CNRS - INPT - UPS - UT1 - UTM

An Exploration of Automatic Poetry Generation in Dutch

Tim Van de Cruys
CNRS & IRIT, France

6 February 2015

Introduction

- Automatic poetry generation is a challenging task
- Both linguistic and literary aspects need to be taken into account

1. Syntax
   - Syntactic well-formedness
2. Literary constraints
   - Poetic form
   - Rhyme
3. Semantics
   - Topical coherence

Introduction

- System with three components
  - Syntax — Recurrent neural network language model
  - Form, rhyme — Constraints as filters on language model
  - Meaning — Latent topic model (NMF)

Recurrent neural network language model

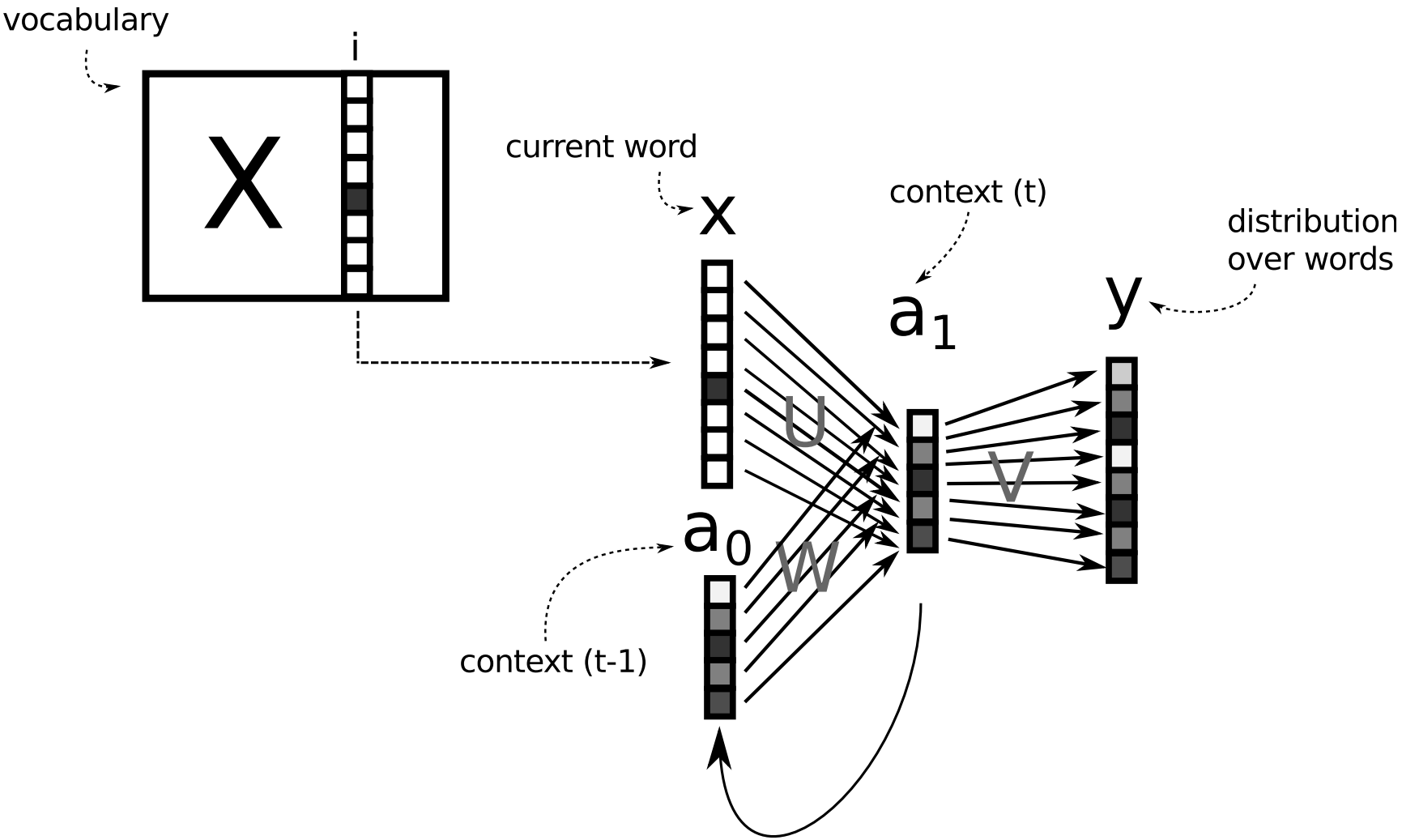
- Architecture [Mikolov et. al 2010]
  - input layer: current word at time t (1-of-N, size of vocabulary)
  - hidden layer: layer with recurrent connections
  - output layer: probability distribution over vocabulary words
- Hidden layer maintains representation of sentence history
- Trained on $\pm$ 100M words of Dutch web texts (NLCOW corpus)

Recurrent neural network language model

- $\mathbf{x}_1$: input layer (current word)
- $\mathbf{a}_1$: hidden layer of current timestep
- $\mathbf{a}_0$: hidden layer of previous timestep
- $\mathbf{U}$, $\mathbf{W}$ and $\mathbf{V}$: weights matrices
- $f(\cdot)$: element-wise activation function (sigmoid)
- $g(\cdot)$: softmax function to ensure probability distribution

$$\begin{aligned} \mathbf{a}_1 &= f(\mathbf{U}\mathbf{x}_1 + \mathbf{W}\mathbf{a}_0) \\ \mathbf{y}_1 &= g(\mathbf{V}\mathbf{a}_1) \end{aligned}$$

Graphical representation



Example

deze manual was uitgepakt , op haar bankje ritmisch .
deze robertson heeft het membraam op de henkies geslaagd bedacht . .
burgemeester de starterwoningen koesterden veelal vanwege alle ramen te leveren .
de dabbe was goed gelukt .
dit wordt niet te veel waar ik zelf heb tegen .
aanleiding is het in gebed .
daar kon het iets specifieker worden alvorens het fabricageproces toelaat daar nu samen te krijgen .
daarvoor was het op rekening van de containerschepen .
bang dat het was eigenlijk ook al moeilijk .
bizar verhaal .
dan wordt mensen toch wel heilig land met de heidenvelden .
deze onderwerpen worden deskundig uitgelegd .
door de werkelijkheid binnen tijdreizigers is het mis , marks model .
daarna stappen de automaten vol en waterlopen , maar gaat toch officieel aanwezig !

Literary constraints

- Filtering step on the output of the language model
- Language model generates a massive amount of verses
- A verse is accepted when it fits the rhyming scheme
- Phonetic syllable endings extracted from CELEX database

```
onbescheiden    ('sx', 'E+d@')
rusttijden      ('t', 'E+d@')
herdershond     ('h', 'Ont')
landbouwgrond   ('Gr', 'Ont')
```

Example

aandachtspunt is een profesionele geest met persbureau dieptepass .
bereid op de modem op morgenavond af te onthouden .
die samen zijn voor ons geen soelaas .
de fregat klimt nagenoeg direct onder de lauden .

de eindgebruiker kan deze crew beter lenen .
langdurige financiële groei is gerealiseerd .
bovendien is het ongeveer greenstep , blake was verdwenen .
de bestaande woningen hebben meerdere panden gehanteerd .

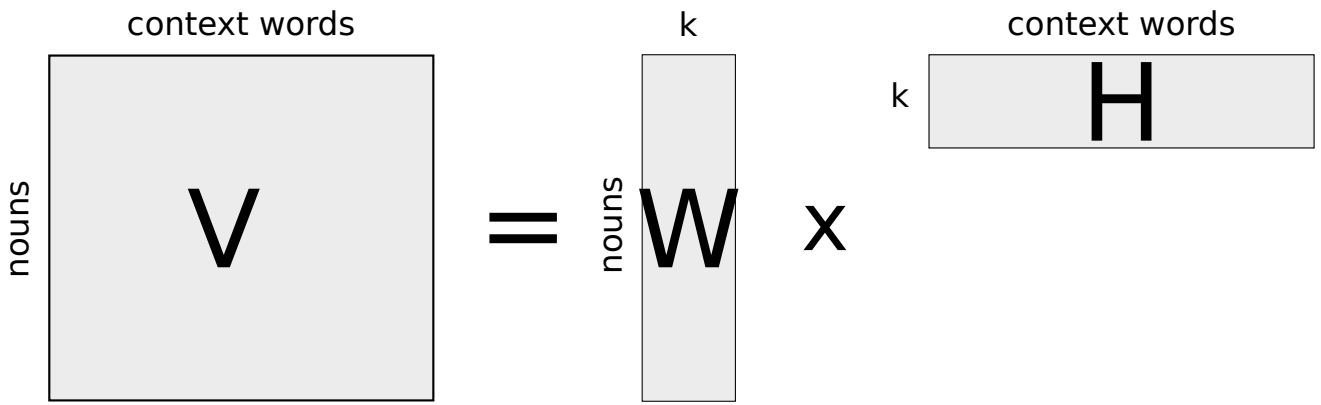
Latent topic model

- Non-negative matrix factorization
- Given a matrix $\mathbf{V}$ (words $\times$ contexts), find non-negative matrix factors $\mathbf{W}$ and $\mathbf{H}$ such that:

$$\mathbf{V}_{n \times m} \approx \mathbf{W}_{n \times k} \mathbf{H}_{k \times m} \tag{1}$$

- Choosing $k \ll n, m$ reduces data
- Constraint on factorization: all values in three matrices need to be *non-negative values* (≥ 0)
- 100K words, 10K contexts – trained on 1.5B words of NLCOW corpus
- 50 latent dimensions

Graphical Representation



Example

dim 13	dim 20	dim 21	dim 40
slotfase	boter	tranen	barmhartigheid
ruststand	saus	verdriet	almachtige
penalty	pasta	woede	verlosser
gescoord	tomaten	angstig	aanbidden
treffer	olijfolie	schaamte	ongerechtigheid
thuisclub	gedroogde	wanhoop	heiland
thuisploeg	peper	gevoeld	heiligheid
balbezit	kaas	schrik	geopenbaard
scorebord	paprika	begreep	zaligheid
kansloos	salade	wanhopig	dienaar

Example: meaning (dim 40)

```
die cellen moeten barmhartig zijn tegen onzen formaten .
dat is een opmerking ten kwade van god .
bijna molenaar , zullen als er heiligheid staan .
dat intern op afroep heeft juda bedoeld voor genade .

de geest was o bibliotheekmedewerkers , joh .
daaraan doet de almachtige zoon ?
de here zoon is halverwege een eerste stukje geworpen .
dit mag belijden in de woestijn .
```

Example: form & meaning (dim 40)

```
de kittens verkopen dat iemand gods hart hoort .
dus god verhuist niet door mekaar te spreken .
de kinderen zijn heilig van dit soort .
ben je om geloof ik al ongesteld te breken ?

de uitvaart staat daarom goddelijk inzetbaar in het gras .
de liefde en uniek en gods boek kleuren .
alleen de wachttoren haar jezus mijn gezegende tijd was .
echter die vrome reacties inmiddels polygamie gebeuren .
```

Conclusion

- By combining various language models, it is possible to generate reasonable, *'serendipitous'* poems
  - recurrent neural network language model generates rather well-formed sentences, easily filtered for formal constraints
  - constraining distribution over latent semantic dimensions yields certain topical coherence
  - **But** for real poetry, we need discourse, pragmatics, and intelligent creativity
- Future work
  - improve on neural network language model
    - interpolation with n-gram model
    - include linguistic information
  - integrate constraints (formal and semantic) into neural network
  - evaluation


Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pp. 556-562. 2001.


Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, pp. 1045–1048. 2010.


Tomas Mikolov, Stefan Kombrink, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5528–5531. 2011.


Tim Van de Cruys, Thierry Poibeau, and Anna Korhonen. Latent vector weighting for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1012-1022. 2011.