CSCI 8450 Reading 3

Mohan Sai Ambati

Web-based Models for Natural Language Processing *Mirella Lapata and Frank Keller, University of Edinburgh*

Summary:

Traditional written corpora are created from printed text such as news, articles and books. With the growth of world wide web as a data resource, it is increasingly used as a source in a wide range of natural language processing (NLP) tasks. Authors in their previous work Keller and Lapata's [2003] demonstrated that web counts can be used to approximate the bigram counts and suggested web based frequencies could be useful for a wide variety of NLP tasks. Since there is only limited number of tasks tested using web based frequencies, this paper investigates performance of several NLP tasks using unsupervised web based models. In this paper authors are concerned about following questions:

- Can Unsupervised web-based methods be a competitive alternative to supervised approaches used in literature?
- Can web based model (noisy, less sparse) and corpus based model (less noisy, sparse) be combined into a single model?

To address these questions authors needs a mechanism to get the web counts. In this paper authors followed their previous work Keller and Lapata [2003] web counts for n-grams using a simple heuristic based on queries to a web search engine (Google and Altavista). The web count is estimated as the number of hits returned by the search engine for queries generated by n-gram. For all tasks web based models are compared with estimated BNC [Burnard 1995] models. Development set and test set data is obtained from gold standard data set. In this paper, authors explored these four models (1) web based model (2) corpus based models (3) a back off model and (4) interpolated model. In back off model authors used a combination of web count and corpus count by using a simple threshold, if the n-gram count for an item in the corpus falls below a threshold θ, then web is used to estimate the frequency else corpus is used. In interpolated model authors used standard interpolation scheme uses combination of web and corpus as follows:

$$f = \lambda f(web) + (1 - \lambda) f(corpus)$$

Authors compared these models to each other to determine the usefulness of web counts for given task and these models are compared to valid base line models in the literature. The obstacles on the way to answer authors questions are (1) to show web counts are useful for wide variety of tasks it need to be applied to a diverse range of NLP tasks both syntactic and semantic, involving analysis and generation. (2) Show this approach scales up to larger n grams and combinations of different parts of speech. To overcome these obstacles authors selected tasks so that they cover both syntax and semantics, both generation and analysis, and wider range of n-grams and parts of speech that have been previous explored. The tasks authors selected are (a) target language candidate selection for machine translation (b) context-sensitive spelling correction (c)ordering of prenominal adjectives (d) compound noun bracketing (e) compound noun interpretation (f) noun countability detection (g)article restoration and (h) PP attachment disambiguation. For all tasks attempted, the web-based models significantly outperform baseline models except for generation and analysis tasks. For all generation tasks authors found web-based models outperform other models except in candidate selection task, where there was no difference between web based models and corpus based models. In analysis tasks authors fail to observe an advantage of web based model over BNC models since it involves decisions that are nor directly observable in data. By this tasks authors observe that unsupervised web models have access to more linguistic information which supervised models in literature lacks. And back off model and interpolation model sometimes showed a small performance gain than individual models. By observations authors argue that unsupervised web-based models should be used as baseline models.