Identifying Coreferent Genotypes in One-Way Cryptographic Hash Using Haplotype Information

Chung-Ning Chang

July 26, 2017

1 Problem Statement and Motivation

We are given some datasets of allele data on a number of genes/regions, where each dataset may have different sets of allele fields (attributes) but all include those in BRCA1 and 2, and each record consists of alleles in that set and represents an individual. I.e., each dataset represents individuals with different attributes. Our goal is to identify individuals using the intersection of the attributes of each dataset. However, as dataset size grow, the collision rate gets higher, and individuals become less identifiable, using only those attributes. We want to find ways to increase the identifiability.

Since the genotype data available to us is not complete, we expect increasingly high collision rate as the dataset size grow. Hence, in addition to the given genotype data, we might also want to take into account other aspects of the individuals such as date of birth. More interestingly, we may be able to infer other genotype information using known haplotype information, in order to express the uncertainty in identifying coreferent entries. With an ultimate goal of performing locally differentially private word prediction in mobile devices using semantic and syntactic information, in the remainder of this article we present and discuss some of the work that has been done on the task of word prediction as well as how these could pave our way to the locally private solution.

2 Current Work

2.1 Word Prediction with Semantics

In [**?**], Pang et al. worked on a particular setting of word prediction, where the text that the user enters is a response to a piece of stimulus text, e.g., replying a text message, and hence there is context in the stimulus in addition to what the user is typing. They employed an n-gram model and mixed in other more sophisticated models such as selection model, where they give more weight to a uniformly randomly selected content word (a word that provides semantic content) from the stimulus, as it is a natural assumption that in order to be semantically coherent, a reply would often repeat certain content words from the stimulus. In this paper, they found that the in addition to the partially typed response, incorporating the stimulus text with a mixture model yielded the best result. This seems to be a closely related work to our application, as we situate our problem

in a context where there is likely a stimulus component.

In [**?**], Hyvonen et al. extended word prediction using several semantic concepts, such as equivalence relations (homonymy, polysemy, etc.), and indirect semantic relations (ontological relations, e.g., meronymy and holonymy). They aimed to complete user text input using labels of semantically similar concepts, which may be different at the surface, but nevertheless shown to be useful. The implementation in this paper would not be directly applicable to our goal, as we want to suggest a completion of the word the user is typing, instead of finding semantically similar but lexically different words.

In [**?**], Wolf et al. abstracted contextual information using topic models such as probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA). These models are used as they are shown to outperform the more canonical latent semantic analysis (LSA) model, and in this paper they indeed found LDA to perform better than the rest. The topic models may be interesting for our application, if we would be able to find the relevant subset of documents, e.g., the sequence of text messages in a dialogue, in which topics are to be modeled.

In [**?**], Ghosh et al. considered the Long Short-term Memory model, an extension of the recurrent neural network, for several NLP tasks including word prediction, and improved it by adding contextual features such as topics. It is shown that with various topic features the model outperforms the traditional LSTM model, and the improvement is reproducible on different datasets. This work is recent and shows us that LSTM would work well on word prediction. In a few other papers on machine translation (MT), which is a sequence-to-sequence task, it is seen that LSTM gives good results. Since our task could be formulated in a similar sequence-to-sequence way, it is not surprising that this model would perform well.

2.2 Word Prediction with Syntax

In [**?**], Fazly et al. incorporated syntactic information in the form of part-of-speech (POS) tags and designed two algorithms to achieve such incorporation: Tags and Words, where a conditional independence among tags and words is assumed, and Linear Combination, where they combine a word model and a tag model linearly. For both algorithms, they considered a trigram model for the tags and a bigram model for the words. Combining POS tags to a statistical baseline model seems to be effective approach and would be useful for us. However, we want to incorporate more structure than simply POS tags.

In [**?**], Wood et al. implemented a syntactic preprocessor to work on top of a statistical model, and they showed that syntactic preprocessing would always improve the keystroke savings, irrespective of the statistical model used. In their work they used data annotated with constituency parsing, where words in a sentence are given tags that give the sentence a tree representation and constituents (such as noun phrases and verb phrases) are subtrees in the tree. This work seems closely related to our projection of the task, as they utilize syntactic parse trees of the document in addition to statistical models such as n-grams. It would be nice if we would be able to find more recent work that uses syntactic parsing.

2.3 Word Prediction with Semantics and Syntax Combined

In [**?**], Wang et al. used a directed Markov Random Field (MRF) model to incorporate models for syntax, semantics, and lexical information. In particular, they modeled syntactic structures of sentences using Probabilistic Context Free Grammar (PCFG). Though in combination with the other models it became context sensitive, they were able to estimate the parameters using a generalized inside-outside algorithm. This is one of the few papers found that incorporated semantics, syntax, as well as lexical information and it seems very promising and relevant to our application. The only obstacle would be in implementation, since they used a rather complex combination of algorithms.

3 Conclusion

The papers and summaries presented in the previous section is only a small portion of work that has been done in the field of word prediction, with or without enhancement of syntactic or semantic information. However, to the extent I was able to find, there does not seem to be much work done using on-the-fly parsing of either syntactic or semantic structure for word prediction, even though there have been work in dynamic parsing of both areas. In most of works found, the use of syntactic information was in the form of POS tags, which do not preserve sentence-level information such as whether a noun phrase is the subject or object of the verb, and could affect the probability of words erroneously in some cases. More research has to be done to have a better understanding of the current state of works in this field. We can then find methods that could integrate in a straightforward way into an appropriate locally private learning algorithm.