LSTM with Shakespeare

Brandon Ustaris
Department of Computer Science
University of California, San Diego
La Jolla, CA 92093
bustaris@ucsd.edu

Brian Preskitt
Department of Mathematics
University of California, San Diego
La Jolla, CA 92093
bpreskit@ucsd.edu

Brian Wilcox
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093
bpwilcox@ucsd.edu

Derek Tran
Department of Computer Science
University of California, San Diego
La Jolla, CA 92093
dtt018@ucsd.edu

John Clara
Department of Computer Science
University of California, San Diego
La Jolla, CA 92093
jclara@ucsd.edu

Abstract

In this paper, we propose using an LSTM neural network to perform supervised learning on English metrical poetry. First we walk through a background on metrical analysis. Then we explain the current metrical tools and research in scansion using machine learning. Finally we will describe the tools and data we plan to use in order to learn how to scan English metrical poetry. We will rely heavily on Agirrezabal1 et al's research in "Machine Learning for Metrical Analysis of English Poetry" 2016.

1 Background

Much of English poetry uses combinations of stressed and unstressed syllables to create rhythm in each line. Scansion is the process of parsing out the stressed and unstressed syllables in each line. Most English poetry starts with a base pattern or meter on which the poet then chooses to make variations. Consider the first line from Shakespeare's Sonnet 18:

/	/	x	/	x	x	x	/	x	/
Shall	I	com	pare	thee	to	a	sum	mer's	day

Here '/' signifies a stressed syllable and 'x' signifies an unstressed syllable. All of Shakespeare's Sonnets follow the iambic pentameter meter. The iambic pentameter line is comprised of 5 couplets of unstressed and stressed syllables: x / x / x / x / x /. However, the first line of Sonnet 18 deviates from this meter by using a stressed 'Shall' and unstressed 'to'. Due to these variations, metrical analysis of English poetry is a non-trivial task.

2 Current Tools and Research

A number of tools exist for metrical analysis based off of linguistic knowledge of scansion rules. *ZeuScansion* (Agirrezaball et al 2016) is the latest tool in this vein for English poetry.
Others have started to use machine learning to automatically parse metrical poetry into stressed and unstressed syllables. First, Estes and Hench 2016 used a CRF model to learn Middle High German epic poetry. MHG meters not only use a combination of stressed and unstressed syllables but also uses long and short syllables. They used 750 lines over 3 different texts as their training data as well as 75 lines from a separate text for their test data. Then they performed syllabification on the data as well as selecting features they thought were informative for scansion. They achieved an F score of .904 on their hold out set.
Inspired by Estes and Hench's good results, Agirrezaball et al 2016 tested a number of different machine learning techniques for English language poetry including Support Vector Machines, Perceptrons, Hidden Markov Models, and Conditional Random Fields. Again, they extracted features that they thought were informative for scansion. They achieved results significantly better than the *ZeuScansion* tool.
Graves and Schmidhuber 2005 used LSTMs and bidirectional LSTMs for phoneme classification. They did not have to do feature extraction and found that while bidirectional LSTMs performed better, unidirectional LSTMs also performed well on the phoneme classification problem. We believe that the phoneme classification problem is a harder version of the metrical analysis problem. If an LSTM can perform well on phoneme classification, it should also be able to perform well on metrical analysis.

3 Our Plan

We plan to perform the same metrical scansion task as Agirrezaball et al using an LSTM neural network. We will also use the same data set: For Better For Verse. This site contains a number of English language poems together with their metrical scansion easily available on github. We will read this data out using an XML parser as well and syllabify it using pyphen. We plan to use Keras' implementation of an LSTM neural network.
We will start out by using only a handful of the features used in the paper and add more features until we can achieve results equivalent to those Agirrezaball et al found using CRFs. We hope to achieve at least a 89.66% accuracy per syllable and 50.16 % accuracy per line. Agirrezabal et al 2016 stated that they began to work on LSTM and RNN models which required less feature extraction with equivalent or better results. Therefore, we are confident that we should be able to achieve almost as good results as their CRF model.

References

[1] Agirrezabal, Manex, Inaki Alegria, and Mans Hulden. "Machine Learning for Metrical Analysis of English Poetry." *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (2016): 772-81. Web. 3 Mar. 2017. http://aclweb.org/anthology/C16-1074.

[2] Estes, Alex, and Christopher Hench. "Proceedings of the Fifth Workshop on Computational Linguistics for Literature." *Supervised Machine Learning for Hybrid Meter* (2016): 1-8. Web. 3 Mar. 2017. http://www.aclweb.org/anthology/W16-0201.

[3] Graves, Alex, and Jrgen Schmidhuber. "Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures." *Neural Networks* 18.5-6 (2005): 602-10. Web.