LINGUISTIC ANALYSIS OF SOCIAL MEDIA

Grefenstette G. (Gregory.Grefenstette@3ds.com)

3DS Exalead, Paris, France

One can look upon the Web as a large corpus that can teach us about language use, and also about the real world. In order to determine what is new or interesting we need to know what the norm for language use is. This involves creating a language model that corresponds to what is found on the web. Since the web is so big, it is impossible to download it all and count appearances of words and phrases, so one must use the technique of probing: generating things to be tested and submitting them to a search engine to find their frequency of occurrence. It has been shown using Google to gather statistics is perilous since Google does not provide exact counts but rather estimates the number of pages containing an expression. These counts can be very far from the reality of what is really in Google's index. Using another search engine, such as Exalead, is one solution, but then the problem of index coverage comes into play. Google has declared having seen 1 trillion unique URLs (in 2008) but estimates of the size of Google's index are about 50 billion pages, so some hidden choice has been made of what is in the index and what is not. This means that frequency based language models derived from search engines are only approximate.

Nonetheless, it is possible to make rough, relative judgments of how often one linguistic phenomenon appears with respect to another, and using probing can provide some information of the relative frequency of these phenomena. Over a long period, it is possible to generate and test a great number of possibilities, some examples of the usefulness of this technique are finding what words commonly occur with other words, what colors are often associated with nouns, what are the most common translation of multiword expressions, what are the most likely transliteration of English terminology and names into Japanese, for example.

The Web is not a uniform corpus, far from it. There are many different language registers even within one language: there are professionally edited well written articles, there are more colloquial blog posts, there are hastily written error-filled comments, all which generate different language models. One recent exploitation of user-generated content on the web has been the mining of opinions concerning some subject, or company, or product. Affect analysis is now a thriving market and a true commercial success for natural language processing. Many other areas of text mining remain to be explored. For example, the particular language used to tag photos in social media sites (such as Panoramio or Flickr) and reveal many things about the user (especially in conjunction with GPS and time data). This language is different from that found in the general web, or on Wikipedia. We can use it to find out the interesting things to visit in a city, we can predict where a tourist can go, we can even guess whether a user is a woman or a man, from their tagging behavior. Mining this information can lead to additional applications that exploit this new knowledge.