlines: Improving named entity recognition

January 22 2020, by David Bradley



Credit: CC0 Public Domain

Data mining and extraction of knowledge from disparate sources is big data, big business. But, how does the search software cope with entities that are mentioned where only part of their name is used or a name is

hyphenated when it normally isn't? Research published in the *International Journal of Intelligent Information and Database Systems* reveals details of a new approach to improving named entity recognition and disambiguation in news headlines.

Jayendra Barua and Rajdeep Niyogi of the Department of Computer Science and Engineering, at the Indian Institute of Technology, in Roorkee, Uttarakhand, India, explain that their approach to such an analysis of current news headlines builds on a trained algorithm that has been taught to remove the hyphens and complete incomplete names to remove ambiguity.

The team's evaluation of their novel approach shows that it works with approximately 10 percent greater accuracy than conventional systems and so could improve the automated retrieval of news associated with particular companies, organizations, events, public figures, and other entities of interest to those data mining the news. The system works well with newsfeeds, such as the RSS type of newsfeed generated by regularly updated websites. Headlines from such sources might commonly be longer than conventional newspaper headlines but are nevertheless succinct, commonly being ten or fewer words long. Each word might then be important in a data mining context and so disambiguation is critical.

More information: Jayendra Barua et al. Improving named entity recognition and disambiguation in news headlines, *International Journal of Intelligent Information and Database Systems* (2020). DOI: 10.1504/IJIIDS.2019.104530

Provided by Inderscience