Summary of Paper 4: - Ramkumar Rajabaskaran (85241493)

This paper talks about the unreasonable effectiveness of data, how with the availability of a big data set, it is possible to come to predict and process natural language, that without this data, will yield to long procedures of writing complex rules and semantics.

One of the authors had the chance to use a corpus at Brown University that contained one million English words, which at that time was a huge deal. Nowadays that has become 100X bigger and Google is now able to produce a trillion-word corpus with instant access to many n-grams as possible. So, this corpus could serve as the basis for a complete model for certain tasks if we can extract the model from the data.

Natural language related machine learning has been successful in speech recognition and statistical machine translation not because it is easier, it is in fact much harder, but it is successful as it is done every day unlike the language processing such as document classification and POS tagging that are seldom done. Hence, they have no large corpus available like the former. A corpus for these tasks requires skilled human annotation and is slow and expensive. Another dis-advantage is the lack of training data that is available for the former. The availability of huge training data as in the case of replacing a photo with pixels from a data set were shown to improve if the data set was larger.

Language is also inherently complex as every day new words are being coined, old usages modified and the set of grammatical rules is also complex. Rare word which are collectively frequent events are more worth than the commonly occurring words. Hence n-grams have been falsely believed to be successful mainly due to people believing that there are only two approaches, a deep approach that relies on hand coded grammar and a statistical approach that relies on n-grams from a large corpus, to natural language processing. But three problems arise due to this assumption, they are choosing a representation language, encoding a model in that language and finally performance inference on that model. Various approaches have been dealt to tackle these problems, such as using a finite state machine and Bayesian models.

Another topic that was discussed in this paper was the difference between the semantic web and semantic interpretation. Semantic Web is a convention for formal representation of languages that lets software services interact with each other, without needing artificial intelligence, whereas Semantic interference deals with the imprecise and ambiguous language where service interoperability deals with data precise enough that will make the function efficient. The Semantic web will allow machines to comprehend semantic documents and not human languages. Still it faces some challenges such as Ontology writing, Difficulty of implementation such as dynamic web pages with databases, competition and inaccuracy and deception by users. The challenges for the semantic interpretation are different, while two words are different, contextually they are the same. For example, “money” and “price” may mean the same depending on the context. Hence making a machine learning algorithm for context recognition is difficult. We can resolve this by extracting a set of schema from the corpus for attributes that rarely occur together. Another is to combine different tables of data from other sources.

My viewpoints: - We can thus see the effectiveness of having a big dataset. From the large dataset, we have a well-defined training base that we can use to adapt machine learning for natural languages through n-grams, statistics as well as to obtaining the set of schemas that we can define a pattern of speech and language to occur in. Hence contextual information can be derived from this data set. Having a big data set though doesn’t guarantee a proper usage of data unless it is properly referenced when used. It should also be regularly updated as per the daily trends and changes. Also, we must choose a model that allows unsupervised learning on unseen data that exists in a larger amount than labelled data. Hence proper storage and identification and retrieval of big data resources is a must for a successful POS tagging and Human Language processing