

The Officer Is Taller Than You, Who Race Yourself!

Using Document Specific Word Associations in Poetry Generation

Jukka M. Toivanen, Oskar Gross, Hannu Toivonen

Department of Computer Science and Helsinki Institute for Information Technology HIIT
University of Helsinki, Finland
jukka.toivanen@cs.helsinki.fi, oskar.gross@cs.helsinki.fi, hannu.toivonen@cs.helsinki.fi

Abstract

We propose a method for automatic poetry composition with a given document as inspiration. The poems generated are not limited to the topic of the document. They expand the topic or even put it in a new light. This capability is enabled by first detecting significant word associations that are unique to the document and then using them as the key lexicon for poetry composition.

Introduction

This paper presents an approach for generating poetry with a specific document serving as a source of inspiration. The work is based on the corpus-based poetry composition method proposed by Toivanen et al. (2012) which uses text mining and word replacement in existing texts to produce new poems. We extend that approach by using a specific news story to provide replacement words to the automatic poetry composition system. New contributions of this work are in constructing a model of document-specific word associations and using these associations to generate poetry in such a way that a single generated poem is always based on a single document, such as a news story.

The method for finding document-specific word associations is based on contrasting them to general word associations. In a given document, some of the document's word associations are long-established and hence well-known links which are part of people's commonsense knowledge, whereas some are new links, brought in by the document. Especially in the case of news stories, these links are exactly the new information the document focuses on, and they can be used in a poetry generation system to produce poems that loosely reflect the topic and content of the specific document. However, the story or message of the document is not directly conveyed by the produced poem as the process of poetry composition is based on the use of word associations. Thus, the generated poetry is roughly about the same topic as the document but it does not contain the actual content of the document. Poetry composed with these word associations may evoke fresh mental images and viewpoints that are related to the document but not exactly contained in it.

The general goal of this work on poetry generation is to develop maximally unsupervised methods to produce poetry

out of given documents. Thus, we want to keep manually crafted linguistic and poetry domain knowledge at minimum in order to increase the flexibility and language independence of the approach.

The next sections present briefly related work on poetry generation, introduce the method of constructing document-specific associations called here foreground associations and outline the procedure of using these associations in a poetry generation system. We also present some examples produced by the method and outline directions for future work.

Related Work

Poetry generation Several different approaches have been proposed for the task of automated poetry composition (Manurung, Ritchie, and Thompson 2000; Gervás 2001; Manurung 2003; Diaz-Agudo, Gervás, and González-Calero 2002; Wong and Chun 2008; Netzer et al. 2009; Colton, Goodwin, and Veale 2012; Toivanen et al. 2012; Toivanen, Järvisalo, and Toivonen 2013). A thorough review of the proposed methods and systems is not in the scope of this paper but, for instance, Colton et al. (2012) provide a good overview.

The approach of this paper is based on the work by Toivanen et al. (2012). They have proposed a method where a template is extracted randomly from a given corpus and words in the template are substituted by words related to a given topic. In this approach the semantic coherence of new poems is achieved by using semantically connected words in the substitution. In contrast to that work, we use document-specific word associations as substitute words to make the new poems around specific stories. Toivanen et al. (2013) have also extended their previous work by using constraint-programming methods in order to handle rhyming, alliteration, and other poetic devices.

Creating poetry from news stories was also proposed by Colton et al. (Colton, Goodwin, and Veale 2012). Their method generates poetry by filling in user-designed templates with text extracted from news stories.

Word association analysis There is a vast number of different methods for co-occurrence analysis. In our work we have been careful not to fall into developing hand-tailored

methods, but to use more general approaches (i.e. statistics), which could be applied to all languages in which different words are detectable in text. Most prominent statistical methods for word co-occurrence analysis are log-likelihood ratio (Dunning 1993), Latent Semantic Analysis (Deerwester et al. 1990), Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) and Pointwise Mutual Information (Church and Hanks 1990; Bouma 2009).

In this work we build on the background association calculation method proposed by Gross et al. (2012) and its recent extension to document specific associations (Gross, Doucet, and Toivonen 2014). We will describe these models in some detail in the next section.

What is Important in a News Story?

To produce a poem from a given news story, we first identify the essential features of its contents. News stories are normally summarized by their headlines, leads, topics, or keywords. For producing a poem, we are less interested in readily written descriptions such as the title and the lead, but more in text fragments such as keywords that we can use in poetry production. This also makes the approach more generic and not limited to just news stories.

Instead of keywords or topics, we propose to search for *pairs of associated words* in the document, as in Gross et al. (2014). The rationale is that often the core of the news content can be better summarized by the links the story establishes e.g. between persons, events, acts etc.

For illustration we use a BBC newspaper article on Justin Bieber drinking and driving on the streets of Miami, published on January 24, 2014¹. As an example, consider the sentence *"Pop star Justin Bieber has appeared before a Miami court accused of driving under the influence of alcohol, marijuana and prescription drugs."* The associations which are rather common in this sentence are, e.g. "pop" and "star", "justin" and "bieber", "miami" and "court" – words which we know are related and which we would think of as common knowledge. The interesting associations in this sentence could be "bieber" and "alcohol", "bieber" and "prescription", "justin" and "alcohol" and so on.

We model the problem of discovering interesting associations in a document as novelty detection, trying to answer the questions "Which word pairs are novel in this document?" In order to judge novelty, we need a reference of commonness. We do this by contrasting the given *foreground* document to a set of documents in some *background corpus*. The idea is that any associations discovered in the document that also hold in the background corpus are not novel and are thus ignored. We next present a statistical method for extracting document-specific word associations.

We use the log-likelihood ratio (LLR) to measure document-specific word associations. LLR is a standard method for finding general associations between words (Dunning 1993). In our previous work, we have used it to build a weak semantic network of words for use in computational creativity tasks (Gross et al. 2012; Toivonen et al.

2013; Huovelin et al. 2013). In contrast to that work, here we look for deviations from the normal associations. This approach, outlined below, seems to be powerful in catching document specific information since it has been used as a central component in a successful document summarization method (Gross, Doucet, and Toivonen 2014).

We count co-occurrences of words which appear together in the same sentence. We do this both for the background corpus and the foreground document. Using LLR, we measure the difference in the relative co-occurrence frequencies. More specifically, the test compares two likelihoods for the observed frequencies: one (the null model) assumes that the probability of co-occurrence is the same as in the background corpus, the other (the alternative model) is the maximum likelihood model, i.e., it assumes that the probabilities are the same as the observed relative frequencies. We will next describe the way to calculate document specific association strengths in more detail.

Counting Co-Occurrences

Consider two words w_1 and w_2 which appear in the document. We denote the number of times w_1 and w_2 appear together in a same sentence by k_{11} . The number of sentences in which w_1 appears without w_2 is denoted by k_{12} , and for w_2 without w_1 by k_{21} . The number of sentences in which neither of them occurs is denoted by k_{22} . In a similar way, we denote the counts of co-occurrences of words w_1 and w_2 in the background corpus by k'_{ij} (cf. Table 1).

Foreground Counts			Background Counts		
	w_1	$\neg w_1$		w_1	$\neg w_1$
w_2	k_{11}	k_{12}	w_2	k'_{11}	k'_{12}
$\neg w_2$	k_{21}	k_{22}	$\neg w_2$	k'_{21}	k'_{22}

Table 1: The foreground and background contingency tables for words w_1 and w_2 .

Probabilities

We use a multinomial model for co-occurrences of words w_1 and w_2 . In the model, each of the four possible combinations (w_1 and w_2 vs. w_1 alone vs. w_2 alone vs. neither one) has its own probability. In effect, we will normalize the values in the contingency tables of Table 1 into probabilities. These probabilities are denoted by p_{ij} such that $p_{11} + p_{12} + p_{21} + p_{22} = 1$.

Let $m = k_{11} + k_{12} + k_{21} + k_{22}$ be the number of sentences in the foreground document. The values of the parameters can then be estimated directly from the document as $p_{ij} = \frac{k_{ij}}{m}$. The respective parameters can also be estimated from the background corpus. Let m' be the number of sentences in the background, and let q_{ij} be the parameters (instead of p_{ij}) of the multinomial model; then $q_{ij} = \frac{k'_{ij}}{m'}$.

Next we will use these probabilities in likelihood calculations.

¹<http://www.bbc.co.uk/news/world-us-canada-25863200>

Log-Likelihood Ratio

To contrast the foreground document to the background corpus, we will compare the likelihoods of the counts k_{ij} in the foreground and background models. The foreground model is the maximum likelihood model for those counts, so the background model can never be better. The question is if there is a big difference between the models.

Let $P = \{p_{ij}\}$ and $Q = \{q_{ij}\}$ be the parameters of the two multinomial probability models, and let $K = \{k_{ij}\}$ be the observed counts in the document. Then, let $L(P, K)$ denote the likelihood of the counts under the foreground model, and let $L(Q, K)$ be their likelihood under the background model:

$$L(P, K) = \binom{k_{11} + k_{12} + k_{21} + k_{22}}{k_{11}, k_{12}, k_{21}, k_{22}} p_{11}^{k_{11}} p_{12}^{k_{12}} p_{21}^{k_{21}} p_{22}^{k_{22}}$$

$$L(Q, K) = \binom{k_{11} + k_{12} + k_{21} + k_{22}}{k_{11}, k_{12}, k_{21}, k_{22}} q_{11}^{k_{11}} q_{12}^{k_{12}} q_{21}^{k_{21}} q_{22}^{k_{22}}.$$

For contrasting the foreground to the background we compute the ratio between the likelihoods under the two models:

$$\lambda = \frac{L(Q, K)}{L(P, K)}. \quad (1)$$

The log-likelihood ratio test D is then defined as

$$D = -2 \log \lambda. \quad (2)$$

Given our multinomial models, the multinomial coefficients cancel out so the log-likelihood ratio becomes

$$D = -2 \log \left(\frac{q_{11}^{k_{11}} q_{12}^{k_{12}} q_{21}^{k_{21}} q_{22}^{k_{22}}}{p_{11}^{k_{11}} p_{12}^{k_{12}} p_{21}^{k_{21}} p_{22}^{k_{22}}} \right), \quad (3)$$

which after further simplification equals

$$D = 2 \sum_{i=1}^2 \sum_{j=1}^2 k_{ij} (\log(p_{ij}) - \log(q_{ij})).$$

The likelihood ratio test now gives higher values for word pairs whose co-occurrence distribution in the document deviates more from the background corpus.

For improved statistical robustness, we include the respective document in the background model, and in the case that the pair only co-exists in the document we estimate their joint co-occurrence probability under the assumption that the words are mutually independent. For more details, see Gross et al. (2014) who refer to these models as a Mixture model and an Independence model.

Given a document, we can now compute the above likelihood ratios for all pairs of words in the document. For poetry composition, we then pick from each document word pairs with the highest likelihood ratios and with $p_{11} > q_{11}$ to find the most exceptionally frequent pairs.

Poetry Composition

We compose poetry using a word substitution method as described by Toivanen et al. (2012). Instead of explicitly representing a generative grammar of the output language or

manually designing templates, the method copies a concrete instance from an existing text (of poetry) and substitute most of its contents by new words. One word of the original text is replaced at a time with a new, compatible word. In this method, compatibility is determined by syntactic similarity of the original and substitute word. Depending on the language, this requires varying degrees of syntactical and morphological analysis and adaptation. For more details on this part, see Toivanen et al. (2012).

In the current method, in contrast to the previous work outlined above, the topics and semantic coherence of the generated poetry are controlled by using the foreground associations. The document-specific foreground associations are used to provide semantically interconnected words for the content of a single poem. These words reflect the document in question but do not convey the actual content of the document. The idea is to produce poetry that evokes fresh mental images and thoughts which are loosely connected to the original document. Thus, the aimed style of the poetry is closely related to the imagist movement in the early 20th-century poetry which emphasised mental imagery as an essence of poetry. In the reported experiments, the corpus from which templates were taken contained mostly Imagist poetry from the Project Gutenberg.²

Examples

Following is an excerpt of the previously introduced BBC news story which we used for generating poems.

Justin Bieber on Miami drink-drive charge after 'road racing'

Pop star Justin Bieber has appeared before a Miami court accused of driving under the influence of alcohol, marijuana and prescription drugs. Police said the Canadian was arrested early on Thursday after racing his sports car on a Miami Beach street. They said he did not co-operate when pulled over and also charged him with resisting arrest without violence and having an expired driving licence. (...)

The article then goes on to discuss the issue in more detail and to give an account of the behaviour of Justin Bieber.

We use Wikipedia as the background corpus, as it is large, represents many areas of life, and is freely available. Contrasting the Justin Bieber story to the contents of Wikipedia, using the model described in the previous section, we obtain a list of word pairs ranked by how specific they are to the news story (Table 2). Pairs with lower scores tend to be quite common associations (e.g. los angeles, sports car, street car, etc). Pairs with top scores seem to capture the essence of the news story well. Clearly the associations suggest that the news story has something to do with Bieber, police, Miami and alcohol (and "saying" something, which is not typical in Wikipedia, our background corpus, but is typical in news stories like this one).

Using words in the top associations, the following sample poem was generated:

Race at the miami-dade justins in the marijuana!

²<http://www.gutenberg.org>

Top pairs	Bottom pairs
say, bieber	los, angeles
say, police	later, jail
miami, bieber	sport, car
miami, say	car, early
bieber, police	thursday, early
beach, bieber	marijuana, alcohol
beach, police	prescription, alcohol
car, say	sport, thursday
bieber, alcohol	car, street
bieber, los	prescription, marijuana

Table 2: The top and the bottom foreground associations for the Justin Bieber’s news story.

*The officer is taller than you, who race yourself
So miami-dade and miami-dade: race how its en-
tourages are said
Co-operate and later in the singer, like a angeles of
alcohols
Racing with jails and singers and co-operate race.*

This poem was one of the many we generated and, in a humorous way, it covers many different aspects of the news story. (Currently, our implementation does not fix capitalization and articles in the results, nor does it recognize compound words. These are left for future work; here we present results in the exact form produced by the implementation without editing them manually in any way.)

In order to illustrate the effect of using document specific associations, we next fix the template used for word substitution and two types of poems: 1) using words related to Justin Bieber in general, using Wikipedia as the background corpus (Toivanen et al. 2012), and 2) and using document specific words from the news story given above.

These poems are generated using words related Justin Bieber:

Is it the youtube, the justin,
the release of second times,
and the second celebrities of our says?
These are but brauns.

Is it the atlanta, the mallette,
the music of first uniteds,
and the song yorks of our defs?
These are but news.

Is it the chart, the braun,
the def of first ushers,
and the musical stratfords of our nevers?
These are but youtubes.

The following three poems have been produced using document specific associations:

Is it the miami, the street,
the jail of co-operate officers,
and the co-operate singers of our prescriptions?
These are but alcohols.

Is it the car, the sport,
the angeles of co-operate justins,
and the early lamborghinis of our entourages?
These are but singers.

Is it the entourage, the sport,
the singer of later lamborghinis,
and the early thursdays of our singers?
These are but justins.

Finally, instead of evaluating the methods with test subjects, we let the readers of this paper decide for themselves by providing a collection of 18 poems at the end of this paper. To make this reader evaluation as fair as possible, *we did not select or edit the poems in any way*. We selected three news stories, of different topics and of sufficiently general interest, based on their original contents but not on the poems produced. Then, without any testing of the suitability of those stories for association extraction and poetry generation, we ran the poetry machinery and added the first poems produced for each of the news stories in the collection at the end of this paper.

The three news stories are the following:

- The aforementioned news story about Justin Bieber.
- A news story *Ukrainian Prime Minister Resigns as Parliament Repeals Restrictive Laws*³ published by NY Times on January 28.
- A news story *The return of the firing squad? US states reconsider execution methods*⁴ published by The Guardian on January 28.

To get some understanding how different background corpora affect the results, we used two different background corpora: the English Wikipedia and the Project Gutenberg corpus. We used each background to generate three poems from each news story: in each collection of six poems, poems 1–3 are generated by using Wikipedia as background, and poems 4–6 using Project Gutenberg as background.

Conclusions and Future Work

In this paper we have proposed a novel approach for using document-specific word associations to provide content words in a poetry generation task. As a novel part of the methodology, we use a recent model that extracts word pairs that are specific to a given document in a statistical sense.

Instead of an objective evaluation with some fixed criteria, we invite the readers of this paper to read the poems generated by the system — called P.O. Eticus — in the next pages and form their own opinions on the methods and results.

Automated methods for poetry generation from given documents could have practical application areas. For instance, the methodology has already been used in an art project exhibited in Estonia and Finland (Gross et al. 2014). Similarly the poems could be used for entertainment or as

³<http://nyti.ms/1k0kj9r>

⁴<http://gu.com/p/3m8p5>

automatically generated thought-provoking mechanisms in news websites or internet forums.

An interesting direction for further developments would be combining together documents on the same topic and then producing poems which give an overview of the *diverse* aspects of the topic. For instance each verse could cover some specific documents, or a step further we could use document clustering for identifying key subtopics and creating verses from these.

Acknowledgments

This work has been supported by the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733 (ConCreTe), the Algorithmic Data Analysis (Algodan) Centre of Excellence of the Academy of Finland, and the Helsinki Doctoral Program in Computer Science (HECSE).

References

- Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.
- Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, 31–40.
- Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22–29.
- Colton, S.; Goodwin, J.; and Veale, T. 2012. Full face poetry generation. In *Proceedings of the Third International Conference on Computational Creativity*, 95–102.
- Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391–407.
- Diaz-Agudo, B.; Gervás, P.; and González-Calero, P. A. 2002. Poetry generation in COLIBRI. In *ECCBR 2002, Advances in Case Based Reasoning*, 73–102.
- Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics* 19(1):61–74.
- Gervás, P. 2001. An expert system for the composition of formal spanish poetry. *Journal of Knowledge-Based Systems* 14(3–4):181–188.
- Gross, O.; Toivonen, H.; Toivanen, J. M.; and Valitutti, A. 2012. Lexical creativity from word associations. In *Knowledge, Information and Creativity Support Systems (KICSS), 2012 Seventh International Conference on*, 35–42.
- Gross, O.; Toivanen, J.; Lääne, S.; and Toivonen, H. 2014. Arts, news, and poetry - the art of framing. Submitted for review.
- Gross, O.; Doucet, A.; and Toivonen, H. 2014. Document summarization based on word associations. In *Proceedings of the 37th international ACM SIGIR conference on Research and Development in Information Retrieval*. ACM.
- Huovelin, J.; Gross, O.; Solin, O.; Lindn, K.; Maisala, S.; Oittinen, T.; Toivonen, H.; Niemi, J.; and Silfverberg, M. 2013. Software newsroom - an approach to automation of news search and editing. *Journal of Print and Media Technology Research* 3(2013):3:141–156.
- Manurung, H. M.; Ritchie, G.; and Thompson, H. 2000. Towards a computational model of poetry generation. In *Proceedings of AISB Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science*, 79–86.
- Manurung, H. 2003. *An evolutionary algorithm approach to poetry generation*. Ph.D. Dissertation, University of Edinburgh, Edinburgh, United Kingdom.
- Netzer, Y.; Gabay, D.; Goldberg, Y.; and Elhadad, M. 2009. Gaiku : Generating haiku with word associations norms. In *Proceedings of NAACL Workshop on Computational Approaches to Linguistic Creativity*, 32–39.
- Toivanen, J. M.; Toivonen, H.; Valitutti, A.; and Gross, O. 2012. Corpus-based generation of content and form in poetry. In *International Conference on Computational Creativity*, 175–179.
- Toivanen, J. M.; Järvisalo, M.; and Toivonen, H. 2013. Harnessing constraint programming for poetry composition. In *International Conference on Computational Creativity*, 160–167.
- Toivonen, H.; Gross, O.; Toivanen, J.; and Valitutti, A. 2013. On creative uses of word associations. In *Synergies of Soft Computing and Statistics for Intelligent Data Analysis, Part 1*, number 190 in *Advances in Intelligent Systems and Computing*. Springer. 17–24.
- Wong, M. T., and Chun, A. H. W. 2008. Automatic haiku generation using VSM. In *Proceedings of ACACOS*, 318–323.

Justin Bieber on Miami drink-drive charge after 'road racing'

Poems by P.O.Eticus

1. *It races at the singer, the later, racing singer, and he is race
within its officer and prescription. Inside is his thursday, his street, his sport, his
lamborghini, and his entourages. He is racing, and the entourages are said with singers
of miami, racing through miami-dade miami-dade. A miami says itself up at
the early entourage, and through the miami-dade miami in the car he can say
miami lamborghini, lazily racing among co-operate singers. A lamborghini in a
early cars and angeleses, and members race into his car, raced,
thursday, saying up like angeleses of member, higher and higher. Justin! The
members say on their later says. The thursday races up in early
later miamis of co-operate marijuana and says into the court. Car! And
there is only the car, the car, the beach, and the racing thursday.*
2. *Fruit can not race
through this co-operate beach:
car can not race into sport
that angeleses up and races
the angeleses of sports
and biebers the singers.*
3. *There is a miami-dade here
within my miami,
but miami-dade and sport....*
4. *I say;
perhaps I have staped;
this is a driving;
this is a incident;
and there is home....*
5. *Oh, he was bieber
Which then was he among the ferrari?
The co-operate, the slow, the medication?
I have transfered a first raymond of thursdays in one
But not this, this sport
Car!*
6. *Make, You! and canadian my driver;
That my ferraris race me no longer;
But thursday in your home.*

Ukrainian Prime Minister Resigns as Parliament Repeals Restrictive Laws

Poems by P.O.Eticus

1. *Water approved and restrictive by repealing building
Which laws and governments it into sundayukraine police weeks
Said with provincial opposition vote.
The repealing of the leader upon the statement
Is like a leader of week oppositions
In a concrete statement new resignation.*

2. *The statement approves into the party, and the party says him in a leader of
leader. But it is said with parliament and restrictive with sundayukraine streets.
The week parliaments. Repealing, repealing, saying, repeal, resigning, resign the
leaders. Over riots, and televisions, and votes, and streets. Approving its region on
the vote the government legislations, blocks itself through the leaders, and ministers
and repeals along the riots.*

3. *The svobodas
police from the resigns,
the televisions at their statements
resign lower through the ukraines.*

4. *And always concrete! Oh, if I could ride
With my week resigned concrete against the repeal
Do you resign I'd have a parliament like you at my television
With your azarov and your week that you resign me? O ukrainian week,
How I resign you for your parliamentary legislation!*

5. *Concrete one,
new and restrictive,
provincial repeal,
region,
concrete and leader you are vote
in our weeks.*

6. *Resigned amid jan
We will avoid all azarov;
And in the government
Resigning forth, we will resign restrictive votes
Over the repealed administration of azarov.*

The return of the firing squad? US states reconsider execution methods

Poems by P.O.Eticus

1. *Many one,
many and lethal,
recent injection,
republican,
recent and drug you are gas
in our electrocutions.*

2. *You are not he.
Who are you, choosing in his justice on the question
And lethal and lethal to me?
His doubt, though he rebuilt or found
Was always lethal and recent
And many to me.*

3. *I die;
perhaps I have began;
this is a doubt;
this is a prisoner;
and there is state....*

4. *You amid the public's pentobarbital longer,
You trying in the josephs of the methods above,
Me, your hanging on the michael, unusual franklins,
Me unusual michael in the states, ending you use
You, your court like a death, proposed, pentobarbital,
You, with your death all last, like the wyoming on a ended!*

5. *Lawmaker and quiet:
a brattin overdoses in the year courts
behind the process with the many new injection
across the brattin.*

6. *The longer rebuilds into the day, and the gas ends him in a supply of
schaefer. But it is divulged with west and powerful with republican penalties. The
process options. Coming, rebuilding, divulging, charles, looming, propose the
news. Over officials, and spectacles, and senators, and burns. Begining its florida on
the dodd the supply spectacles, franklins itself through the propofols, and burns and
proposes along the gases.*