Invited Presentation

Repetition and Language Models and Comparable Corpora

Ken Church
Human Language Technology Center of Excellence
Johns Hopkins University
Kenneth.Church@jhu.edu

I will discuss a couple of non-standard features that I believe could be useful for working with comparable corpora. Dotplots have been used in biology to find interesting DNA sequences. Biology is interested in ordered matches, which show up as (possibly broken) diagonals in dotplots. Information Retrieval is more interested in unordered matches (*e.g.*, cosine similarity), which show up as squares in dotplots. Parallel corpora have both squares and diagonals multiplexed together. The diagonals tell us what is a translation of what, and the squares tell us what is in the same language. I would expect dotplots of comparable corpora would contain lots of diagonals and squares, though the diagonals would be shorter and more subtle in comparable corpora than in parallel corpora.

There is also an opportunity to take advantage of repetition in comparable corpora. Repetition is very common. Standard bag-of-word models in Information Retrieval do not attempt to model discourse structure such as given/new. The first mention in a news article (*e.g.*, "Manuel Noriega, former President of Panama") is different from subsequent mentions (e.g., "Noriega"). Adaptive language models were introduced in Speech Recognition to capture the fact that probabilities change or adapt. After we see the first mention, we should expect a subsequent mention. If the first mention has probability p, then under standard (bag-of-words) independence assumptions, two mentions ought to have probability p^2, but we find the probability is actually closer to $p/2$. Adaptation matters more for meaningful units of text. In Japanese, words (meaningful sequences of characters) are more likely to be repeated than fragments (meaningless sequences of characters from words that happen to be adjacent). In newswire, we find more adaptation for content words (proper nouns, technical terminology and good keywords for information retrieval), and less adaptation for function words, clichés and ordinary first names. There is more to meaning than frequency. Content words are not only low frequency, but likely to be repeated.

Proceedings of the 2nd Workshop on Building and Using Comparable Corpora, ACL-IJCNLP 2009, page 1,
Suntec, Singapore, 6 August 2009.