CNRS - INPT - UPS - UT1 - UTM

Generating Topically Coherent Text with Recurrent Neural Networks

Tim Van de Cruys
CNRS & IRIT, France

26 January, 2018

Introduction

- Recurrent neural networks provide adequate language models
- Impressive performance for natural language generation
- Uninformed language models generate non-coherent language utterances
- Research question: can we inform language models in order to generate **topically coherent** sentences?

Introduction

deze is bij ons ook verkrijgbaar voor de grijze massa

een heel prettige dag en wat een zegen in het nieuwe jaar

het is absoluut niet zo dat je het uitstekend hebt gedaan met dit initiatief

ze zitten op het hoogste punt van de plattegrond , op verschillende posities in de haven

voortaan is dit een veel belangrijker element in het ontwerp van het nieuwe stelsel

het was een goede vraag , maar geen antwoord op de vraag over de hoogte van die vergoeding

ik laat mij dan ook niet vertellen dat ik adhd heb

Introduction

Het treinverkeer tussen Kontich en Mechelen is woensdagochtend verstoord door een aanrijding. Sinds 8.10 uur zijn alle sporen weer beschikbaar tussen Kontich en Mechelen. Dat laten de NMBS en Infrabel weten. Het treinverkeer tussen Antwerpen en Brussel zal wel nog een tijd zware hinder ondervinden, met vertragingen die kunnen oplopen tot een uur.

Introduction

Het treinverkeer tussen Kontich en Mechelen is woensdagochtend verstoord door een aanrijding. Sinds 8.10 uur zijn alle sporen weer beschikbaar tussen Kontich en Mechelen. Dat laten de NMBS en Infrabel weten. Het treinverkeer tussen Antwerpen en Brussel zal wel nog een tijd zware hinder ondervinden, met vertragingen die kunnen oplopen tot een uur.

$\rightarrow$ topical coherence

Introduction

Het treinverkeer tussen Kontich en Mechelen is woensdagochtend verstoord door een aanrijding. Sinds 8.10 uur zijn alle sporen weer beschikbaar tussen Kontich en Mechelen. Dat laten de NMBS en Infrabel weten. Het treinverkeer tussen Antwerpen en Brussel zal wel nog een tijd zware hinder ondervinden, met vertragingen die kunnen oplopen tot een uur.

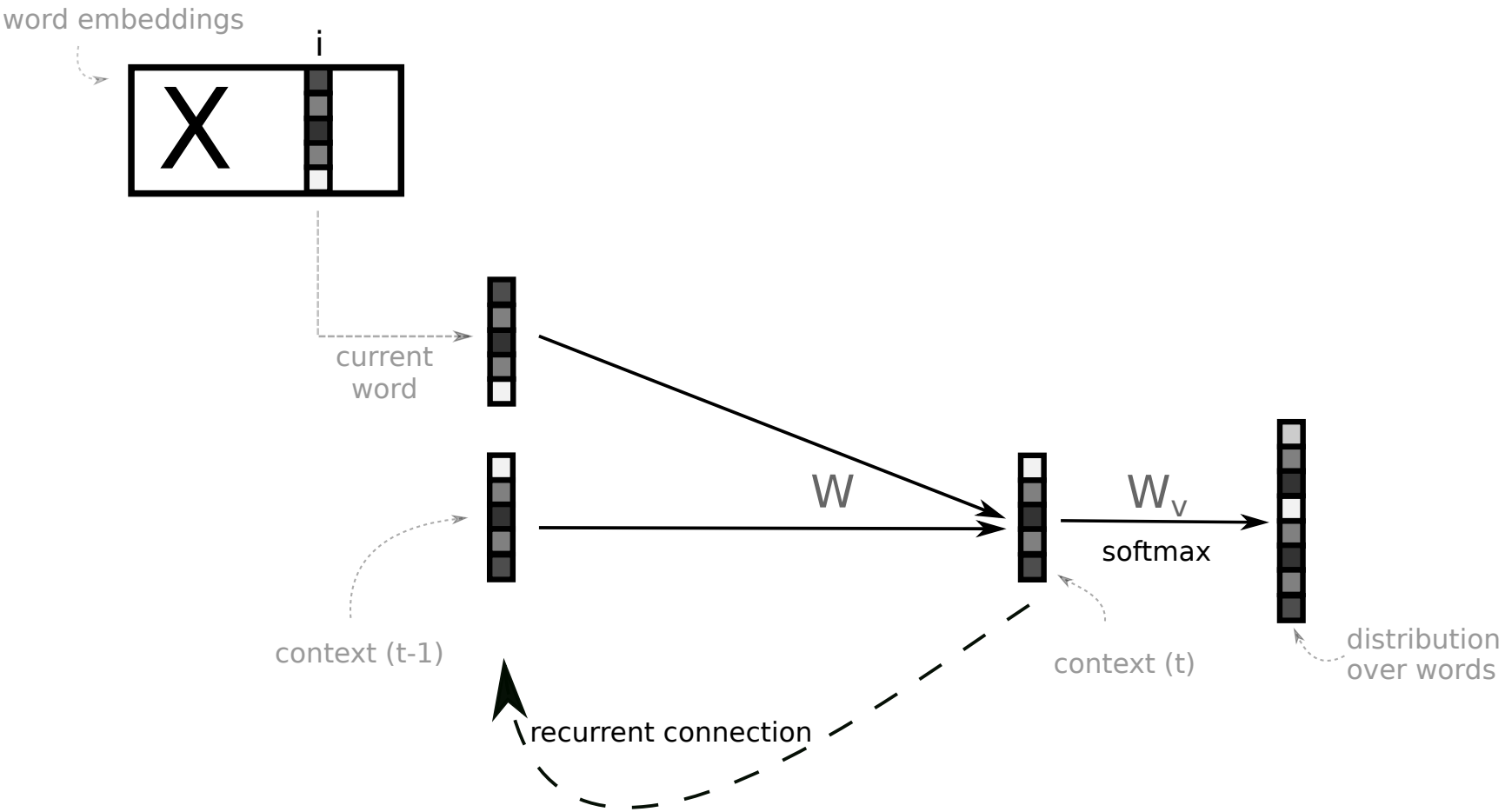
$\rightarrow$ discourse coherence

Recurrent neural network language model

- Long Short Term Memory (LSTM) Network
  - Recurrent neural network with memory cell state
  - input layer: current word at time t, represented as word embedding
  - output layer: probability distribution over vocabulary words
  - input, forget, output gate: control flow of information
- Recurrent connections allow construction of entire sentence history
- Gates allow fine-grained control of what information the network remembers

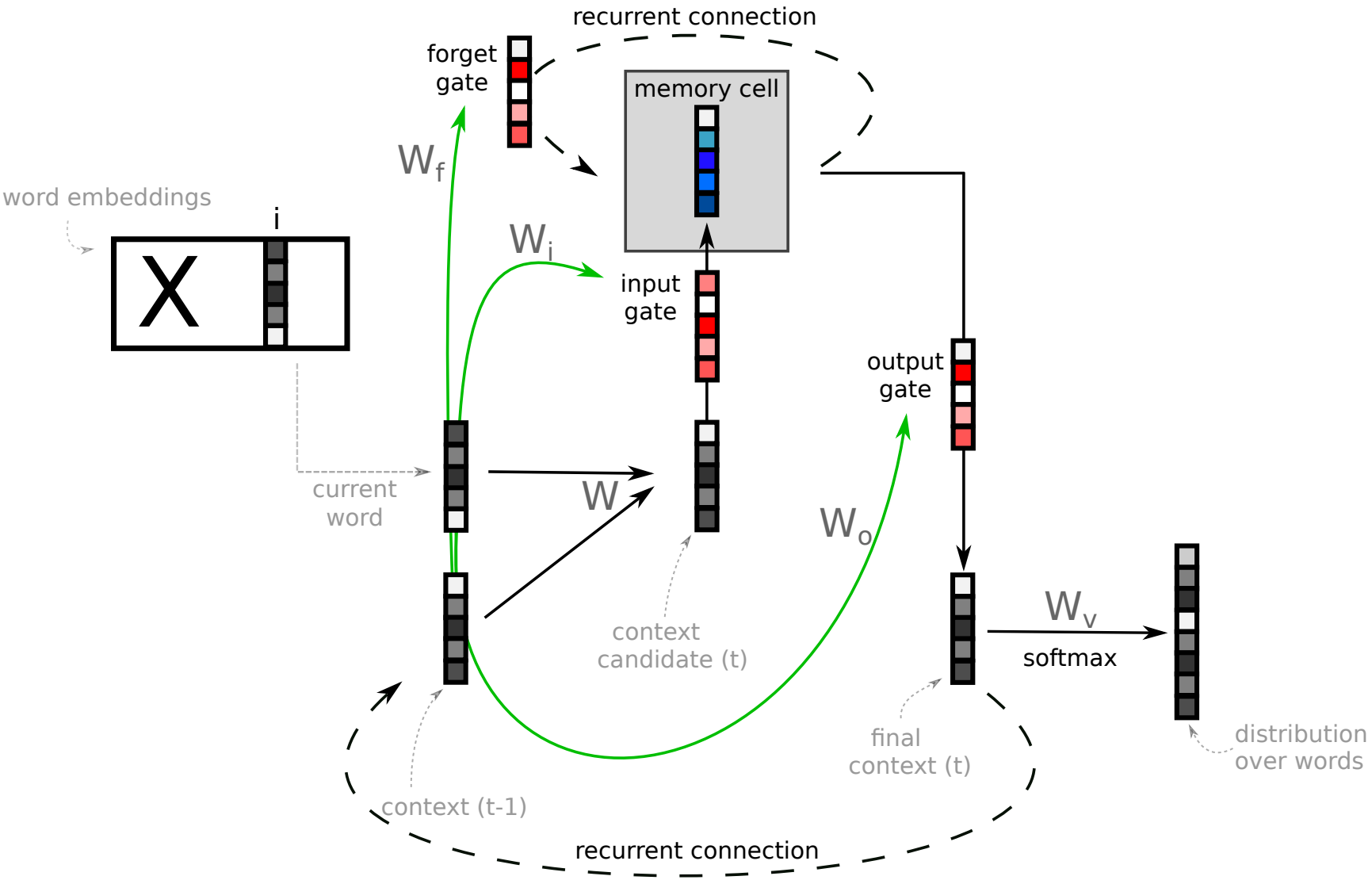
Graphical representation

RNNLM



Graphical representation

RNNLM with LSTM



Latent topic model

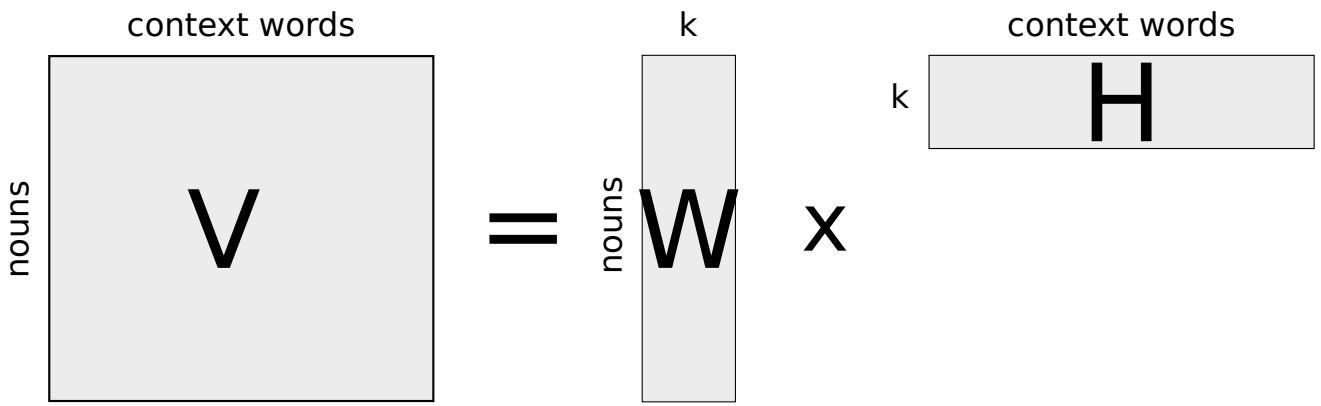
- Non-negative matrix factorization
- Given a matrix $\mathbf{V}$ (words $\times$ contexts), find non-negative matrix factors $\mathbf{W}$ and $\mathbf{H}$ such that:

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

- Choosing $k \ll n, m$ reduces data
- Constraint on factorization: all values in three matrices need to be *non-negative values* (≥ 0)

Latent topic model

Non-negative matrix factorization



Example

dim 14	dim 21	dim 57	dim 66
baai	astma	nek	beschrijft
dorpje	ziekte	buik	geschiedenis
rivier	aandoening	borst	roman
vallei	chronische	benen	schetst
natuurgebied	diabetes	vingers	schets
fietspad	infectie	heupen	gebeurtenissen
duinen	syndroom	keel	mythe
park	symptomen	tenen	belicht
dorp	depressie	spieren	fictie
dal	migraine	schouder	geschetst

Inserting topical information

- Neural network's output is a probability distribution
- NMF can be interpreted as probability distribution
- Probabilities can be combined according to each model's preferences
- **Sum of experts** (mixture model) and **product of experts** [Hinton 2002]
  - $P_{soe}(y_i) = \frac{1}{K} \sum_{k=1}^{K} P_k(y_i)$
  - $P_{poe}(y_i) = \frac{1}{Z} \prod_{k=1}^{K} P_k(y_i)$

Examples

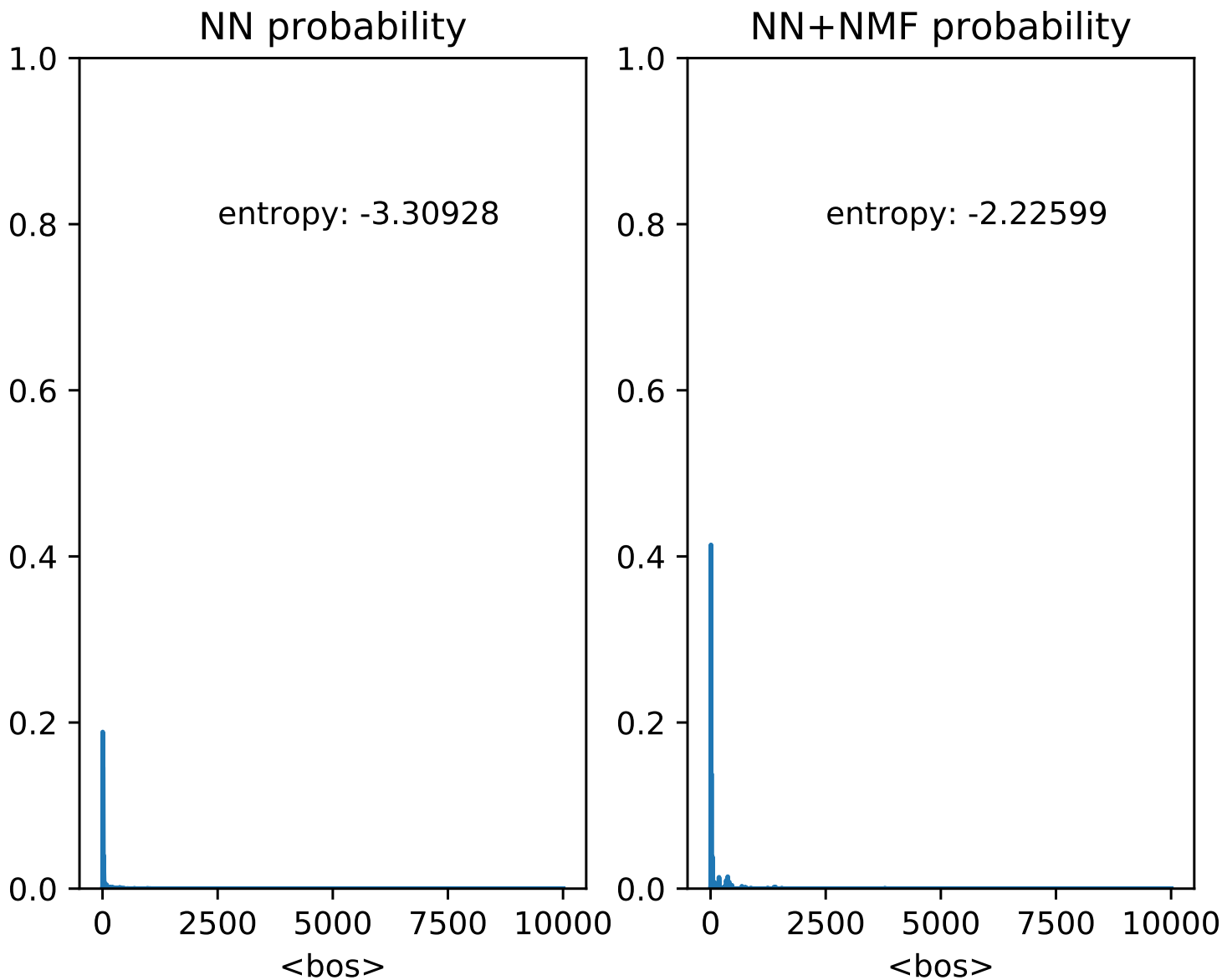
- Sum of experts

```
de vermaak uien leg taart eitjes in de warme oven
vlees dorst naar aanbevolen twee maal per dag droog en het is beton
ik tablet spul bouillon , dieet , en schep het gebakken brood
```

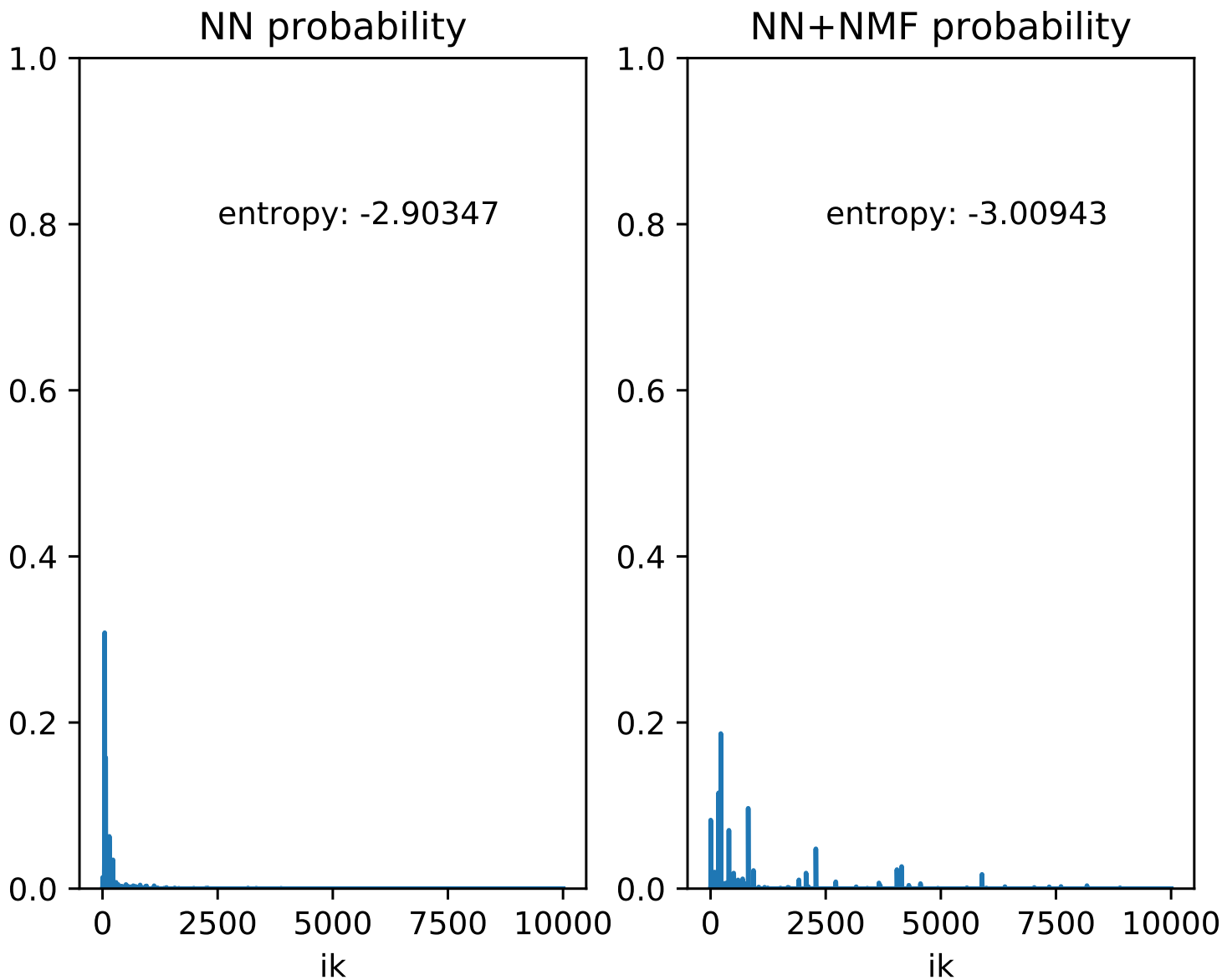
- Product of experts

```
het liefst kook ze lekker gewoon , en ze werken met verse groenten en
fruit
het stukje vlees of vis zit gewoon lekker en eten heerlijk
of het een beetje een zure saus of pasta saus , wordt pittig en pittig
gehakt
```

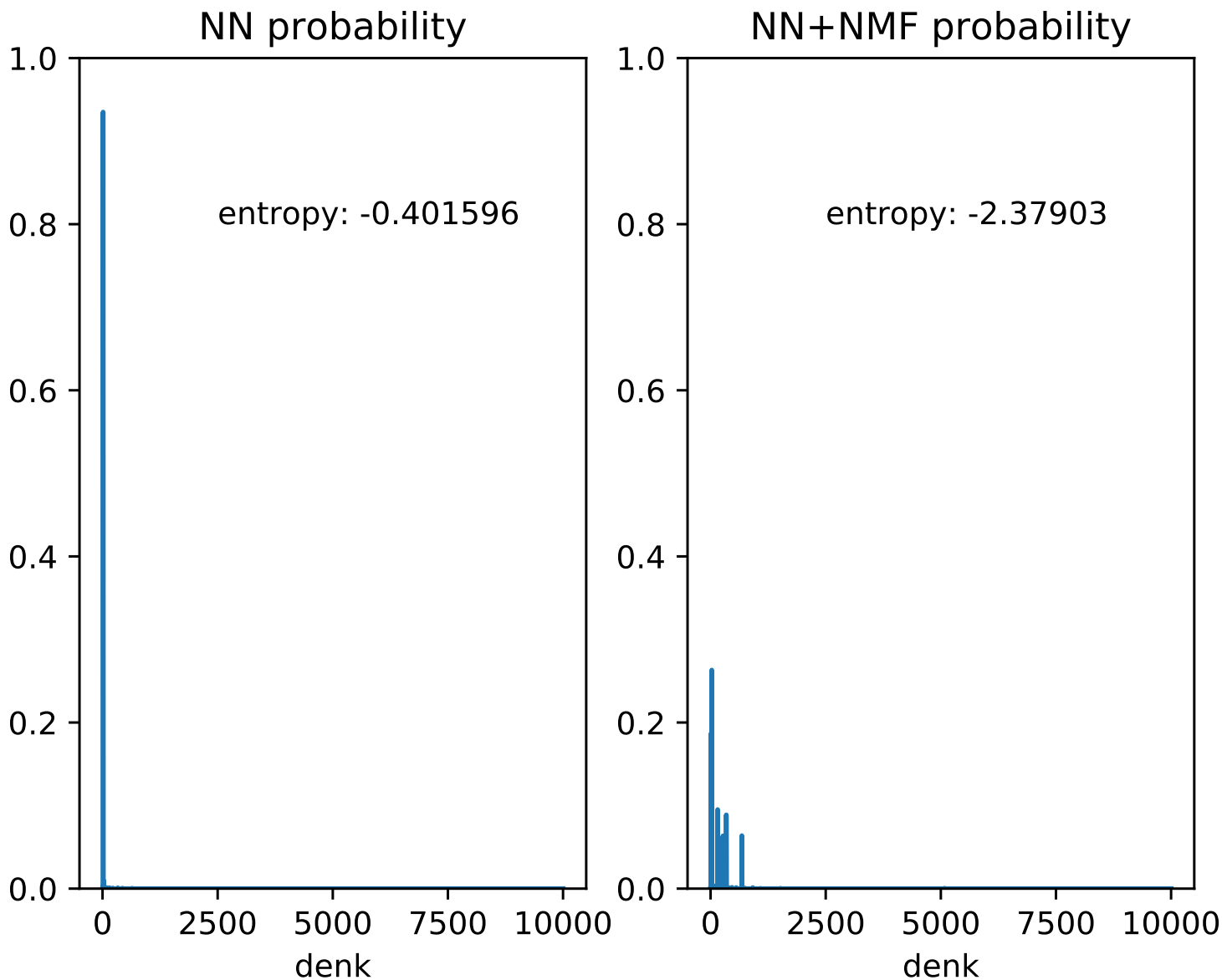
Word sequence probabilities



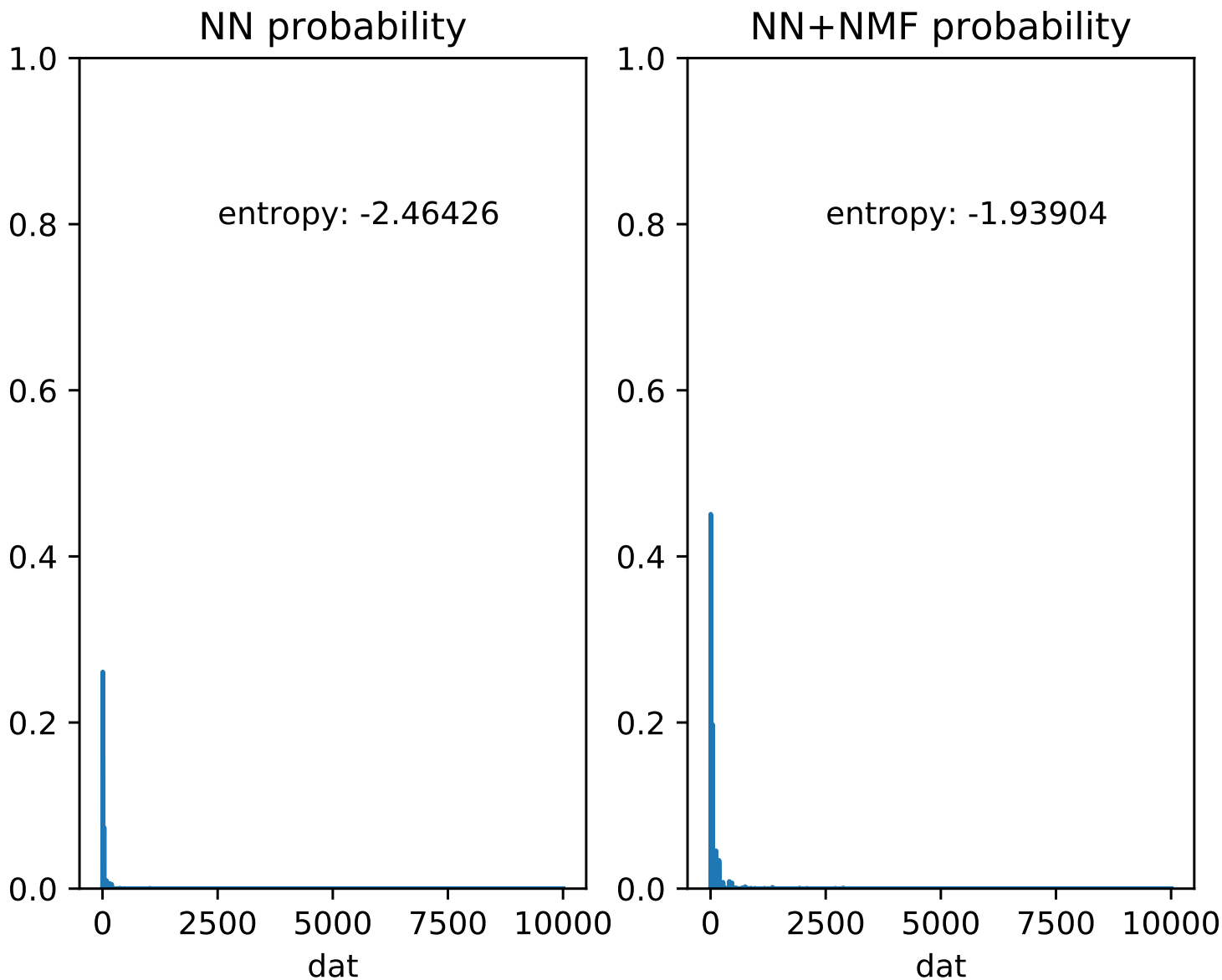
Word sequence probabilities



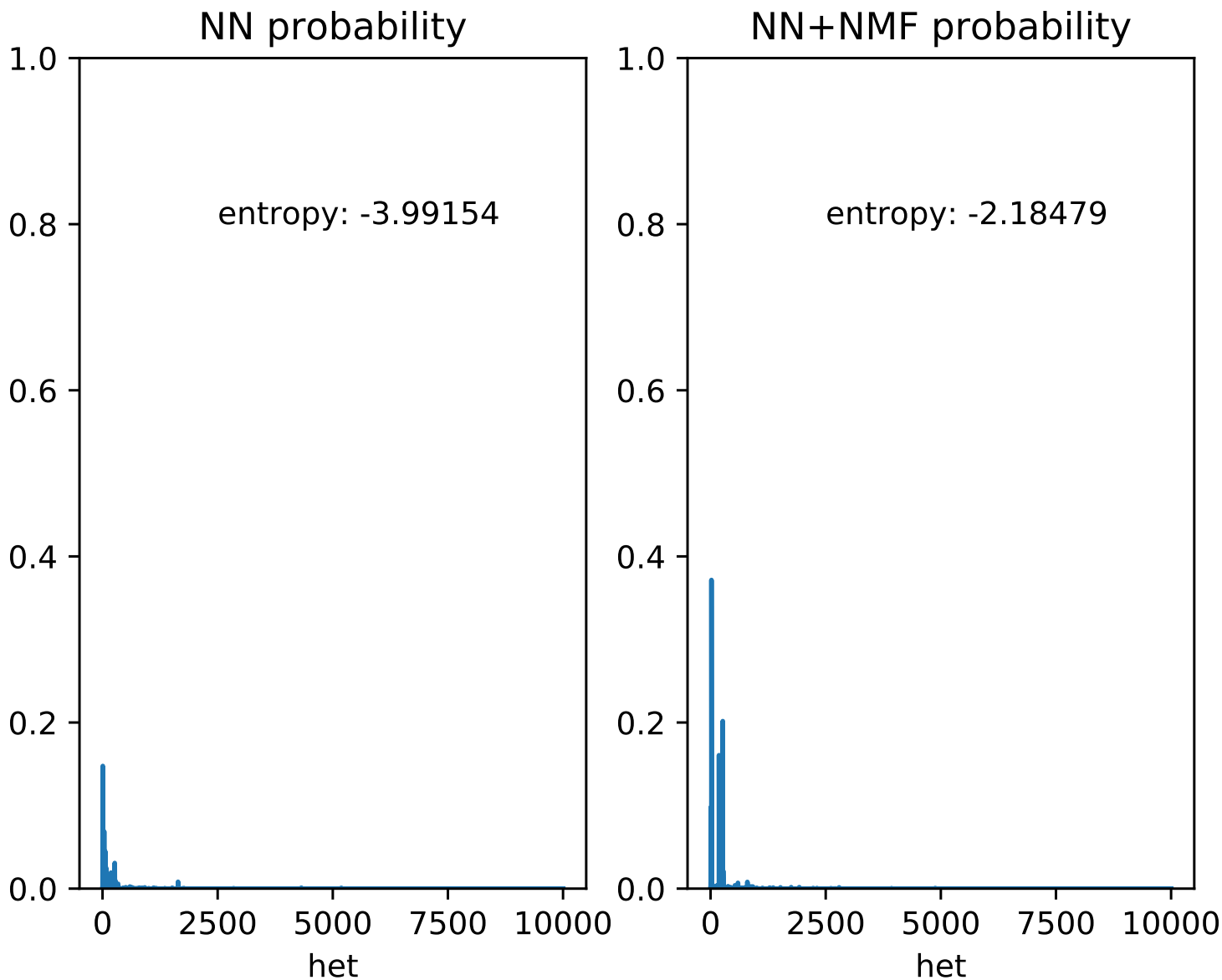
Word sequence probabilities



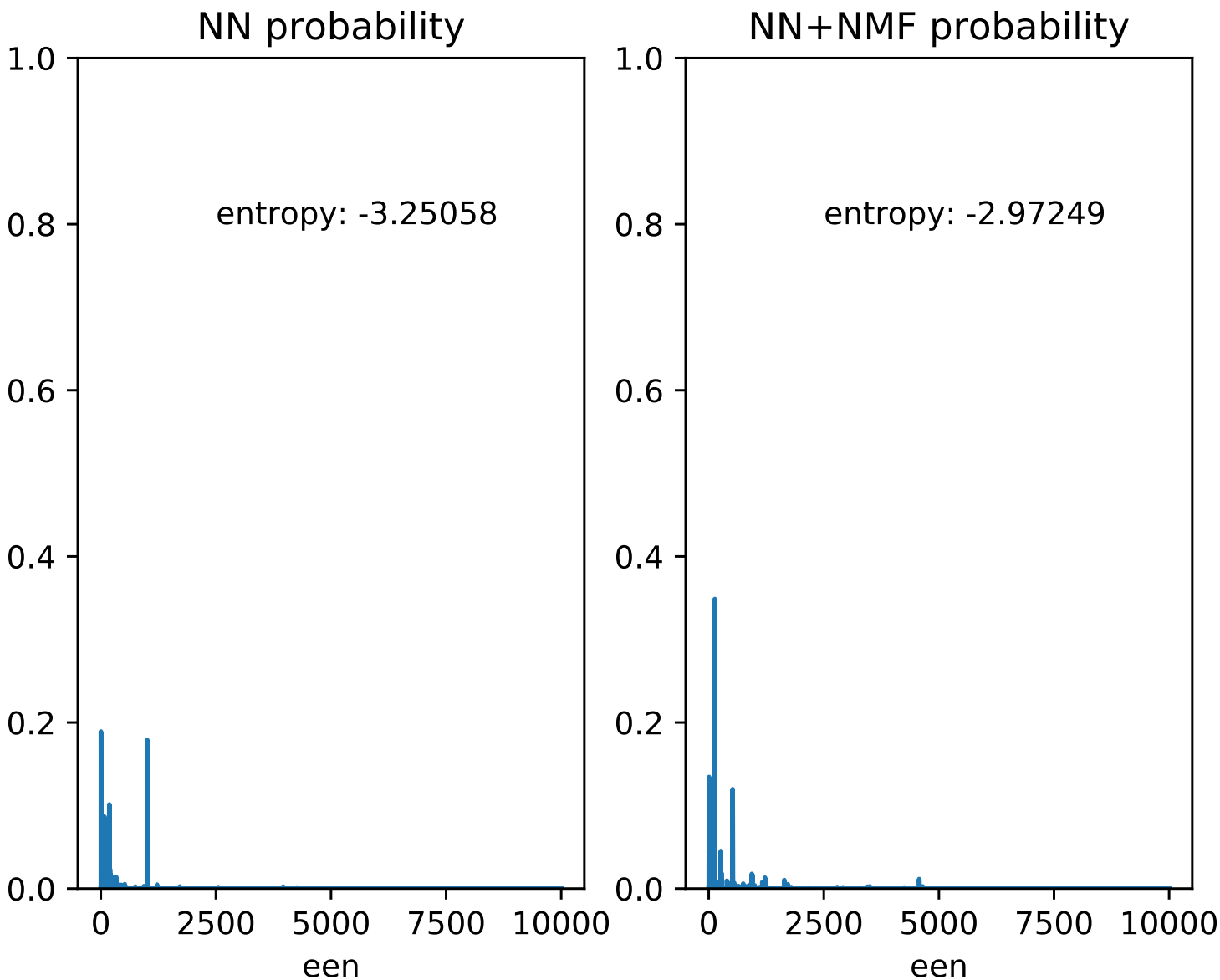
Word sequence probabilities



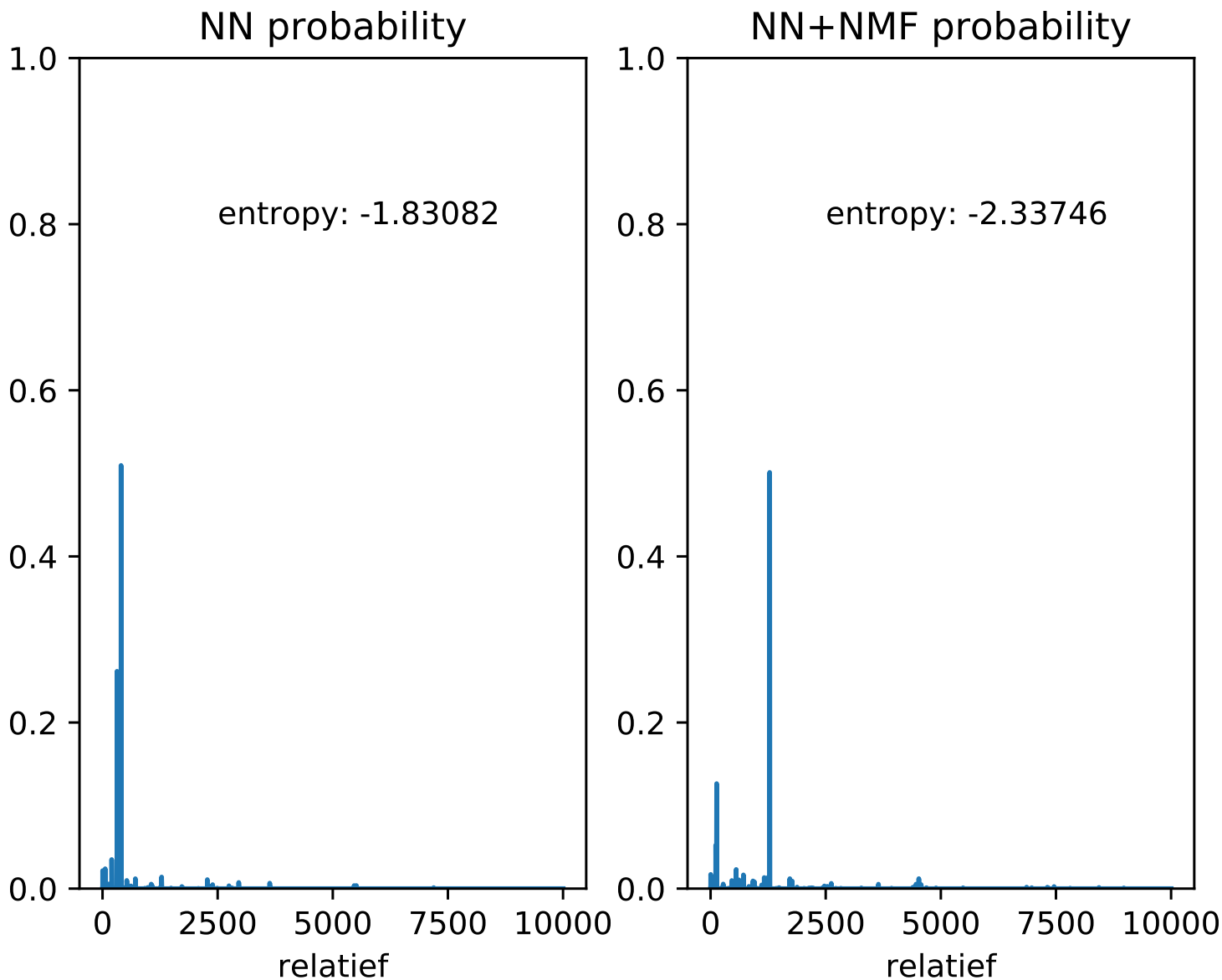
Word sequence probabilities



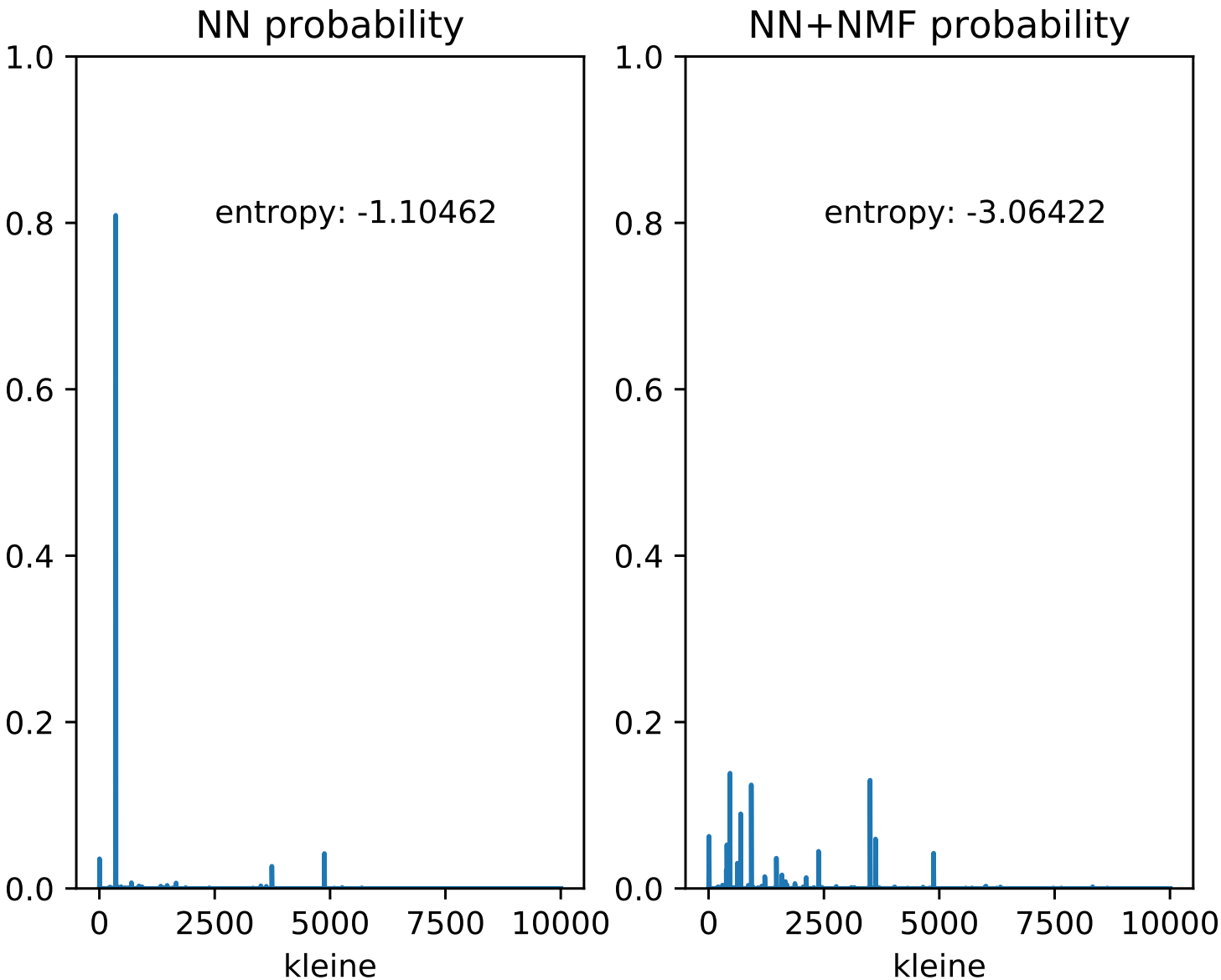
Word sequence probabilities



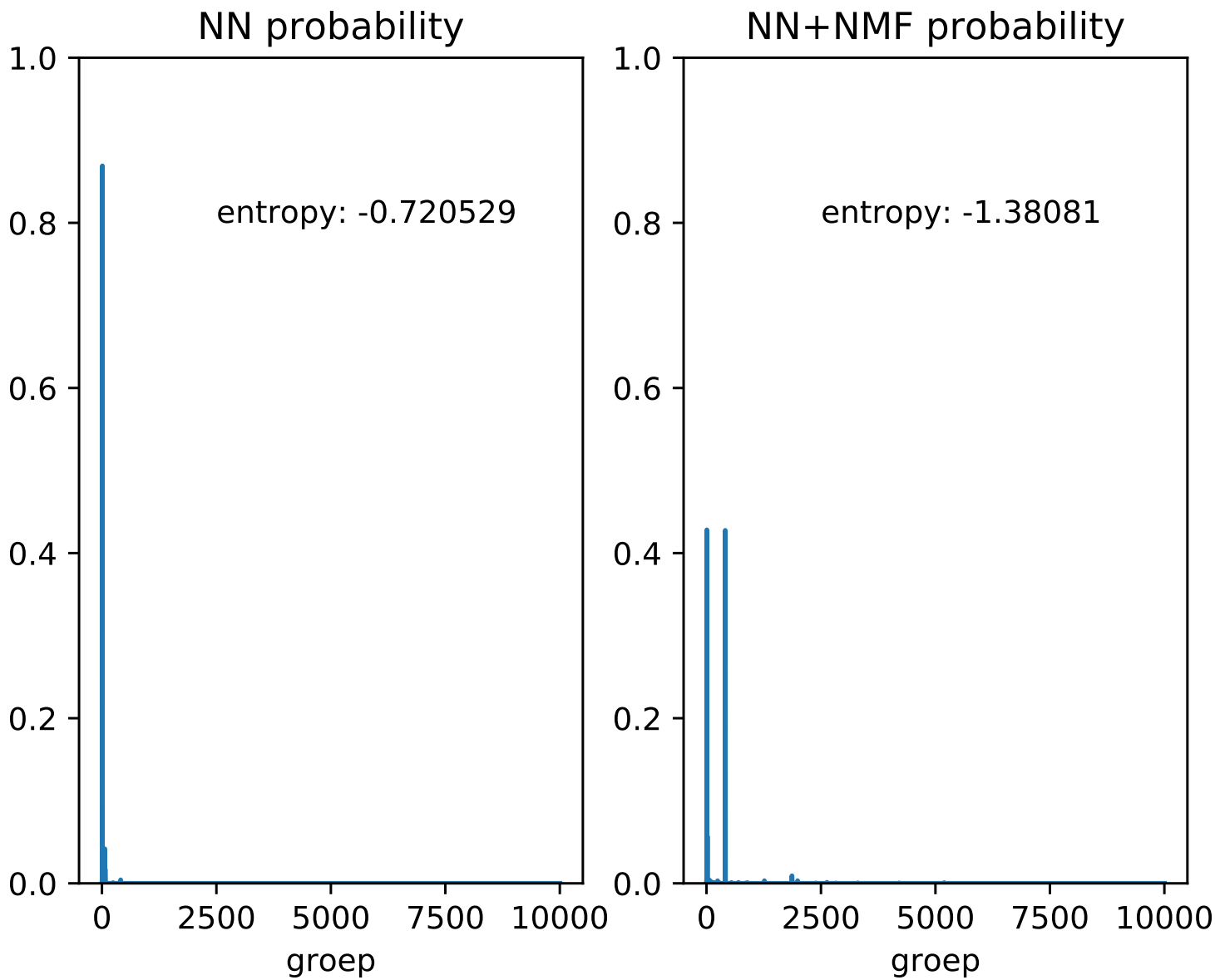
Word sequence probabilities



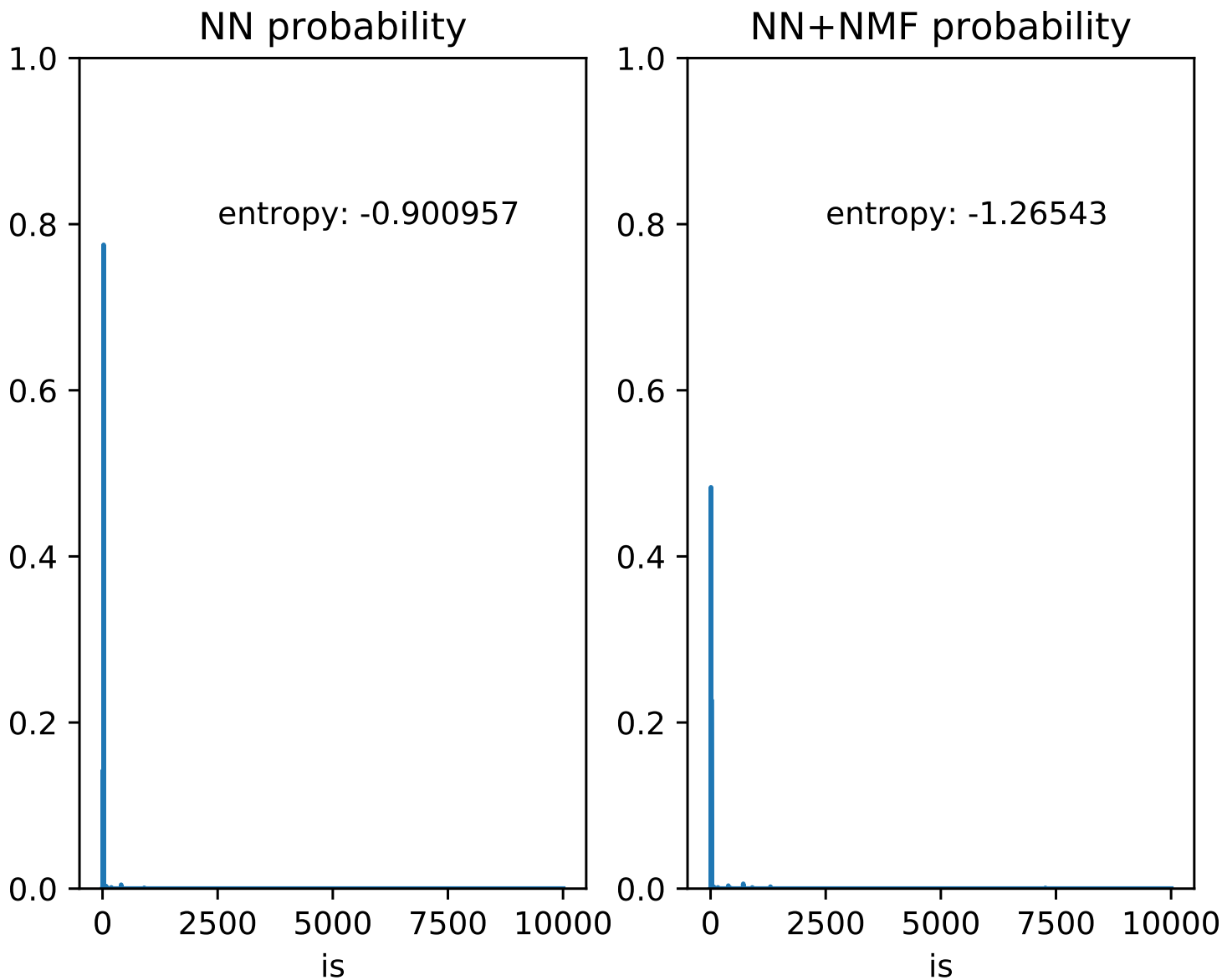
Word sequence probabilities



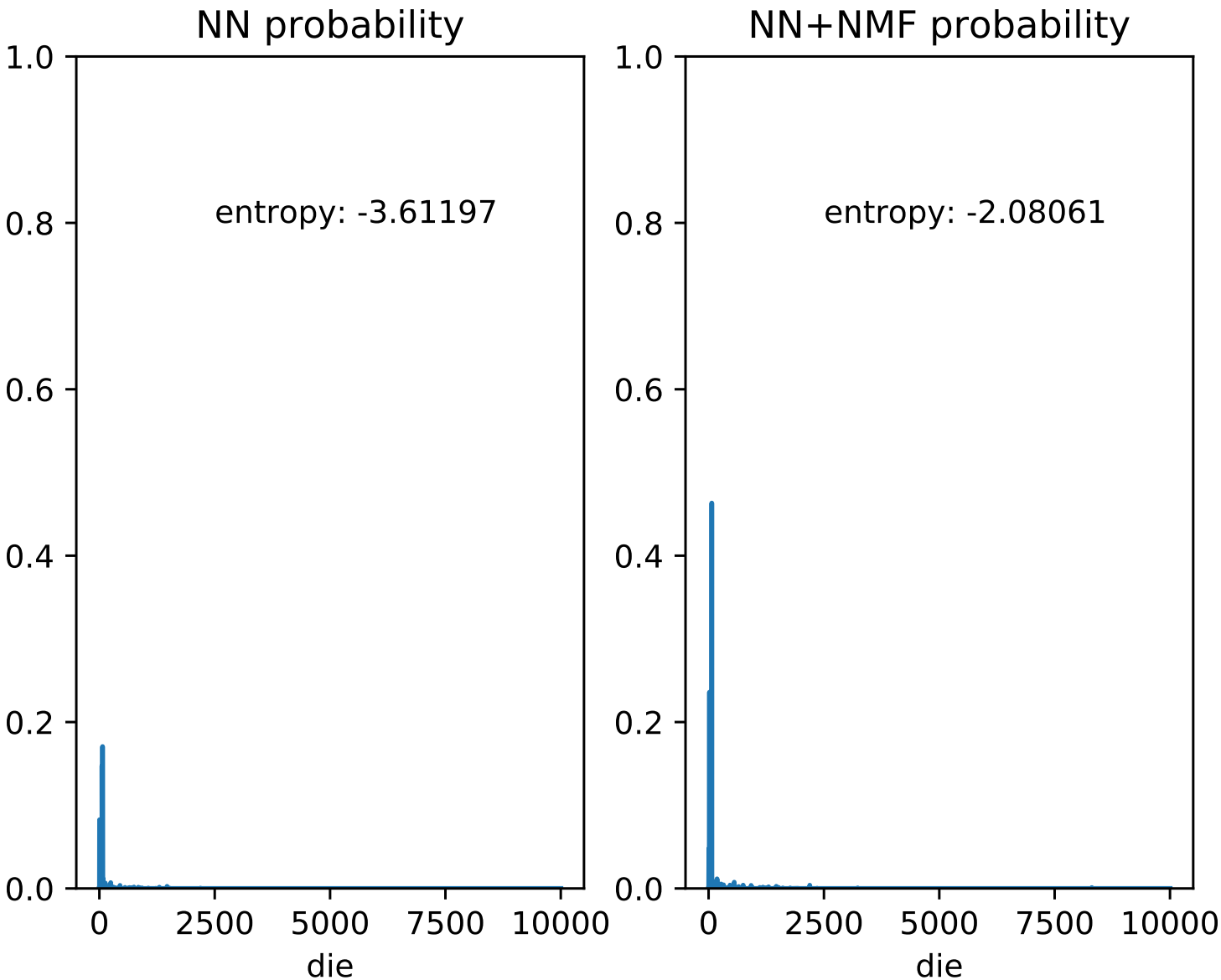
Word sequence probabilities



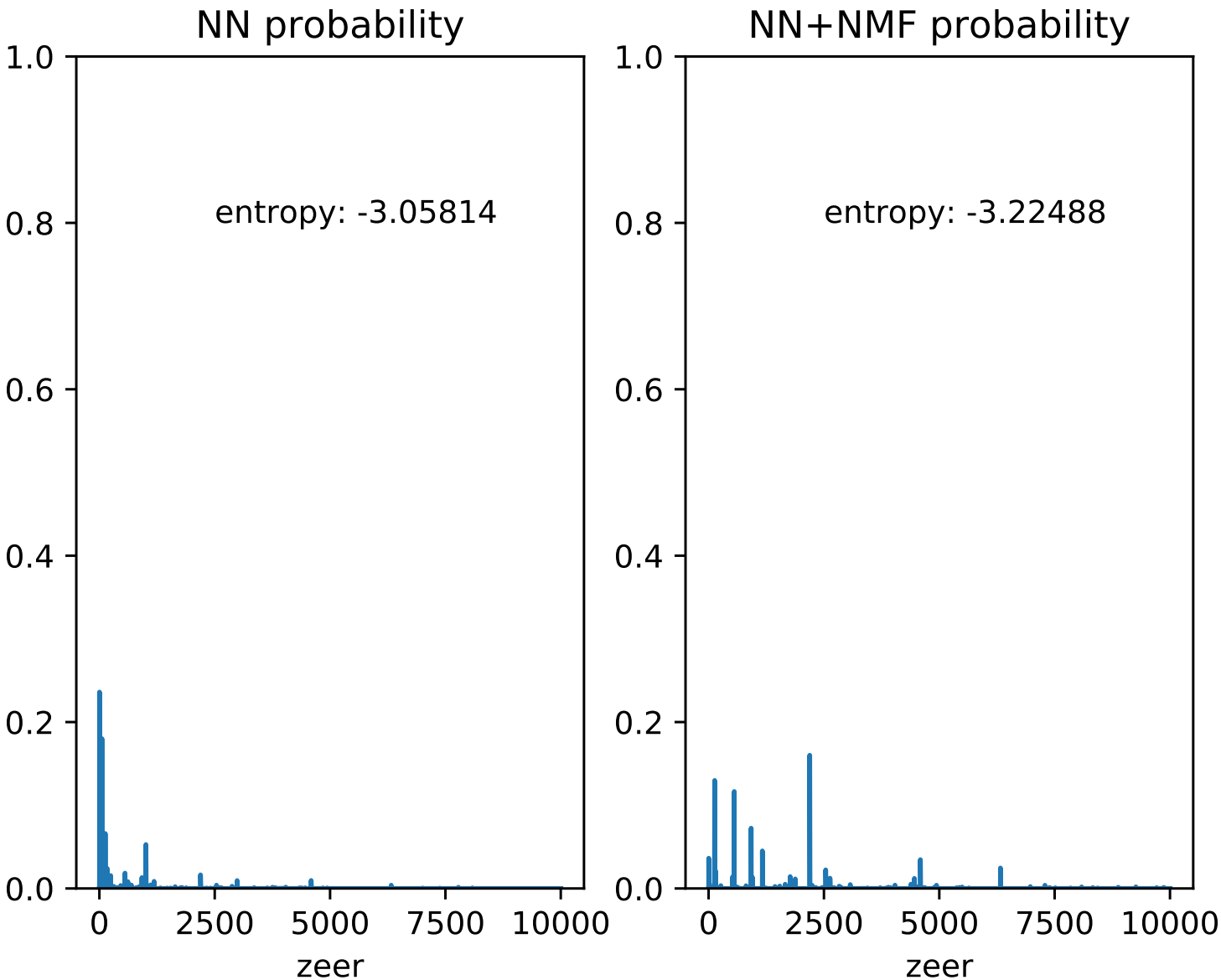
Word sequence probabilities



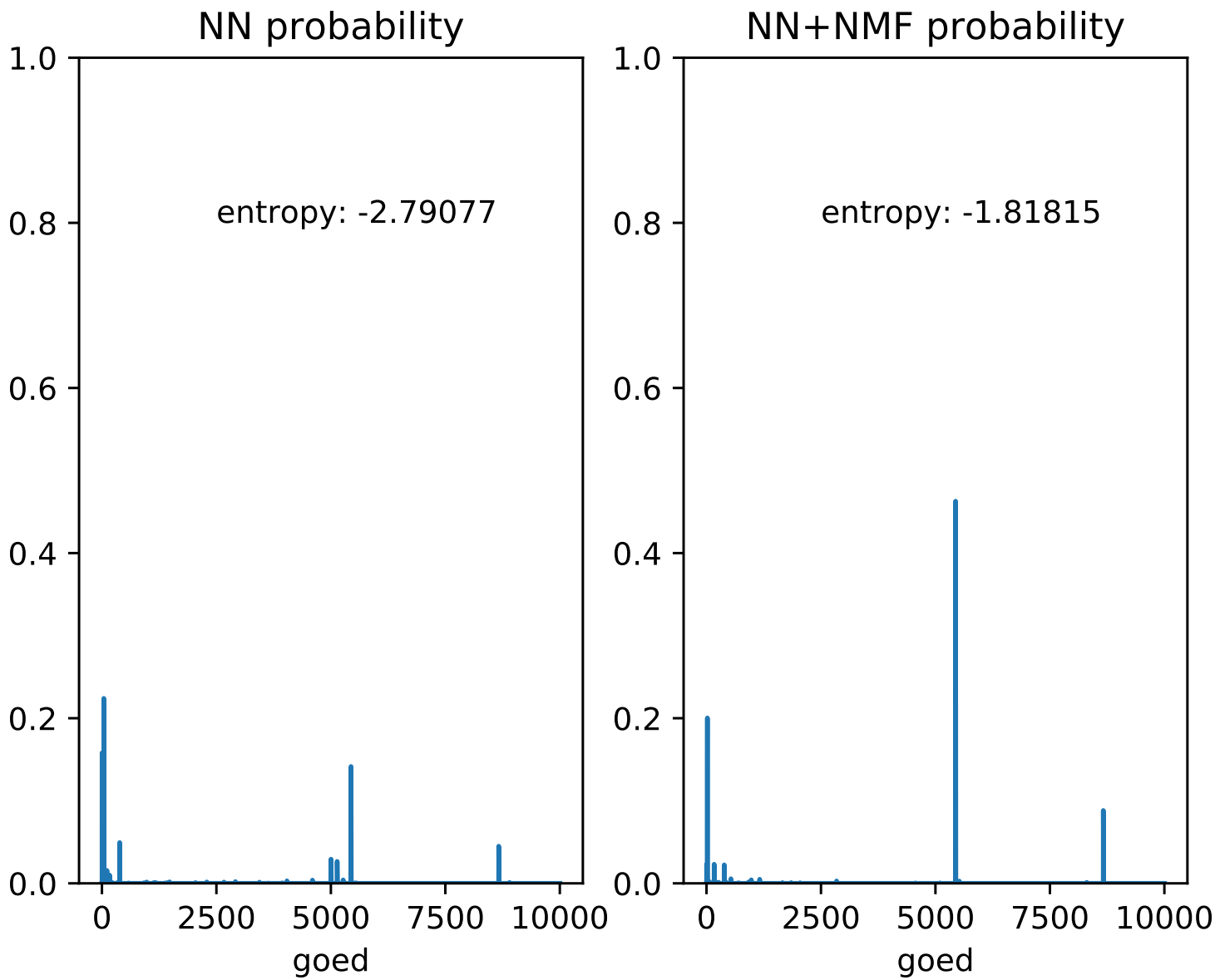
Word sequence probabilities



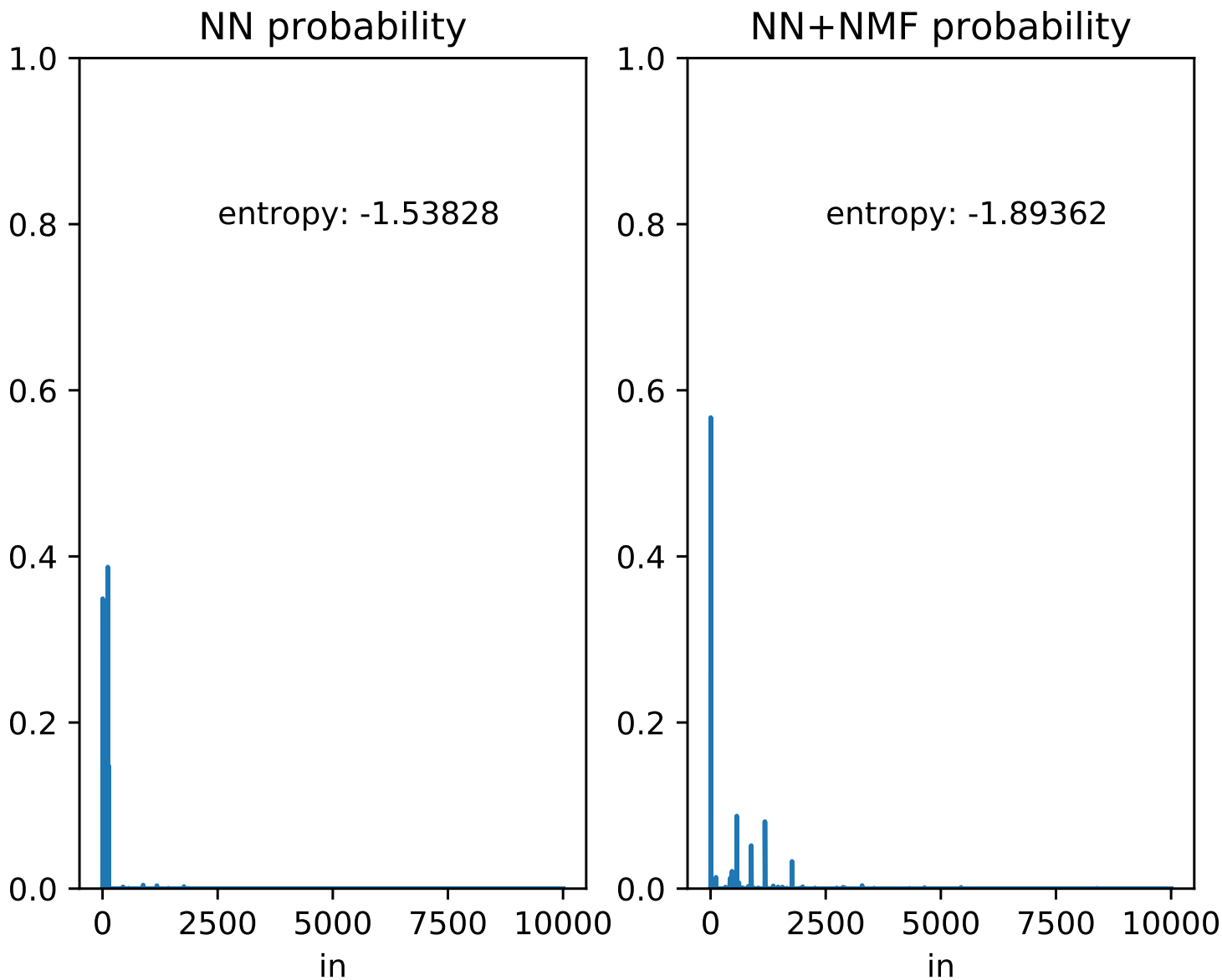
Word sequence probabilities



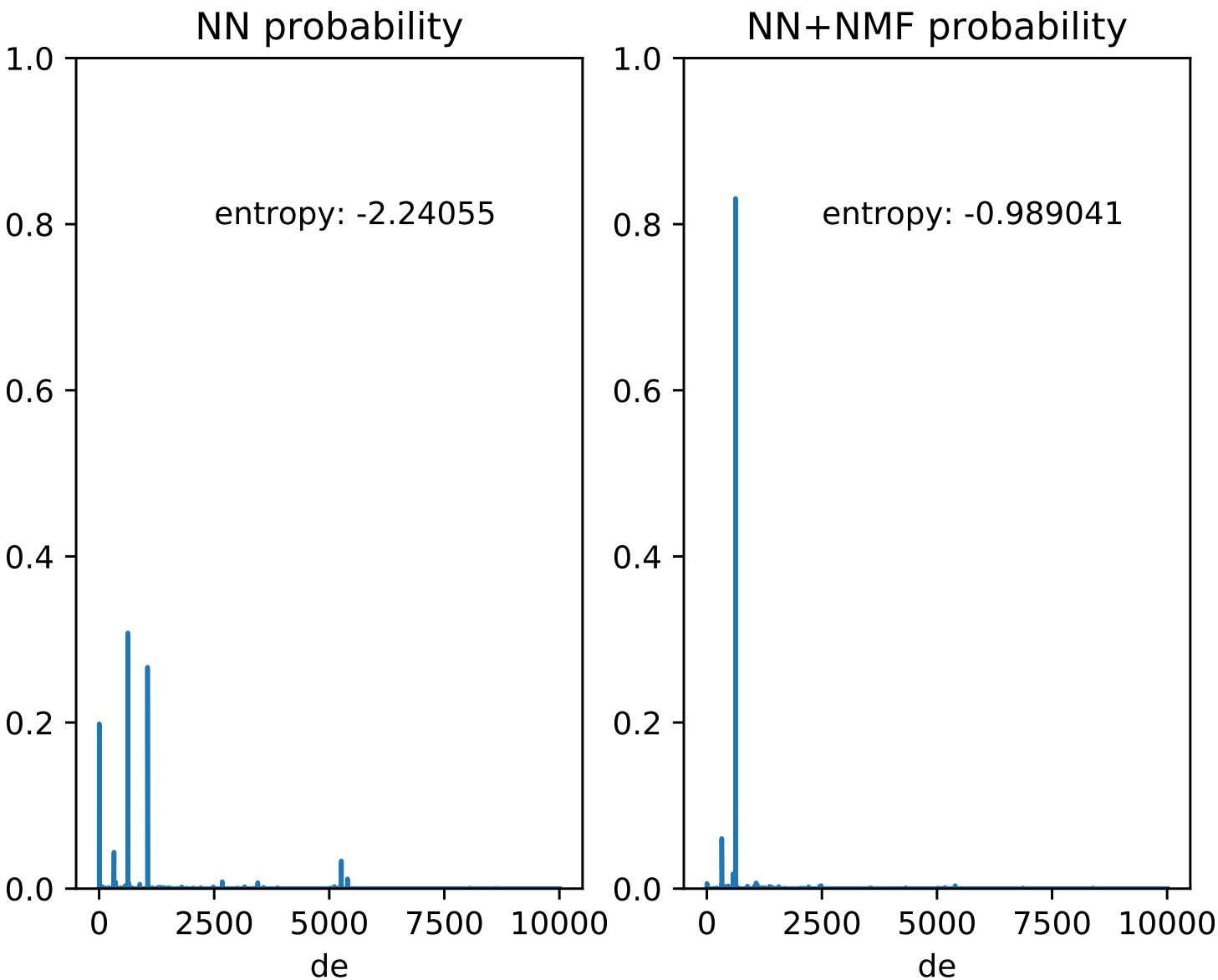
Word sequence probabilities



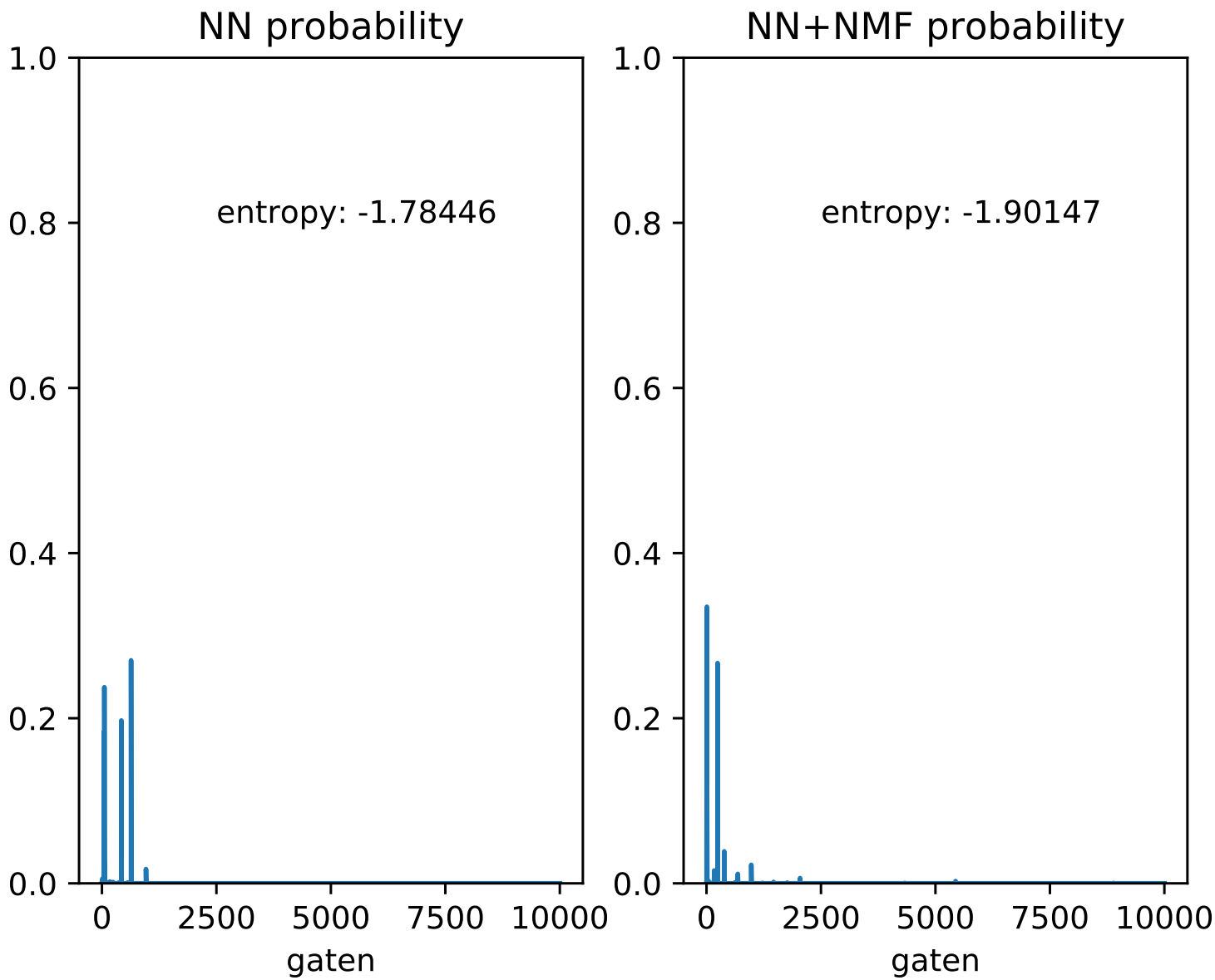
Word sequence probabilities



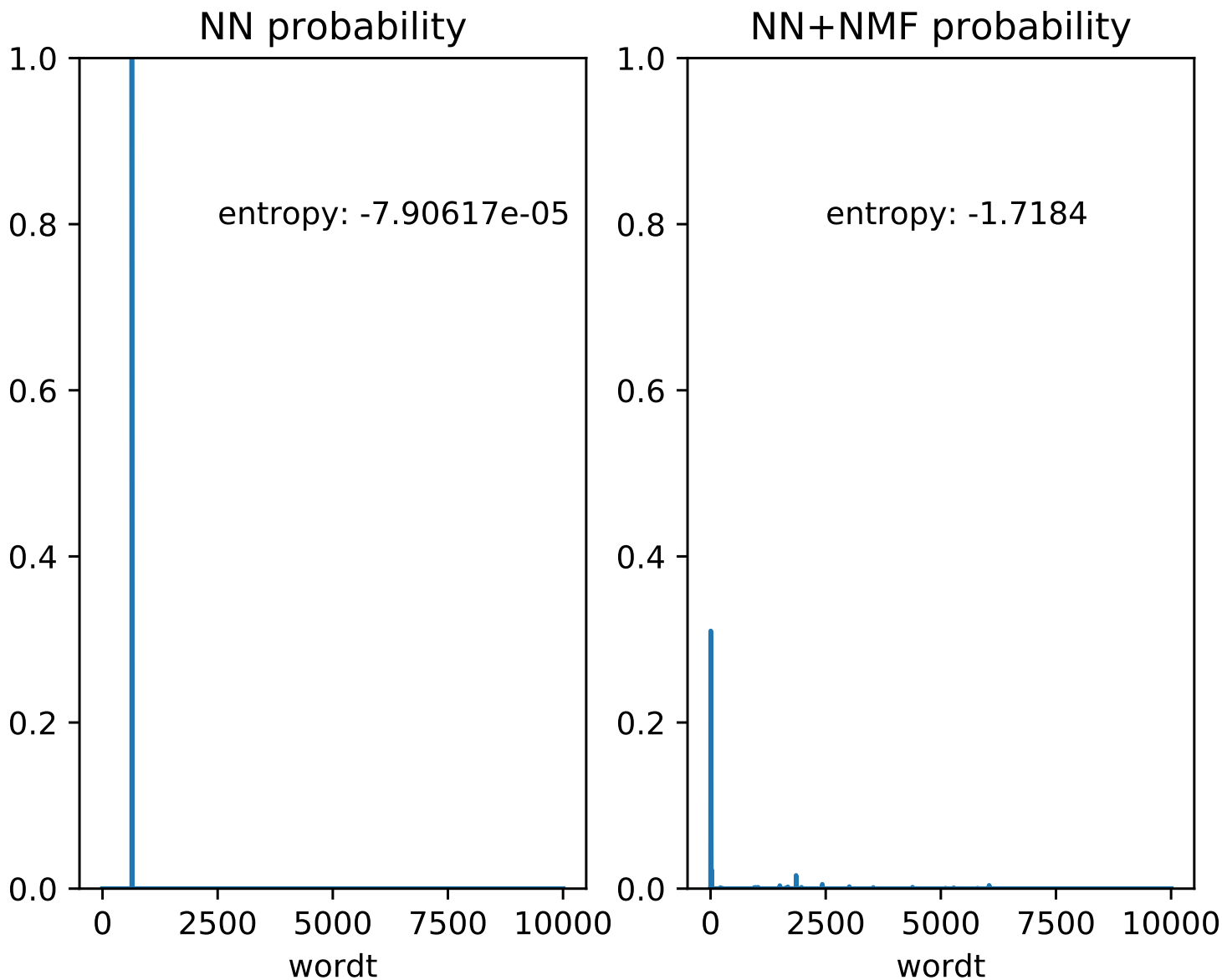
Word sequence probabilities



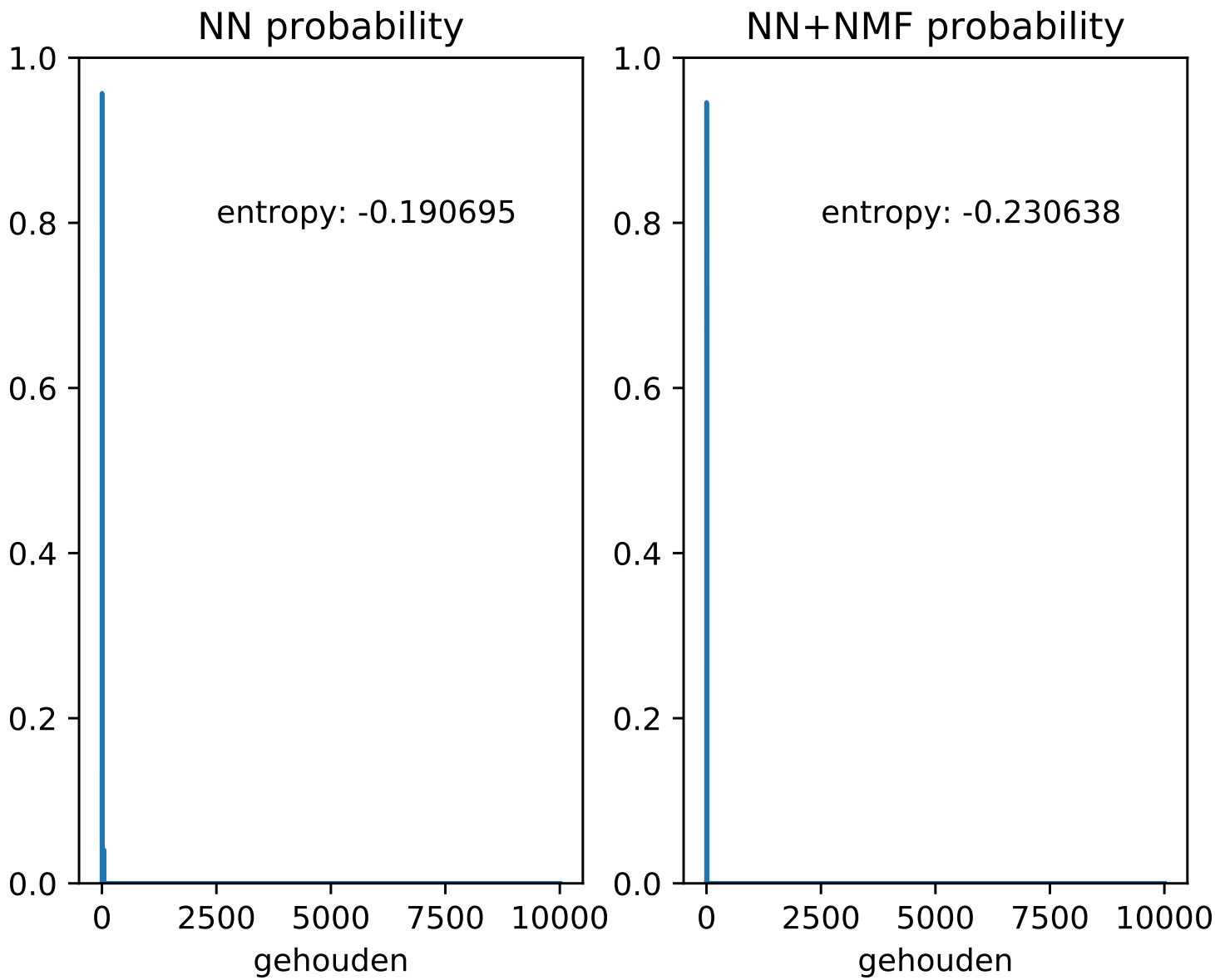
Word sequence probabilities



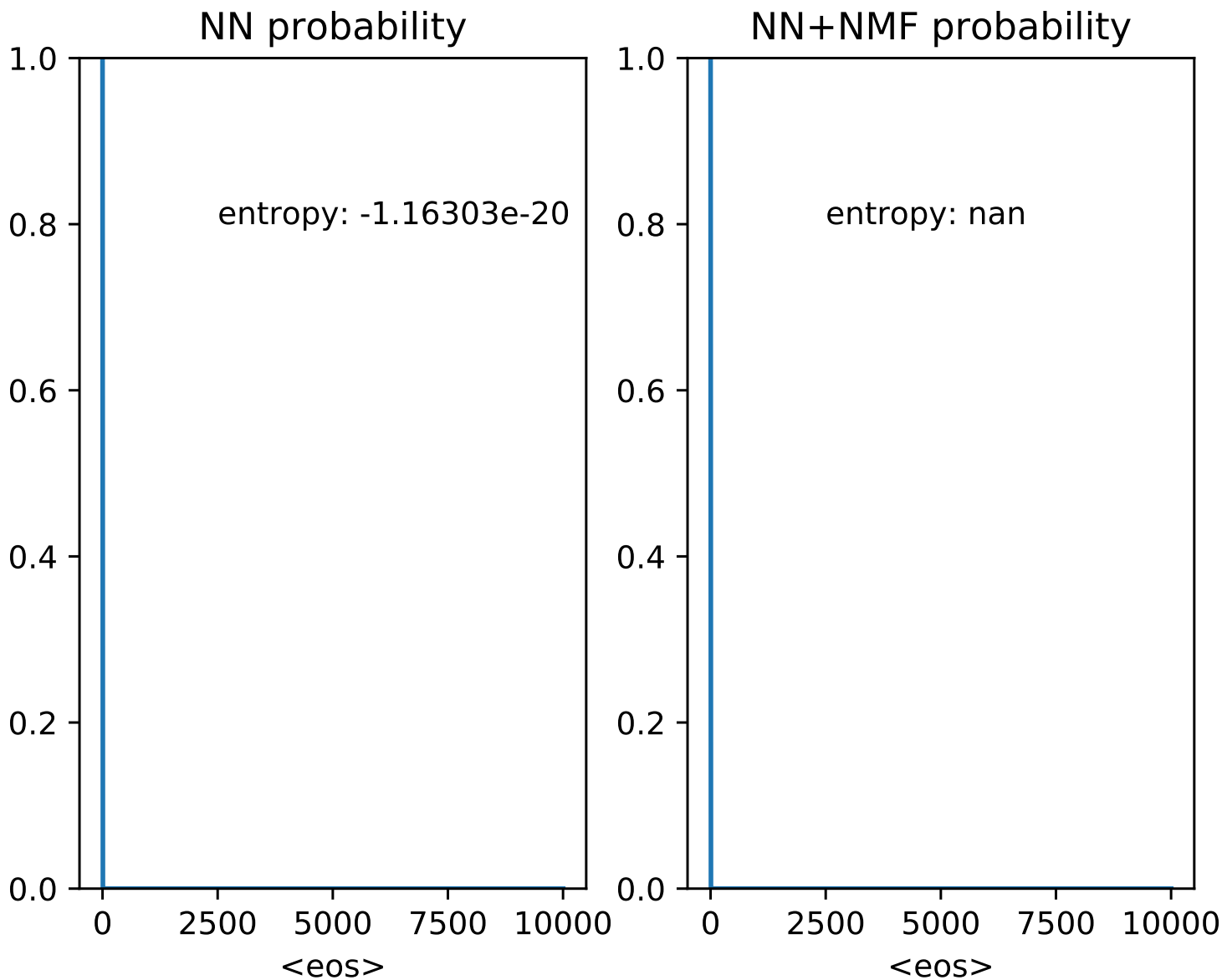
Word sequence probabilities



Word sequence probabilities



Word sequence probabilities



Entropy-mediated word sequence distribution

- When entropy of NN distribution is *low*, the NN knows the correct word choice in order to generate a well-formed sentence
- When entropy of NN distribution is *high*, use product of NN expert and NMF expert in order to insert relevant topicality
- Entropy threshold θ set experimentally

Implementation details

- Models trained on $\pm$ 100 million words of web text (NLCOW)
- Vocabulary: 10K words
- NN: 2-layer LSTM, embedding size: 512, hidden layer size: 2048
- Batch stochastic gradient descent with Nesterov momentum
- Learning rate annealed on validation data
- NMF: 2000 context words, sentence window, 100 dimensions
- Entropy threshold $\theta := 1.5$

Example ('political' dimension)

met name omdat wij voor onze fractie juist duidelijk willen maken dat marktwerking geen trend is ten koste van

dat betekent dat alle partijen er met elkaar voor moeten zorgen dat de politici zich kritisch zullen opstellen

zijn enige partij om kiezers te winnen , sp stemt tegen pvv en wilders gaat voor het cda

met name dat laatste is voor mij onbegrijpelijk en lijkt mij een verstandige zet van leefbaar om te overwegen

omdat mensen zich op dat moment schamen voor het feit dat zij geen moslim zijn

en daarmee zijn de verschillen tussen de partijen duidelijker geworden

Example (‘policing dimension’)

de man weigert het slachtoffer te waarschuwen , waarna hij opgeroepen wordt om door de politie te worden gebeld

het was zijn eerste slachtoffer die hij had gedood tijdens een poging tot zelfmoord

het is de taak van de politie om de slachtoffers te beschermen en hen bescherming te bieden

de politie waarschuwt terroristen die zich verzetten tegen de terugkeer van turkse troepen

zij zijn bang dat de politie hen zelf verdacht maakt en illegale prostitutie gebruikt

door zijn gedrag is het gerechtvaardigd dat de dader zich schuldig voelt tijdens de ruzie

Evaluation

- Two aspects to evaluate: **syntactic fluency** and **topical coherence**
- This evaluation: focus on first aspect
- Generate 10k sentences for each model
- Compute perplexity scores using standard trigram model (Kneser-Ney smoothed, trained on 1 billion words)

model	perplexity
LSTM	56.79
Sum of Experts	7791.15
Product of Experts	134.58
entropy ($\theta = 1.5$)	97.71
entropy ($\theta = 3.0$)	66.91

Conclusion

- NN distribution is good at capturing general sentence structure, but does not know what to talk about
- NMF distribution knows nothing about syntax, but knows what subject to talk about
- Product of experts works rather well in order to combine both distributions, but still muddles with syntactic fluency
- By tracking the neural network distribution's entropy, we can determine when syntax necessitates giving priority to NN distribution
- Future work
  - Evaluation of topical coherence
  - Explore neural network architectures that incorporate topical coherence


Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. Neural computation 14(8): 1771–1800. 2002.


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.