Crowdsourced Social Media Poetry

Carolyn Elizabeth Lamb
David R. Cheriton School of Computer Science, University of Waterloo
200 University Avenue West
Waterloo, Ontario, Canada N2L 3G1

Abstract

We describe our program of PhD work in which a computer program creates topical poems out of text found on Twitter. These poems are made using a combination of natural language processing and crowdsourcing and are part of a general research plan involving the creation and evaluation of computer-generated poetry, grounded in domain-specific research on human creativity.

Studies of Poetry Evaluation

A key issue in all computational creativity is evaluation (Jordano 2011). It is important to us that evaluations be falsifiable, evidence-based, and grounded on an interdisciplinary level–drawing both from the psychology of creativity and from domain-specific knowledge about the art, science, or craft being generated by the computer. Computational creativity often suffers from methodological confusion in which evaluations fail to meet these standards—for example, evaluating through a rhetorical argument by the authors (Jordano 2011; Pearce, Meredith, and Wiggins 2002).

To help remedy this, we performed a thorough interdisciplinary survey which also served as the depth requirement of my PhD. The survey compiles information from psychology and philosophy which is applicable to the measurement of creativity, and discusses its possible applicability, or lack thereof, to computational creativity. In updated and modified form, it is now out on submission to relevant journals, as is an expanded version of a smaller published survey of techniques used in generative poetry (Lamb, Brown, and Clarke 2016b).

We performed two of our own experiments, specific to the domain of poetry. First, by studying non-expert judges’ responses to human-generated poems, we showed that expert and non-expert preferences in this domain can oppose each other, and that popular metrics for evaluating computational creativity fail to remedy this (Lamb, Brown, and Clarke 2015a). This study highlights the need either to use expert judges for creativity, or to separate experts from non-experts and be clear about which group is the intended audience.

Second, by studying expert judges’ responses to computer-generated poems, we developed our own domain-specific criteria for assessing generative poems (Lamb, Brown, and Clarke 2016a). Our criteria are as follows:

- **Reaction:** The poem should evoke feelings of enjoyment and/or interest.
- **Meaning:** The poem should intentionally convey a specific idea. Even if the poem is difficult to understand, its difficulty should enhance the underlying meaning.
- **Novelty:** The poem should be unusual or surprising in some way.
• Craft: The poem should make effective use of poetic techniques in service of the other three criteria.

**TwitSonnet and Future Work**

One can only do so much with evaluation if there is no system to evaluate. Our system, TwitSonnet, is a found poetry system based on Twitter. The idea of TwitSonnet is to construct entertaining poetic summaries of news and other trending events by compiling poems based on rhyming, topical tweets. Twitter found poetry has been previously performed by Pentametron (Bhatnagar 2012) and Mobtwit (Hartlová and Nack 2013), but Pentametron creates poems based only on rhyme/meter and without a topic, while Mobtwit focuses on putting tweets together based on location rather than a specific topic.

TwitSonnet uses a step-by-step process: harvest tweets based on time, keyword, and number of syllables; detect sets of appropriately rhyming tweets; rate the harvested tweets based on various desired criteria; and put the highest rated tweets together into a sonnet, or, if not enough rhyming tweets are available, a shorter poem.

In 2015, we built a TwitSonnet prototype which crowdsourced the ratings (Lamb, Brown, and Clarke 2015b). Crowdflower workers rated the tweets based on topicality, sentiment, and imagery. TwitSonnet put the rated tweets together, and we then evaluated the poems. Poems which were strongly topical and had positive sentiment were given a significantly better reception than baseline poems (those for which only rhyme and syllable criteria had been used, and not line ratings).

Later we expanded TwitSonnet into a version which could perform its own tweet rating, and thus avoid the need to hire crowd workers for each poem, while also giving more creative responsibility (and therefore, perhaps, more creativity) to the system itself (Lamb, Brown, and Clarke 2017). This time the tweet ratings were based on our set of four poetic criteria (Lamb, Brown, and Clarke 2016a). We posted the results on http://twitsonnet.tumblr.com. Unfortunately, this time the poems which were created using automatic line rating did not do better in the evaluation than the baseline poems. In other words, the automated version of the tweet ratings was not adding any poetic value, and was not useful in its current form.

A possible future avenue of work would be to increase the sophistication of TwitSonnet’s line rating techniques. However, we are planning to go the opposite route and to find new applications for the crowdsourced version of the system. There is an established tradition of making found poetry out of more substantive parts of news items, such as Hart Seely’s “The Poetry of D.H. Rumsfield” (Seely 2003)—as opposed to Twitter commentary and users’ arguments with each other. We feel that in the present political climate, a system like this, with a text source other than Twitter, would provide more focused and relevant news poems, and we are interested in using the crowdsourced version of TwitSonnet to mine especially poignant or relevant lines from substantive public documents regarding current events.

**References**


