Adding semantics to statistical generation for poetic creativity

Maximilian Droog Hayes and Geraint A. Wiggins
Computational Creativity Lab
School of Electronic Engineering and Computer Science
Queen Mary University of London
Mile End Road, London E1 4FZ, UK
m.drooghayes@se11.qmul.ac.uk,geraint.wiggins@qmul.ac.uk

Introduction

We present a preliminary study that provides some evidence for the IDyOT model of creative perceptual cognition (Wiggins, 2012; Wiggins and Forth, 2015) by using distributional semantics to influence a statistical model of surface form language. We apply this in the context of a haiku generator, which uses a random walk through the syllables of words to generate its poems. We compare two versions of the generator: one without semantic influence and one with, and evaluate in terms of human judgments of semantic coherence of the poetry. The semantic influence yielded poems that were significantly more likely ($p < .05$) to be judged coherent. The work also begins to address considerations to do with semantic content raised in previous poetry generation research (e.g., Gonçalo Oliveira et al., 2014; Toivanen, Gross, and Toivonen, 2014).

Background

Haiku

A Haiku is a three line Japanese poem having exactly seventeen syllables, five then seven and then five again. Haiku in English more or less follow the original form and style of Japanese Haiku, which often cover the subject of nature. The limited length and constraints on the number of syllables make Haiku a good aim for the computational generation of poetry. These features confine the form of the output while still admitting creativity. Various on-line Haiku generators exist[1], though not all produce outputs that are of the correct form, and most are template based, restricting the creativity possible.

Markov modelling of language

In this preliminary study, we use the very simplest form of Markov model, the *n-gram* model (Manning and Schütze, 1999). We used 3-grams, groups of 3 words drawn from an independent data source (see below). Given a starting word, it is then possible to stochastically generate sequences of words based on the distributions produced by the given context. Such techniques have proven effective in language processing in general, though the focus is usually on parsing and/or speech processing, rather than the generation of semantic coherence, as in our case.

Distributional Semantics

Distributional Semantics, in general, is an approach to the extraction of meaning from large bodies of linguistic data. The essential principle is that words with related meanings are more likely to co-occur in language than words with unrelated meanings. The inverse likelihood of co-occurrence is used as a distance in a multidimensional space, which may or may not be dimensionally reduced to avoid redundancy. Normalisation of the vectors so produced yields a subspace of the multidimensional space in which the points are projected onto the surface of a hypersphere centred on the origin; geodesic distances on the surface of this sphere (which can alternatively be thought of as the angle between vectors) give a measure of the semantic distance between words. It is important to understand that this method does not give *definitions* of words, but *clusters* of words with related meaning. In the current work, we used an open source library called *SemanticVectors*[2], which creates the semantic space but does not normalise or reduce it (though the number of dimensions is set in advance, and makes a substantial difference to the output). Given the space, one can query a word, and get a list of related words in response.

Data

We used an independent data source, the English One Million trigrams from the Corpus of Contemporary American English (COCA)[3].

Syllabification

Because the structure of haiku depends on syllables, we sought a syllabification library to enable our system to construct lines of the correct length. We were unable to find one that was consistently reliable, and so we constructed our own, which we believe to work better than all of the existing systems that we found. This will be reported elsewhere.

[1] http://www.languageisavirus.com/interactive-haiku-generator.html; http://www.everypoet.com/haiku/; http://www.smalltime.com/haiku.html.

[2] https://code.google.com/p/semanticvectors/

[3] corpus.byu.edu/coca/

The System

The system generates from either a given or a randomly chosen word, using the trigrams of COCA in the conventional Markovian way (Manning and Schütze, 1999). As each new word is selected, by roulette wheel selection from the distribution given in the Markov model by the previous word, it is decomposed into syllables, which are counted. The process continues until the required pattern of 5, 7, and 5 is reached; hyphenation is not allowed, so words do not span line breaks. In the event that a string of 17 syllables, broken up this way, is not found, the generation fails; an obvious and probably helpful future addition would be the ability to backtrack during this process.

The two versions of our system are differentiated by the formation of the distribution from which words are chosen. In the statistical syntax only version, v1, the Markov model generated from the COCA trigrams is used directly. In the version that uses statistical syntax and semantics, v2, these same distributions are biased towards words in the top ten results generated by the Semantic Vectors similarity space. A parameter allows control of the amount of bias; too much bias causes discontinuity in the predictions of the Markov model (by effectively ruling out intermediate words). If the probability of a word in the top ten list is is zero after the current word, then it is added with low probability to the distribution before the random walk proceeds.

The overall effect of the semantic bias in v2 is to favour words that the semantic space clusters around the initial word. Because the semantic space is sometimes idiosyncratic, this can result in entertaining word play, as well as more coherent haiku.

Evaluation

To evaluate the effect of the extra semantic information, we asked human readers to choose between pairs of haiku, one of which was generated by v1, and one of which was generated by v2. There were 18 participants, and we used a balanced block forced choice design. The participants were asked to choose the haiku that was the more meaningful. The results are shown in Fig. 1. The haiku were based on the words shown in the figure, with the first word always being the keyword. The haiku used for the study were chosen at random, and not curated. Our null hypothesis was that there would be no reliable, perceptible influence of the semantic generation on the haiku.

More participants selected the semantically generated haiku in 8 of the 10 pairs, and one haiku produced a tie. A two-tailed T-test over the results gave $p = .0488$; thus, the participants were significantly more likely to identify the semantically generated haiku as more meaningful ($p < .05$), demonstrating that our null hypothesis was false. We can conclude that this method is capable of generating more meaningful poems than mere syntactic generation.

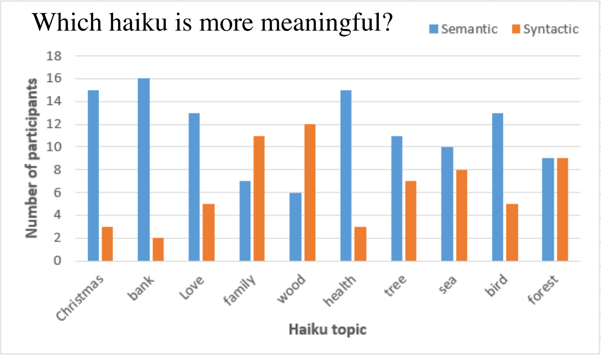

Figure 1: Experimental results: number of participants choosing semantically- or syntactically-generated haiku in response to the question "Which haiku is more meaningful?"

Conclusion

We conclude that learned semantics can be used to statistically bias generated text to be more meaningful. As a demonstration of this, the following two haiku are generated from the word "Kentucky", the left with statistical syntax only, and the right with semantics. We have selected this example for its humorous quality.

Syntax only	**Syntax and Semantics**
Kentucky is free	Kentucky Derby
in an insect resistance	in agriculture in sin
of economic	of Evil Fried Chicken

Acknowledgments

The second author was funded by the project ConCreTe, which acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733.

References

Gonçalo Oliveira, H.; Hervás, R.; Díaz, A.; and Gervás, P. 2014. Adapting a generic platform for poetry generation to produce spanish poems. In *5th International Conference on Computational Creativity*.

Manning, C. D., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Toivanen, J. M.; Gross, O.; and Toivonen, H. 2014. The officer is taller than you, who race yourself! using document specific word associations in poetry generation. In *Proceedings of ICCC 2014*.

Wiggins, G. A., and Forth, J. C. 2015. IDyOT: A computational theory of creativity as everyday reasoning from learned information. In *Computational Creativity Research: Towards Creative Machines*, Atlantis Thinking Machines. Atlantis/Springer. In preparation.

Wiggins, G. A. 2012. The mind's chorus: Creativity before consciousness. *Cognitive Computation* 4(3):306–319.