Reading Report - Week 5 **Varad Meru**
CS 221 - Information Retrieval Student # 26648958
Prof. Cristina Lopes Due Date: 02/07/2015

The unreasonable effectiveness of data*: Summary

By Alon Halevy, Peter Norvig, and Fernando Pereira (Google).

Summary

The authors start with a historical perspective on how mathematics has enabled physics to model the world in an elegant and understandable manner, which has driven the science world for a long period of time. This elegance is the result of perfections being added to the model overtime an anomaly is seen in the previous model. The discussion then moves towards an example of *Brown Corpus*, which had a million English words, and the modern era trillion-word corpus by Google. The authors showcase the use and effectiveness of more data, even though it may have its own inherent flaws.

The discussion is then focused on text and the difficulties it brings with itself. The success of statistical speech recognition and statistical machine translation in the field of natural-language-related machine learning is due to the large training set of the input-output behavior. This is in contrast with the state of natural language processing problems such as document classification, part-of-speech tagging, etc. Memorization also helps in presence of large amount of training data. The use of n-gram models also help, by counting each of the n-gram sequences and computing probability distribution of probable future n-grams. An example of *scene completion* elaborated the application of a data-centric and data-driven system with section of a photo hidden and making the system decide the best-fit cropping of a photo for that section.The quality of the system grew as the data grew, with the results being poor for thousands of photos, but with millions, the quality improved dramatically. But this success of the n-gram model directs us to make false assumptions, with many people believing that there are two approaches to natural language processing: A *deep* approach relying on hand-coded grammars and ontologies, and *statistical* approach that relies on n-gram statistics. But this gives rise to three orthogonal problems: Choosing a representation language, encoding a model in that language, and performing inference on the model. Various ways have been devised to deal with each of the problems.

The discussion then takes an interesting turn, diving into the discussion of *Semantic Web* versus *Semantic Interpretation*. Semantic Web is a convention for formal representation languages for interaction between softwares. On the other hand, the problem of understanding human speech and writing - the semantic interpretation problem - is quite different from the problem of service interoperability. Building Semantic Web services is an engineering and sociological challenge and we must deal with significant hurdles - Ontology Writing, Difficulty of implementation, Competition, Inaccuracy and deception. The challenges for achieving accurate semantic interpretation are different. We have already solved the sociological and technological problems of creating content with semantic web framework and are only left with a scientific problem of interpreting the content. The authors then dive into some of the examples of the shortcomings of the standard *Semantic Web* framework in helping with *Semantic Interpretation* problem.

Commentary

The article gives a nice introduction on data-centric systems and how data is driving the applications of the modern era. This article also articulates the fallacies of the unstructured nature of web how it gives rise to the modern issues maintaining the semantic nature of the data.

*Halevy, A., Norvig, P., and Pereira, F., *The unreasonable effectiveness of data*, Intelligent Systems, IEEE 24.2 (2009): 8-12. Video at `https://www.youtube.com/watch?v=yvDCzhbjYWs`