p(conclusions|Skipping {*2*})

Bayesian Language Modelling with Skipgrams

LANGUAGE MACHINES

CLS | Centre for Language Studies
Radboud University

Bayesian Language Modelling with Skipgrams

Louis Onrust
Centre for Language Studies, Radboud University
Center for Processing Speech and Images, KU Leuven

l.onrust@let.ru.nl
github.com/naiaden

Language Models

Applications

- Input assists on telephones
- Automatic translation of search results
- Digital court reporting

Language Models

Applications

- Input assists on telephones
- Automatic translation of search results
- Digital court reporting

Flavours

- Frequentist language models
- Bayesian language models
- Neural language models
- ...

The Task at Hand

After all , tomorrow is another [. . .]

The Task at Hand

After all , tomorrow is another [. . .]

Word prediction **Word probability** **Pattern probability**

The Task at Hand

After all , tomorrow is another [. . .]

Word prediction	**Word probability**	**Pattern probability**
• day		

The Task at Hand

After all , tomorrow is another [. . .]

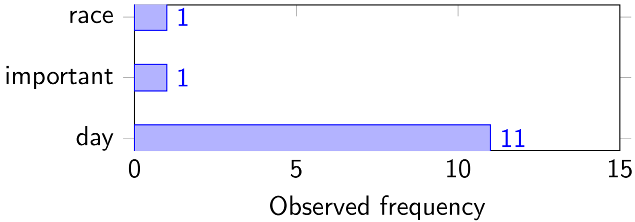
Word prediction

- day

Word probability

- cow
- day
- vegetable
- . . .

Pattern probability



The Task at Hand

After all , tomorrow is another [. . .]

Word prediction

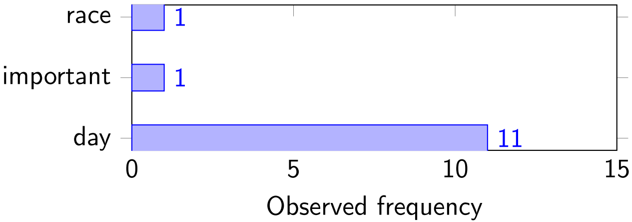
- day

Word probability

- cow
- day
- vegetable
- . . .

Pattern probability

- tomorrow is another day
- tomorrow is another important
- tomorrow is another race



Generalising the *n*-gram

***n*-grams**

- Continuous sequence of *n* words

Skipgrams

- *n*-gram with at most $n - 2$ skips of length 1

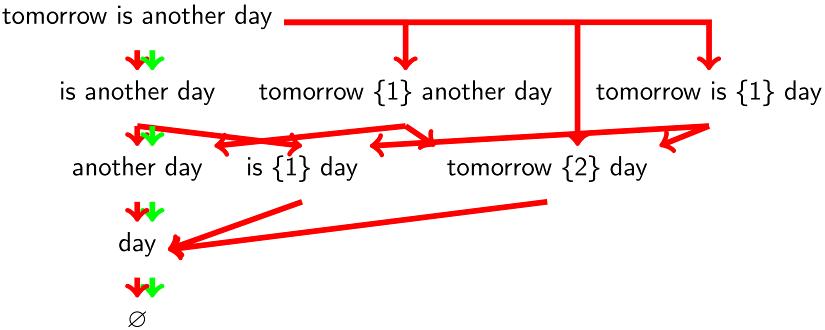
Flexgrams

- *n*-gram with any number of skips of any length

Skipgrams in RNN

- Based on embeddings rather than co-occurrence
- Co-occurrence of embeddings

Backoff Patterns with Skipgrams



Probability Estimation

Maximum Likelihood Estimate

$$p_{\mathrm{ML}}(w_i | w_{i-N+1}, \ldots, w_{i-1}) = \frac{C(w_{i-N+1}, \ldots, w_{i-1})}{C(w_{i-N+1}, \ldots, w_i)}$$

- Parameter estimation is impossible for $N > 2$
- Naïve priors assuming independent parameters fail as well

Probability Estimation

Maximum Likelihood Estimate

$$p_{\mathrm{ML}}(w_i | w_{i-N+1}, \ldots, w_{i-1}) = \frac{C(w_{i-N+1}, \ldots, w_{i-1})}{C(w_{i-N+1}, \ldots, w_i)}$$

- Parameter estimation is impossible for $N > 2$
- Naïve priors assuming independent parameters fail as well

Smoothing

$$p_{\mathrm{SM}}(w_i | w_{i-N+1}, \ldots, w_{i-1}) = \sum_{n=1}^{N} \lambda(n) Q_n(w_i | w_{i-N+1}, \ldots, w_{i-1})$$

- Chen and Goodman found that interpolated and modified Kneser-Ney are best under virtually all circumstances

Bayesian Probability Estimation

Parametrise conditional probabilities

$$p(w_i = w | w_{i-N+1}, \dots, w_{i-1} = u) = G_u(w)$$

$$G_u = [G_u(w)]_{w \in W}$$

$$\pi(w_{i-N+1}, \dots, w_{i-1}) = w_{i-N+2}, \dots, w_{i-1}$$

- G_u is a probability vector associated with context u

Bayesian Probability Estimation

Parametrise conditional probabilities

$$p(w_i = w | w_{i-N+1}, \ldots, w_{i-1} = u) = G_u(w)$$

$$G_u = [G_u(w)]_{w \in W}$$

$$\pi(w_{i-N+1}, \ldots, w_{i-1}) = w_{i-N+2}, \ldots, w_{i-1}$$

- G_u is a probability vector associated with context u

Hierarchical Dirichlet language model

- What is $p(G_u | G_{\pi(u)})$?
- Standard Dirichlet distribution over probability vectors: does not outperform ikn and mkn (MacKay and Peto, 1994)

Bayesian Probability Estimation

Parametrise conditional probabilities

$$p(w_i = w | w_{i-N+1}, \ldots, w_{i-1} = u) = G_u(w)$$

$$G_u = [G_u(w)]_{w \in W}$$

$$\pi(w_{i-N+1}, \ldots, w_{i-1}) = w_{i-N+2}, \ldots, w_{i-1}$$

- G_u is a probability vector associated with context u

Hierarchical Dirichlet language model

- What is $p(G_u | G_{\pi(u)})$?
- Standard Dirichlet distribution over probability vectors: does not outperform ikn and mkn (MacKay and Peto, 1994)

Hierarchical Pitman-Yor process

- Two-parameter extension of the Dirichlet distribution
- PYP produces power-law distributions
- Outperforms ikn and mkn (Teh, 2006)

Everlasting feud

Frequentists

- Unconditional perspective: inferential methods should give good answers in repeated use
- "pessimist": let's protect ourselves against bad decisions given that our inferential procedure is inevitably based on a simplification of reality
- $R(\theta) = \mathbb{E}_\theta l(\delta(X), \theta)$

Bayesians

- Conditional perspective: inferences should be made conditional on the current data
- "optimistic": let's make the best possible use of our sophisticated inferential tool
- $\rho(X) = \mathbb{E}[l(\delta(X), \theta)|X]$

Comparing HPYLM to MKN: the Numbers

The reported values are entropy values

HPYLM

	jrc	1bw	emea	wp
jrc	3.65	10.56	10.08	10.34
1bws	9.58	7.31	9.89	8.94
emea	9.59	10.60	1.88	10.10
wps	9.12	8.83	9.97	7.76

Modified Kneser-Ney

jrc	1bw	emea	wp
4.43	10.70	10.35	10.84
10.51	7.72	10.57	9.96
10.12	10.77	3.28	10.70
9.72	9.23	10.63	8.81

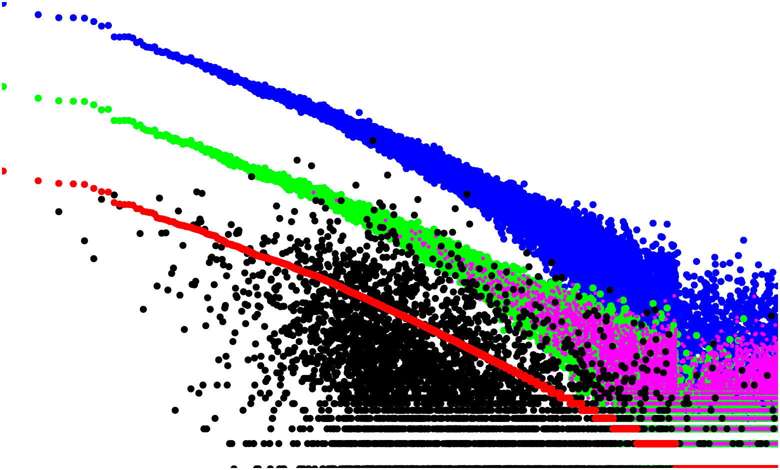
Relative reduction in entropy (in %)

jrc	17.66	1.31	2.67	4.67
1bws	8.82	5.37	6.44	10.19
emea	5.30	1.57	42.51	5.65
wps	6.20	4.28	6.23	11.92

Relative reduction in perplexity (in %)

41.87	9.25	17.43	29.63
47.41	24.99	37.62	50.51
31.04	11.08	61.92	34.27
34.14	23.95	36.81	51.72

Comparing HPYLM to MKN: the Figure



Adding skipgram features alongside *n*-grams

HPYLM

	jrc	1bw	emea	wp
jrc	3.65	10.22	9.91	9.98
1bws	9.58	7.31	9.89	8.94
emea	9.23	10.16	1.88	9.72
wps	9.12	8.83	9.97	7.76

Modified Kneser-Ney

jrc	1bw	emea	wp
3.68	10.18	9.87	9.98
9.55	7.34	9.85	8.98
9.18	10.17	1.89	9.72
9.14	8.88	9.95	7.82

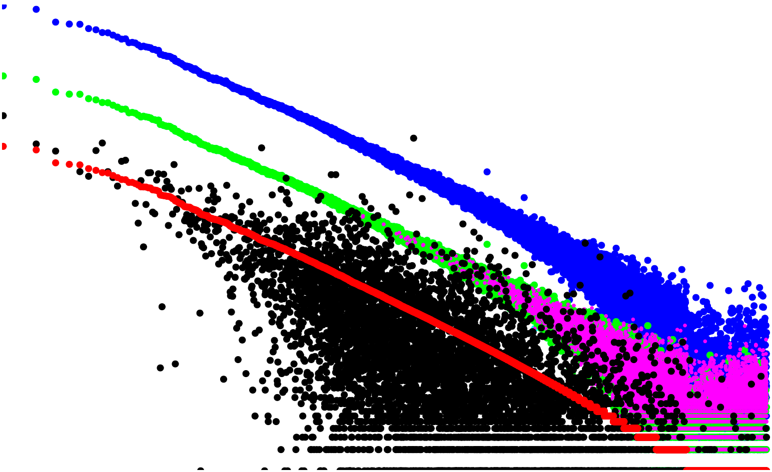
Relative reduction in entropy (in %)

jrc	-0.81	0.40	0.34	0.03
1bws	0.34	-0.47	0.39	-0.45
emea	0.51	-0.15	-0.41	0.01
wps	-0.31	-0.53	0.21	-0.82

Relative reduction in perplexity (in %)

-2.07	2.80	2.3	0.23
2.23	-2.38	2.63	-2.81
3.20	-1.09	-0.54	0.08
-1.98	-3.30	1.43	-4.48

Comparing HPYLM to MKN: the Figure



Choosing a Language Model

Quick turnaround

- Modified Kneser-Ney (Kneser and Ney, 1995)

Best results

- Hierarchical Pitman-Yor process language model (Teh, 2006)
- Recurrent neural network language model (Mikolov et al., 2010)

Newest

- Sparse non-negative matrix language models (Shazeer, Pelemans, and Chelba, 2014)
- Power low rank ensembles (Parikh et al., 2014), Gaussian embedding (Vilnis and MacCallum, under review), ...