“The Officer Is Taller Than You, Who Race Yourself!”

Using Document-Specific Word Associations in Poetry Generation

Jukka M. Toivanen, Oskar Gross, Hannu Toivonen

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Goals of this Work

- Generating poetry that reflects loosely a specific news story or other document
- Evoking fresh mental images and viewpoints that are related to the document but not exactly contained in it
  - Using document-specific words in poetry generation
  - Extension of the basic word substitution based poetry generation method by Toivanen et al. (2012)

Outline

- Our Aims in Poetry Generation
- Related Work
- Extracting Document-Specific Words
- Overview of the Generation
- Experiment and Examples
- Summary and Future Work

Our Aims in Poetry Generation

- Maximally unsupervised
  - Minimum amount of hand-crafted linguistic, world, and poetry domain knowledge
    - No explicit grammars
    - No manually generated templates
    - No knowledge bases
  - Use of statistical corpus-based methods
- Fresh mental images evoked by novel associations

Benefits and Restrictions of this Approach

- Pros:
  - Flexibility
  - Language independence
  - Direct learning from corpora, minimal amount of hand-crafted rules
- Cons:
  - The quality of the results varies a lot

Related Work

- Full-FACE poetry generation system (Colton, Goodwin, and Veale 2012)
- The system by Manurung et al. (2003)
- ASPERA system (Gervás 2001)
- Many others

Word Association Analysis

- Finding content words for replacement poetry
- General association calculation method proposed by Gross et al. (2012)
- Recent extension to document specific associations (Gross, Doucet, and Toivonen 2014)
  - Which word pairs are novel in a specific document?
  - A background corpus as a reference of novelty
  - Contrasting a specific document (called foreground) to a set of documents in the background corpus
  - log-likelihood ratio (LLR) used to measure document-specific word associations

Example News Story

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. (…)

BBC News, 23 January 2014

Document-Specific Word Associations

- Descriptive associations could be:
  - “bieber” and ”alcohol”
  - ”bieber” and ”prescription”
  - ”justin” and ”alcohol”
  - …
- Not so descriptive associations include:
  - “justin” and “bieber”
  - “sports” and “car”
  - “driving” and “licence”

How to Find Document-Specific Word Associations

Counts in the News Story

	Bieber	¬ Bieber
Alcohol	2	0
¬ Alcohol	4	22

Counts in the Background Corpus

	Bieber	¬ Bieber
Alcohol	0.	19419.
¬ Alcohol	244.	33967685.

How to Find Document-Specific Word Associations

Counts in the News Story

	Justin	¬ Justin
Bieber	3	3
¬ Bieber	0	22

Counts in the Background Corpus

	Justin	¬ Justin
Bieber	5.	239.
¬ Bieber	3747.	33983357.

How to Find Document-Specific Word Associations

Foreground Counts

	w_1	$\neg w_1$
w_2	k_{11}	k_{12}
$\neg w_2$	k_{21}	k_{22}

Background Counts

	w_1	$\neg w_1$
w_2	k'_{11}	k'_{12}
$\neg w_2$	k'_{21}	k'_{22}

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

Document-Specific Word Associations

- Find word pairs whose co-occurrence distribution in the document deviates most from the background corpus
- These words are descriptive for the novel content of the document in question
- Use these words as replacements in the poetry generation phase

Example Associations

Most novel pairs	Least novel 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

Overview of the Generation

- Word substitution method as described by Toivanen et al. (2012)
  - A piece of text from a corpus (e.g. poetry)
  - Replacing most of the words with words relevant to a specific news story
  - Morphological analysis and synthesis
    - Stanford POS-tagger (Toutanova et al., 2003)
    - morpha & morphg tools (Minnen, Carroll, and Pearce, 2001)

Overview of the Generation

Is it the dirt, the squalor,

the wear of human bodies,

and the dead faces of our neighbours?

These are but symbols.

Project Gutenberg, Imagist Poetry

Is it the entourage, the sport,

the singer of later lamborghinis,

and the early thursdays of our singers?

These are but justins.

P.O. Eticus-Apparatus

Experiment

- The corpus from which templates were taken contained mostly Imagist poetry from the Project Gutenberg
- Background corpus was the English Wikipedia
- Several different news stories, e.g.
  - Justin Bieber drinking and driving
  - Huawei profits surging
  - Ukrainian prime minister resigning
  - US states reconsidering execution methods
- The following poems were selected randomly and presented as they are

Huawei Profits Surge...

Oaks

and impact technologies,

rise buying with transfer, rise:

their comfortable technology.

Ukrainian Prime Minister Resigning...

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!

US States Reconsidering Execution Methods

I die;

perhaps I have begun; this is a doubt;

this is a prisoner;

and there is state....

Summary and Future Work

- A novel method for identifying document-specific words
- Document-specific word associations to provide content words in a poetry generation task
- Future Work:
  - Evaluation
  - More statistical natural language analysis to improve the results
  - Combining different methods to generate poetry
  - Producing poems which give an overview of a set of similar documents

References

- Colton, S.; Goodwin, J.; and Veale, T. 2012. Full-face poetry generation. In International Conference on Computational Creativity, 95–102.
- Gervás, P. 2001. An expert system for the composition of a formal spanish poetry. Journal of Knowledge-Based Systems, 14(3–4):181–188.
- 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.
- Minnen, G.; Carroll, J.; and Pearce, D. 2001. Applied morphological processing of English. Natural Language Engineering 7(3):207–223.
- 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.
- 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.
- Toutanova, K.; Klein, D.; Manning, C.; and Singer, Y. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proceedings of HLT-NAACL, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 252–259.