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

Using Document-Specific Word Associations in Poetry Generation

Jukka M. Toivanen, Oskar Gross, Hannu Toivonen
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
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

<table>
<thead>
<tr>
<th></th>
<th>Bieber</th>
<th>¬ Bieber</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>¬ Alcohol</td>
<td>4</td>
<td>22</td>
</tr>
</tbody>
</table>

### Counts in the Background Corpus

<table>
<thead>
<tr>
<th></th>
<th>Bieber</th>
<th>¬ Bieber</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol</td>
<td>0.</td>
<td>19419.</td>
</tr>
<tr>
<td>¬ Alcohol</td>
<td>244.</td>
<td>33967685.</td>
</tr>
</tbody>
</table>
How to Find Document-Specific Word Associations

<table>
<thead>
<tr>
<th>Counts in the News Story</th>
<th>Justin</th>
<th>¬ Justin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bieber</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>¬ Bieber</td>
<td>0</td>
<td>22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Counts in the Background Corpus</th>
<th>Justin</th>
<th>¬ Justin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bieber</td>
<td>5.</td>
<td>239.</td>
</tr>
<tr>
<td>¬ Bieber</td>
<td>3747.</td>
<td>33983357.</td>
</tr>
</tbody>
</table>
How to Find Document-Specific Word Associations

**Foreground Counts**

<table>
<thead>
<tr>
<th></th>
<th>$w_1$</th>
<th>$\neg w_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_2$</td>
<td>$k_{11}$</td>
<td>$k_{12}$</td>
</tr>
<tr>
<td>$\neg w_2$</td>
<td>$k_{21}$</td>
<td>$k_{22}$</td>
</tr>
</tbody>
</table>

**Background Counts**

<table>
<thead>
<tr>
<th></th>
<th>$w_1$</th>
<th>$\neg w_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_2$</td>
<td>$k'_{11}$</td>
<td>$k'_{12}$</td>
</tr>
<tr>
<td>$\neg w_2$</td>
<td>$k'_{21}$</td>
<td>$k'_{22}$</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Most novel pairs</th>
<th>Least novel pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>say, bieber</td>
<td>los, angeles</td>
</tr>
<tr>
<td>say, police</td>
<td>later, jail</td>
</tr>
<tr>
<td>miami, bieber</td>
<td>sport, car</td>
</tr>
<tr>
<td>miami, say</td>
<td>car, early</td>
</tr>
<tr>
<td>bieber, police</td>
<td>thursday, early</td>
</tr>
<tr>
<td>beach, bieber</td>
<td>marijuana, alcohol</td>
</tr>
<tr>
<td>beach, police</td>
<td>prescription, alcohol</td>
</tr>
<tr>
<td>car, say</td>
<td>sport, thursday</td>
</tr>
<tr>
<td>bieber, alcohol</td>
<td>car, street</td>
</tr>
<tr>
<td>bieber, los</td>
<td>prescription, marijuana</td>
</tr>
</tbody>
</table>
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 lamborghini,
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


