Automatic Poetry Generation from Prosaic Text

Tim Van de Cruys
Institut de Recherche en Informatique de Toulouse (IRIT)
Artificial and Natural Intelligence Toulouse Institute (ANITI)
CNRS, Toulouse
tim.vandecruys@irit.fr

Motivation

- language models based on neural networks have improved the state of the art with regard to predictive language modeling
- topic models are successful at capturing clear-cut, semantic dimensions.
- In this paper, we explore how these approaches can be adapted and combined to model the linguistic and literary aspects needed for **poetry generation**.

Innovation

- Even though it only uses standard, non-poetic text as input, the system yields state of the art results for poetry generation.
- we make use of a latent semantic model to model topic coherence

Model Framework

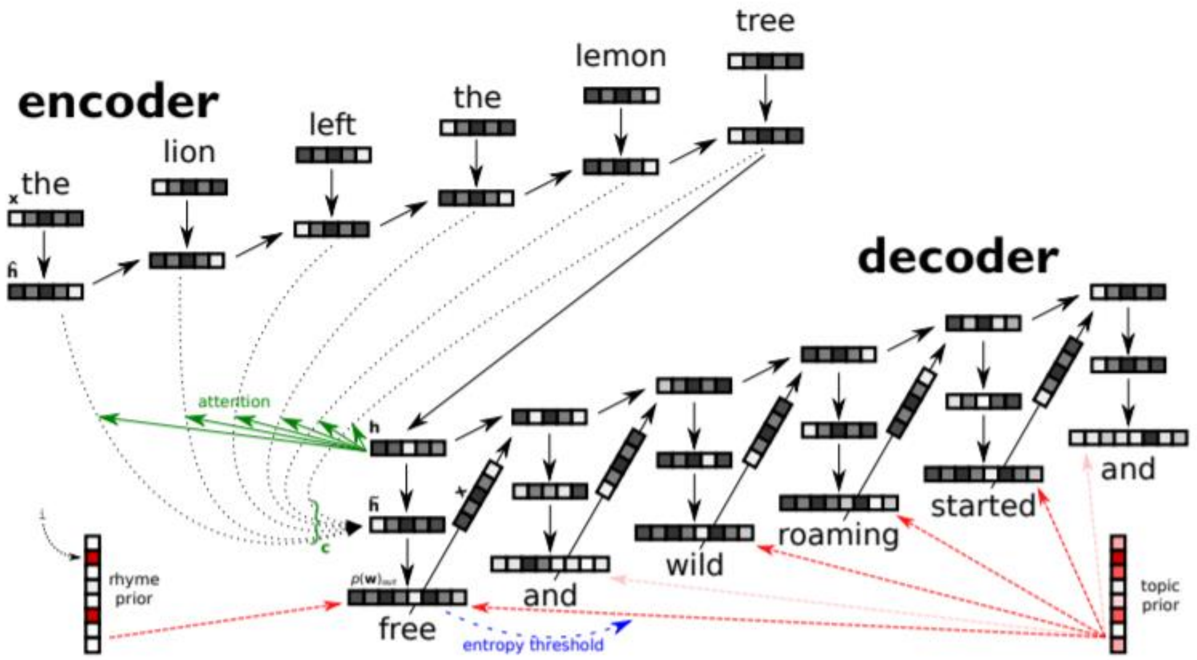


Figure 1: Graphical representation of the poetry generation model. The encoder encodes the current verse, and the final representation is given to the decoder, which predicts the next verse word by word in reverse. The attention mechanism is represented for the first time step. The rhyme prior is applied to the first time step, and the topic prior is optionally applied to all time steps, mediated by the entropy threshold of the network's output distribution.

Neural architecture

- The core of our poetry is a gated recurrent network (GRU) equipped with attention mechanism.

$$
\begin{aligned}
\mathbf{r}_t &= \sigma(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \hat{\mathbf{h}}_{t-1}) && (1)\\
\mathbf{z}_t &= \sigma(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \hat{\mathbf{h}}_{t-1}) && (2)\\
\bar{\mathbf{h}}_t &= \tanh(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r}_t \odot \hat{\mathbf{h}}_{t-1})) && (3)\\
\hat{\mathbf{h}}_t &= (1-\mathbf{z}_t) \odot \hat{\mathbf{h}}_{t-1} + \mathbf{z}_t \odot \bar{\mathbf{h}}_t && (4)
\end{aligned}
$$

$$\mathbf{a}_t(i) = \frac{\exp(\mathrm{score}(\mathbf{h}_t, \hat{\mathbf{h}}_i))}{\sum_{i'} \exp(\mathrm{score}(\mathbf{h}_t, \hat{\mathbf{h}}_{i'}))}$$

$$\mathrm{score}(\mathbf{h}_t, \hat{\mathbf{h}}_i) = \mathbf{h}_t^T \mathbf{W_a} \hat{\mathbf{h}}_i$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W_c}[\mathbf{c}_t; \mathbf{h}_t])$$

$$p(\mathbf{w}^t | w^{<t}, S_i) = \mathrm{softmax}(\mathbf{W_s} \tilde{\mathbf{h}}_t)$$

Neural architecture

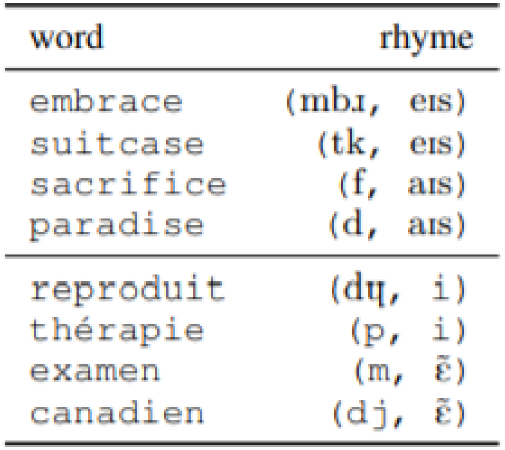
- As an objective function, the sum of the logprobabilities of the next sentence is optimized:

$$J_t = \sum_{(S_i, S_{i+1}) \in C} - \log p(S_i | S_{i+1})$$

- At inference time, each word is then sampled randomly according to the output probability distribution
- Crucially, the decoder is trained to predict the next sentence in **reverse**, which is important for the incorporation of rhyme.

Rhyme constraint

- In order to adequately model the rhyme constraint, we make use of a phonetic representation of words, extracted from the online dictionary Wiktionary.
- It help us determine its rhyme sound for each word.



word	rhyme
embrace	(mbɹ, eɪs)
suitcase	(tk, eɪs)
sacrifice	(f, aɪs)
paradise	(d, aɪs)
reproduit	(dɥ, i)
thérapie	(p, i)
examen	(m, ɛ̃)
canadien	(dj, ɛ̃)

Rhyme constraint

- We create a probability distribution for a particular rhyme sound:

$$p_{rhyme}(\mathbf{w}) = \frac{1}{Z}\mathbf{x} \text{ with } \begin{cases} x_i = 1 & \text{if } i \in R \\ x_i = \epsilon & \text{otherwise} \end{cases} \tag{10}$$

- Where R is a set of words that contain the required rhyme sound,ϵ is a small value close to zero. We can now use $p_{rhyme}(\boldsymbol{w})$ as a prior probability distribution in order to reweight the neural network's standard output probability:

$$p_{out}(\mathbf{w}) = \frac{1}{Z}(p(\mathbf{w}^t|w^{<t}, S_i) \odot p_{rhyme}(\mathbf{w})) \tag{11}$$

Rhyme constraint

- As we noted before, each verse is generated in reverse; the reweighting of rhyme words is applied at the first step of the decoding process, and the rhyme word is generated first.

Topical constraint

- As input to the method, we construct a frequency matrix $\boldsymbol{A}$, which captures cooccurrence frequencies of vocabulary words and context words. we use nonnegative matrix factorization (NMF) to factorize $\boldsymbol{A}$:

$$\mathbf{A}_{i\times j} \approx \mathbf{W}_{i\times k}\mathbf{H}_{k\times j} \qquad (12)$$

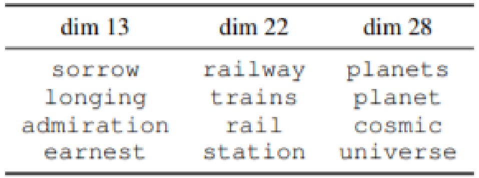
- Where k is much smaller than i and j. Matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ are randomly initialized, and the rules in 13 and 14 are iteratively applied:

$$\mathbf{H}_{a\mu} \leftarrow \mathbf{H}_{a\mu}\frac{\sum_i \mathbf{W}_{ia}\frac{\mathbf{A}_{i\mu}}{(\mathbf{WH})_{i\mu}}}{\sum_k \mathbf{W}_{ka}} \qquad (13)$$

$$\mathbf{W}_{ia} \leftarrow \mathbf{W}_{ia}\frac{\sum_\mu \mathbf{H}_{a\mu}\frac{\mathbf{A}_{i\mu}}{(\mathbf{WH})_{i\mu}}}{\sum_v \mathbf{H}_{av}} \qquad (14)$$

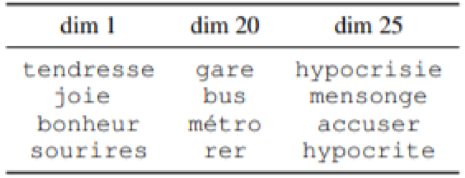
Topical constraint

- The following is the number of examples induced by the model:



dim 13	dim 22	dim 28
sorrow	railway	planets
longing	trains	planet
admiration	rail	cosmic
earnest	station	universe

Table 2: Three example dimensions from the NMF model for English (4 words with highest probability)



dim 1	dim 20	dim 25
tendresse	gare	hypocrisie
joie	bus	mensonge
bonheur	métro	accuser
sourires	rer	hypocrite

- The factorization that comes out of the NMF model can be interpreted probabilistically, matrix $\boldsymbol{W}$ can be considered as $p(\boldsymbol{w}|k)$, i.e. the probability of a word given a latent dimension k.

Topical constraint

- it would be straightforward to simply use $p(\boldsymbol{w}|k)$**,**
- t. Initial experiments, however, indicated that such a blind modification of the output probability distribution for every word of the output sequence is **detrimental** to syntactic fluency
- To combine syntactic fluency with topical consistency, we condition the weighting of the output probability distribution on the entropy of that distribution: when the output distribution's entropy is low, the model is confident of the choice of the next word , then we don't change it; otherwise, we will modify the distribution by using the topical distribution $p(\boldsymbol{w}|k)$**,**

A global optimization framework

- For each final verse, the model generates a considerable number of candidates; each candidate verse is then scored according to the following criteria:
- 1) the log-probability score of the generated verse
- 2) compliance with the rhyme constraint, give a higher score to rhyme words with disparate preceding consonant groups,
- 3) compliance with the topical constraint, modeled as the sum of the probabilities of all words for the defined dimension
- 4) the log-probability score of a standard n-gram model.

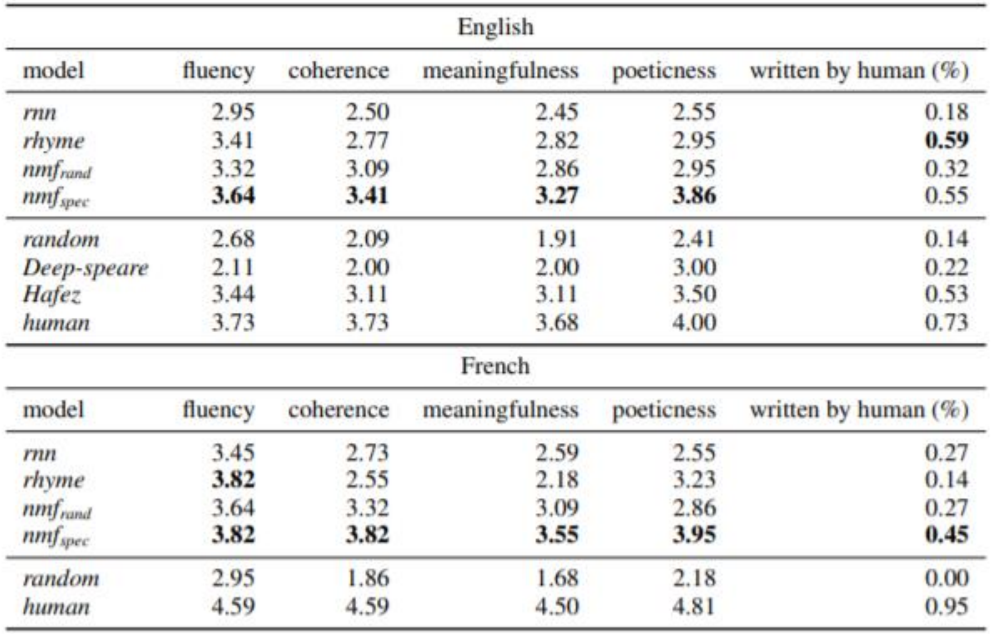
A global optimization framework

- The score for each criterion is normalized to the interval [0,1],
- and the harmonic mean of all scores is taken as the final score
- We keep the candidate with the highest score

Results and evaluation

- We train two different models for the generation of poetry in both English and French.
- The neural architecture is trained on a large corpus of generic web texts, called CommonCrawl corpus.
- We adopt the evaluation framework by Zhang and Lapata (2014), human annotators are asked to evaluate poems on a five point scale with regard to four aspects: 1)fluency 2)coherence 3)meaningfulness 4)poeticness

Results and evaluation



English

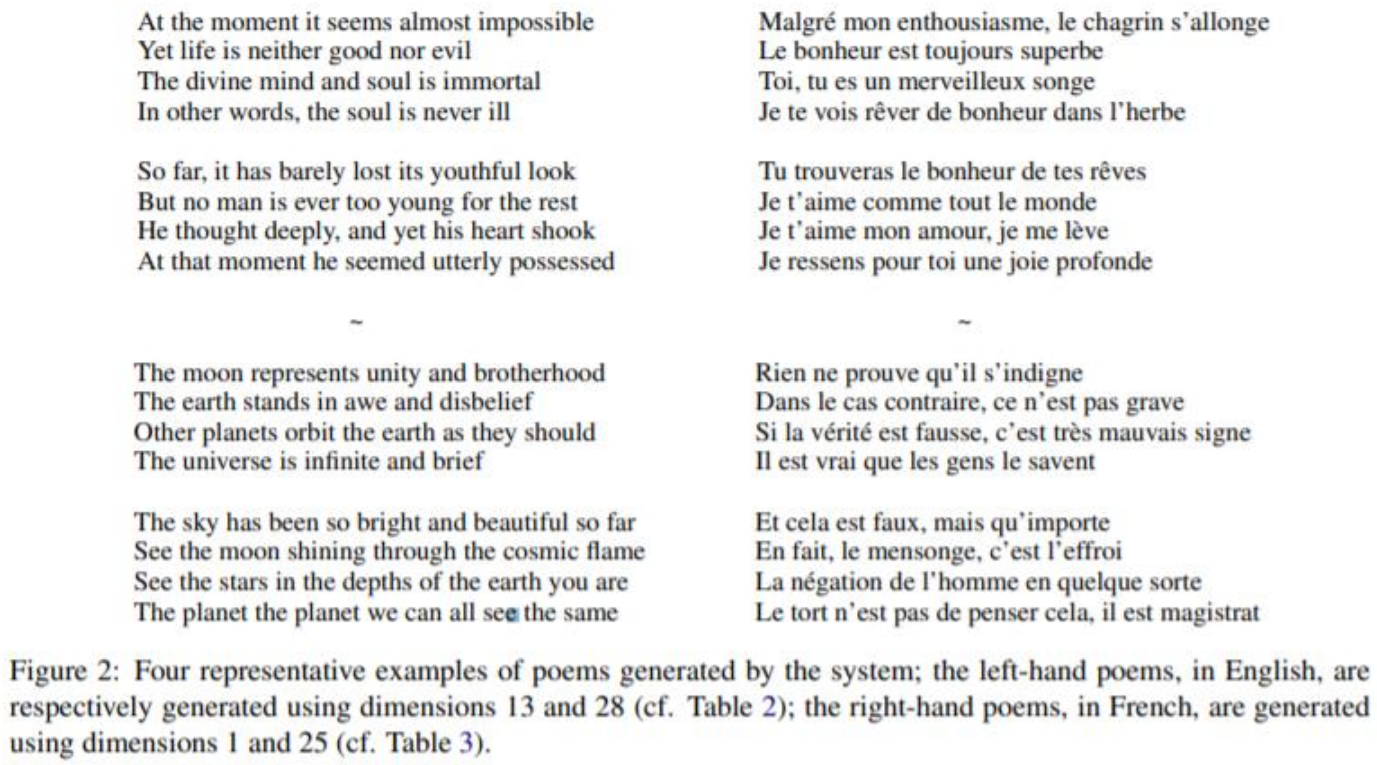
model	fluency	coherence	meaningfulness	poeticness	written by human (%)
rnn	2.95	2.50	2.45	2.55	0.18
rhyme	3.41	2.77	2.82	2.95	**0.59**
nmf$_{rand}$	3.32	3.09	2.86	2.95	0.32
nmf$_{spec}$	**3.64**	**3.41**	**3.27**	**3.86**	0.55
random	2.68	2.09	1.91	2.41	0.14
Deep-speare	2.11	2.00	2.00	3.00	0.22
Hafez	3.44	3.11	3.11	3.50	0.53
human	3.73	3.73	3.68	4.00	0.73

French

model	fluency	coherence	meaningfulness	poeticness	written by human (%)
rnn	3.45	2.73	2.59	2.55	0.27
rhyme	**3.82**	2.55	2.18	3.23	0.14
nmf$_{rand}$	3.64	3.32	3.09	2.86	0.27
nmf$_{spec}$	**3.82**	**3.82**	**3.55**	**3.95**	**0.45**
random	2.95	1.86	1.68	2.18	0.00
human	4.59	4.59	4.50	4.81	0.95

Table 4: Results of the human evaluation (mean score of all annotators) for English and French; values in **bold** indicate best performance of all generation models

Results and evaluation



At the moment it seems almost impossible
Yet life is neither good nor evil
The divine mind and soul is immortal
In other words, the soul is never ill

So far, it has barely lost its youthful look
But no man is ever too young for the rest
He thought deeply, and yet his heart shook
At that moment he seemed utterly possessed

~

The moon represents unity and brotherhood
The earth stands in awe and disbelief
Other planets orbit the earth as they should
The universe is infinite and brief

The sky has been so bright and beautiful so far
See the moon shining through the cosmic flame
See the stars in the depths of the earth you are
The planet the planet we can all see the same

Malgré mon enthousiasme, le chagrin s'allonge
Le bonheur est toujours superbe
Toi, tu es un merveilleux songe
Je te vois rêver de bonheur dans l'herbe

Tu trouveras le bonheur de tes rêves
Je t'aime comme tout le monde
Je t'aime mon amour, je me lève
Je ressens pour toi une joie profonde

~

Rien ne prouve qu'il s'indigne
Dans le cas contraire, ce n'est pas grave
Si la vérité est fausse, c'est très mauvais signe
Il est vrai que les gens le savent

Et cela est faux, mais qu'importe
En fait, le mensonge, c'est l'effroi
La négation de l'homme en quelque sorte
Le tort n'est pas de penser cela, il est magistrat

Figure 2: Four representative examples of poems generated by the system; the left-hand poems, in English, are respectively generated using dimensions 13 and 28 (cf. Table 2); the right-hand poems, in French, are generated using dimensions 1 and 25 (cf. Table 3).