WIKISONNET PROCEDURAL POEM GENERATOR

SAM TARAKAJIAN
Girlfriends, Brooklyn, NY, USA
starakajian@gmail.com

CASSIE TARAKAJIAN
Girlfriends, Brooklyn, NY, USA
ctarakajian@gmail.com

ANA GIRALDO-WINGLER
Girlfriends, Brooklyn, NY, USA
agiraldow@gmail.com

Keywords
Generative poetry
Natural language processing
Wikipedia

We present Wikisonnet, a procedural poem generator using text drawn from Wikipedia (Wikipedia 2016). Occasionally, without the author's intent, a string of words in a Wikipedia article will follow iambic pentameter. Wikisonnet extracts these, storing them along with rhyme and grammar information. Then it stitches them together, composing poems on the fly. To focus the poem on a particular subject, the algorithm can favor lines from a starting page, from pages linked to that page, or from pages similar to the starting page. Sometimes nonsensical, sometimes full of surprising juxtaposition, Wikisonnet poems each have several authors—like Wikipedia itself. But unlike Wikipedia, the poem as a whole is a product of the algorithm. A Wikisonnet poem is therefore a collaboration between human and machine sources, where the reader ascribes meaning to the end result through interpretation.

2016.
xCoAx
.org
Computation
Communication
Aesthetics
& X
Bergamo, Italy

Fig. 1. Wikisonnet as an installation https://s3.amazonaws.com/wikisonnet/Wikisonnet-2.mov

INTRODUCTION

The move from print to web transformed the written word from static text to dynamic hypertext. In the words of Kenneth Goldsmith, a contemporary American poet, “language, once ‘locked onto a page,’ has become ‘completely fluid.’”(Perloff 2010) Generative poetry grows naturally out of this language pool. Through it, we investigate possibilities for recombining text, examining the impact of medium on meaning. Through the resulting juxtaposition we re-encounter familiar language in an unfamiliar context.

Wikisonnet attempts to create such an encounter with Wikipedia. While reading an article on Wikipedia, it’s easy to forget that millions of individual authors have contributed to Wikipedia as a whole. Occasionally, one of those authors writes a revision in accidental prose—a few words in iambic pentameter. Wikisonnet seizes this text, extracts it from the original source, and grafts many such samples together into an Elizabethan sonnet. The result, whether nonsensical, satirical, or eerily poignant, excavates the language itself from the information it conveys, offering the reader an unexpected opportunity to revisit the latent poetry of Wikipedia.

RELATED WORK

The line between author and curator started to blur in the twentieth century, as authors experimented with reducing personal contribution to their work. Marjorie Perloff traces this development in *Unoriginal Genius: Poetry by Other Means in the New Century* (Perloff 2010). With its extensive interpolation of source material, T.S. Eliot's "The Waste Land" shows the earliest rumblings of this transition (Eliot and Vendler 1998). Though rightly celebrated, Eliot's masterpiece did meet with some critical rebuke, or at least confusion, for its long passages of apparently undigested citation. But as Perloff describes, this poetic collage represented a deliberate attempt to effect "coordination rather than subordination, likeness and difference rather than logic or sequence."

Later in the twentieth century, Raymond Queneau (a founding member of the experimental writing collective *Oulipo*, the *Ouvroir de littérature potentielle*) would write *Cent Mille Milliards de Poèmes*(Queneau 1982). Completed in 1961, the work contains ten sonnets, each printed on fourteen separate, moveable cards. Since each line has the same rhyme sound, they are interchangeable, allowing for 1014 possible poems. In *The Wasteland*, Eliot lets the reader interpret each citation; Queneau continues the trajectory: in allowing the reader to pick and choose particular lines, he invites her into the role of author.

In *Moveable Type*, Ben Rubin and Mark Hansen offer a digital update to *Cent Mille Milliards de Poèmes*, extracting and re-assembling digital text into poetic form (Rubin and Hansen 2007). The piece extracts text from *The New York Times*, either citing content directly or pulling from live web searches related to *The Times*. Then, one line at a time, it displays the results in the lobby of the newspaper's headquarters. Pulling together data from distant sources, *Moveable Type*'s audience gets a holistic feel for a complex system that otherwise escapes human sensitivity. Rubin: "We want it to feel almost like an organism that is living and breathing and consuming the news." (Kennedy 2007)

Today, authors and readers are moving to the Internet, where dynamic content enables new poetic encounters. The massive databases that power Google are making their way into the hands of computational poets, who use algorithms to scan vast stores of speech and text. Queneau asked the reader to arrange pre-written text, but algorithmic authors can work so fast that the reader can create poems on demand according to his or her specifications. "The US political poetry generator" (Unknown 2016) lets the reader choose a US politician along with a well-known poet, then generates poems combining text from the two sources. As is the case with Wikisonnet, the original text is unaltered, only

excised from its source and spliced together with other fragments. Interestingly, the early 20th-century desire to remove the author's voice from the final poem may here be realized. The reader chooses a politician and a poet, their words appear on the page, but the author who enables the entire experience has been reduced to the role of an engineer, a technician.

METHOD

Wikisonnet arranges Wikipedia into an Elizabethan sonnet, a fourteen-line poetic form with its own particular rhyme scheme. Each line conforms to iambic pentameter, consisting of ten alternating stressed and unstressed syllables. To compose a poem, Wikisonnet proceeds in two phases: a scraping phase and an authoring phase.

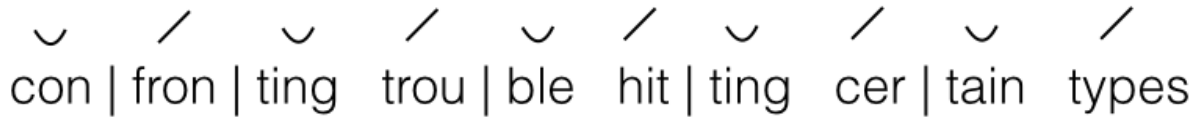


Fig. 2. An example of words in iambic pentameter. The half circle above the syllable denotes a weak accent, and the slash a strong one.

Fig. 3. The same example text block, with part of speech labels added to the words that will be stored by the scraping algorithm. Note that only the bold text is in iambic pentameter. The two words preceding and following the iambic pentameter text are analyzed and their parts of speech stored, but the words themselves are ignored.

SCRAPING PHASE

In the scraping phase, Wikisonnet downloads Wikipedia's monthly XML dump, the most current revision the Wikimedia Foundation provides (Wikimedia 2016). Next, using the *Textblob* package (Loria 2015), it parses each article, scanning for consecutive words in iambic pentameter. The sentence, "He finished his rookie season with a .255 batting average, despite confronting trouble hitting certain types of pitches," from professional baseball player Robert Clemente's Wikipedia article, returns "confronting trouble hitting certain types."

When the algorithm finds a block of interest, it analyzes its grammatical structure. The *Pattern* package (De Smedt and Daelemans 2012) for Python constructs a parse tree for the sentence, labeling each word with a part of speech tag. The algorithm, however, only stores the part of speech tag for the first two words of the iambic pentameter text block, the last two words, and the two words immediately preceding and following the text block.

Finally, the algorithm stores the last word in the text block, along with its rhyme class. It then uses the *NLTK* (Natural Language Tool Kit)'s pronunciation dictionary to determine rhyme(Loper and Bird 2002).

AUTHORING PHASE

In the authoring phase, the algorithm pieces lines together to write a poem, which must: 1) be a sonnet, and 2) satisfy the constraints of English grammar. To focus the poem's content, the algorithm starts with a given Wikipedia article. Once finished scanning this page, it moves to related ones, determined using Latent Dirichlet Allocation (Řehůřek and Sojka 2010).

To generate grammar constraints, the algorithm uses a part of speech "seam matching" technique, aligning the overhanging parts of speech from one fragment with those from the next. This helps Wikisonnet write poems with speed, a high degree of variety, and little sacrifice in grammatical accuracy.

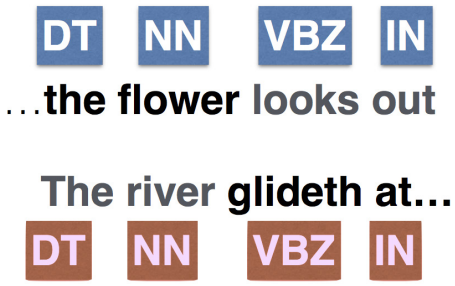
Fig. 4. Rows from the MySQL table storing blocks of iambic pentameter text scraped from Wikipedia articles.

id	page_id	word	rhyme_part	line	starts	ends	pos_m2	pos_m1	pos_0	pos_1	pos_len_m2	pos_len_m1	pos_len	pos_len_p1
1	25041	could	UHD	convergence, meaning that good pilots could	0	0	IN	NN	NN	VBG	NNS	MD	VB	RB
2	25041	steel	IYL	to make extensive use of stainless steel	0	0	JJ	NN	TO	VB	JJ	NN	CC	DT
3	25041	air	EHR	the basis of the record flight, the Air	0	0	RB	IN	DT	NN	DT	NN	NN	VBD
4	25041	those	OWZ	exert tremendous leverage under those	0	0	NN	MD	VB	JJ	IN	DT	NULL	NULL
5	25041	mass	AES	configurations of external mass	0	0	NULL	JJ	NNS	IN	JJ	NN	NNS	VBD
6	25041	plant	AENT	or to the new Lockheed assembly plant	0	0	NN	NN	CC	TO	NN	NN	NN	NN

Fig. 5. Choosing the next line in the poem. **5a** The first line, "We watch the bud of promise; and the flower" ends with the parts of speech DT and NN; and the two words following the iambic text – "looks out" – have the parts of speech VBZ and IN. **5b** To continue the poem, the algorithm looks for a line that begins with the same parts of speech as ended the previous line. **5c** The parts of speech at the "seam" of the previous and next line. When such a line is found, it becomes the next line of the poem **5d**.

(a) Line of text, with parts of speech labeled

(b) A candidate poem continuation



(c) The seam between two lines

(d) The completed continuation

EXAMPLES

HAMBURGER

In other places in the country there
in the United States, the Middle East
or pewter with the help of spoons or bare
unwanted side reactions are decreased.

The motor is supplied directly from
the side including mustard, mayonnaise,
explorer, author and inventor, some
supplies are meant to last for several days.

Around this time, Sukarno had begun
to aid in eating sauce in French cuisine
or just the patties served without a bun,
that is related to the kidney bean.

Donations and affiliation fees
in the United States and overseas.

JOHN CLEESE

Among the most important is the fact
A. Crockett, Jr. “Henry Louis Gates
his overall objection toward abstract
expatriates in the United States

It is considered scripture, classified
in British advertisements for Compaq
enthusiasts; it could be used to guide
the publication of his almanac.

That there is a substantial likelihood
of Doctor in the House (and later Cleese
did not object to starting the statehood
for making war to forge a lasting peace.

“The Universal Language” skit from All
about postponing love until the fall).

REFERENCES

De Smedt, T. and Daelemans, W. 2012. "Pattern for Python." *Journal of Machine Learning Research* 13:2031?2035.

Eliot, T.S. and H. Vendler. 1998. *The Waste Land and Other Poems*. Signet classic. Signet Classic.

Kennedy, R. 2007. "News Flows, Consciousness Streams: The Headwaters of a River of Words." *The New York Times*.

Loper, Edward and Bird, Steven. 2002. "NLTK: The Natural Language Toolkit." *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 63–70.

Loria, S. 2015. Textblob. https://textblob.readthedocs.org/en/dev/.

Perloff, M. 2010. *Unoriginal Genius: Poetry by Other Means in the New Century*. Chicago, IL, USA: The University of Chicago Press.

Queneau, R. 1982. *Cent Mille Milliards de Poèmes*. Paris, France: Gallimand.

Řehůřek, Radim and Sojka, Petr. 2010, May. "Software Framework for Topic Modelling with Large Corpora." *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 45–50. http://is.muni.cz/publication/884893/en.

Rubin, B. and Hansen, M. 2007. Moveable Type.

Unknown. 2016. The US Political Poetry Generator. http://www.elmcip.net/creative-work/us-political-poetry-generator.

Wikimedia. 2016. Wikimedia Foundation. https://dumps.wikimedia.org/.

Wikipedia. 2016. Wikipedia. http://www.wikipedia.org.