Research Statement

Big data is ubiquitous today [**?**]. Many science and engineering applications in astronomy, biology, chemistry, physics and medicine are generating peta-and tera-bytes of data. With the evolution of large and complex data archives, machine learning algorithms need to be designed to extract patterns from them. This class of machine learning algorithms which can scale to massive databases are expected to rely significantly on well established techniques of parallelization and distributed computing and are called *large scale* learning algorithms.

Problem Statement: Our research primarily focuses on developing algorithms for large scale machine learning for text data. Of particular interest is noisy text generated from Optical Character Recognition devices.

The Library of Congress(LoC[1]) has initiated the digitization of newspapers from early 1800-to date[2]. Newspapers are the first draft of history – they are a rich source of information for historians, researchers, and scholars. With the advent of *digitized* newspapers, the accessibility to old historic papers has increased. The usability of archives storing these newspapers depends on the imaging technology, Optical Character Recognition (OCR) devices, zoning and segmentation, metadata extraction, search ability and web delivery systems developed to make them accessible. It is a well known fact that the OCR technology is far from perfect and often, the text generated is garbled affecting the efficiency of search and retrieval. Since the human mind excels in visual cognition and language processing tasks and machines are not able to completely recreate these skills, manual labor is involved in correcting garbled OCR. Thus building an automated OCR text correction tool is desirable – more importantly, such a tool should scale to very large databases since archives storing newspaper articles are usually very large.

1 Review

Golding and Roth [**?**] uses a Winnow-based algorithm for OCR Correction and achieve accuracy levels of 99% for 265 confusion sets. This approach was originally meant to correct predetermined words. It is unclear to what extent it could be scaled to correct all words in a dictionary.

[1] www.loc.gov
[2] 3

Another approach used in literature [**?**] is based on Hidden Markov Models (HMMs) which are trained on wikipedia edits and further augmented with perceptron re-ranking. This algorithm also is not able to scale to large textual databases and caused performance bottleneck. Structured learning is the umbrella term for supervised machine learning techniques that involve predicting structured objects, rather than single or real valued outputs.

Structured machine learning refers to learning structured hypotheses from data with rich internal structure usually in the form of one or more relations. In general, the data might include structured inputs as well as outputs, parts of which maybe missing, noisy or uncertain. Structured prediction models are typically trained by means of observed data in which the true prediction value is used to adjust model parameters. Structured Output Learning: f : X → Y. inputs X can be any kind of objects outputs y ? Y are complex (structured) objects images, parse trees, folds of a protein Due to the complexity of the model and the interrelations of predicted variables the process of prediction using a trained model and of training itself is often computationally infeasible and approximate inference and learning methods are used.

Structured learning is most likely to pay off in large domains, because in small ones it is often not too difficult to hand-engineer a good set of features. Structured prediction models mainly consist of probabilistic graphical models, Baysian Networks, random fields, inductive logic programming, structured SVMs, Markov logic networks and constrained conditional models.

2 Problem Statement

The research problems include the following:

- Developing OCR text correction algorithm which can be applied to a large data repository. Evaluate the performance of the algorithm against current industry standards.
- Since humans are much more adept at correcting OCR errors manually – how can we use annotations provided by humans to model corrections? How does the subjectivity of human annotated data affect the text correction process?

3 Scope

With the current trend in the growth of data and information, large-scale learning problems can not be overlooked. There is a need to develop algorithms that deal with these problems efficiently.
The objective of this research is to train a model on a large-scale data which can be used further to predict the errors in the unseen OCR text. This would help in building a clean online repository of historic newspapers providing rich source of information to historians, researchers, scholars and general users.

References

[1] Jeffrey D. Ullman Anand Rajaraman, Jure Leskovec. *Mining Of Massive Datasets*. Cambridge University Press, 2013.

[2] Andrew Carlson, Chad Cumby, Je Rosen, and Dan Roth. The snow learning architecture. Technical report, Technical Report UIUCDCS, 1999.

[3] Ibm Corporation. Bigdata: Why it matters to the midmarket. *Forward View: The magazine for mid-size business*, 2012.

[4] ThomasG. Dietterich, Pedro Domingos, Lise Getoor, Stephen Muggleton, and Prasad Tadepalli. Structured machine learning: the next ten years. *Machine Learning*, 73(1):3–23, 2008.

[5] AndrewR. Golding and Dan Roth. A winnow-based approach to context-sensitive spelling correction. *Machine Learning*, 34(1-3):107–130, 1999.

[6] Thorsten Joachims. Advances in kernel methods. chapter Making large-scale support vector machine learning practical, pages 169–184. MIT Press, Cambridge, MA, USA, 1999.

[7] Aditya Krishna Menon. Large-scale support vector machines: algorithms and theory. *Research Exam, University of California, San Diego*, 2009.

[8] Elif Yamangil and Rani Nelken. Scalable lexical correction from wikipedia edits using perceptron reranking. *unpublished*, 2008.