Topic-Based Modeling of Finnish and English News Data

Teemu Hirsimäki
Teemu.Hirsimaki@hut.fi

January 3, 2002

1 Introduction

Gildea and Hofmann introduced a new probabilistic topic-based language model in [2] and [3], and showed how combining the topic model with the traditional trigram model lead to better perplexity results on English data. The goal of this work is to find out how well the topic model performs on Finnish data when compared to English data. Real-world news data is used for both languages.

In the next section the implemented topic model is described. In Section 3 the experimental setup and the results are presented with discussion. Finally, Section 4 concludes the work.

2 Topic Model

The basic idea of the topic model [2] is to decompose the word probability in topics:

$$P(w|h) = \sum_t P(w|t)P(t|h),$$

where w is the word, h is the history of the observed words, and t is the topic. The topic specific unigrams $P(w|t)$ are estimated from training data using the following EM-iterations:

$$P(t|w,d) \propto P(w|t)P(t|d), \tag{1}$$

$$P(w|t) \propto \sum_d n(w,d)P(t|w,d), \tag{2}$$

$$P(t|d) \propto \sum_w n(w,d)P(t|w,d), \tag{3}$$

where d denotes a training document, and $n(w,d)$ the number of occurences of word w in document d. Equation (1) constitutes the E-step of the EM algorithm, while Eqs. (2) and (3) constitute the M-step.

The topic distribution of the observed history, $P(t|h)$, can be estimated during the testing phase as follows:

$$P(t|h_i) = \frac{1}{i+1}\frac{P(w_i|t)P(t|h_{i-1})}{\sum_{t'} P(w_i|t')P(t'|h_{i-1})} + \frac{i}{i+1}P(t|h_{i-1}), \tag{4}$$

$$P(t|h_1) = \frac{\sum_{w,d} n(w,d)P(t|d)}{\sum_{w,d} n(w,d)}, \tag{5}$$

where h_i is the observed word history $(w_1, ..., w_{i-1})$.

3 Experiments

3.1 Data

The experiments were run on two real-world news data sets, the first data set consisting of 9752 Finnish articles (STT). The mean length of the articles was 55.3 words, the standard deviation was 17.7 words, and in total there were 539 589 words. In order to overcome the problems with the inflected forms in Finnish language, all inflected forms were restored to base forms using TWOL, a morphological analyser by Lingsoft [4]. The data was used to create 4 different sets by selecting random 9000 articles for training, and leaving 752 articles for testing.

The second data set consisted of 3102 English articles (TREC). The mean length was 248 words, the standard deviation was 364 words, and there were 771 795 words in total. Again, 4 sets were created, using 3,000 random articles for training, and 102 articles for testing.

In both cases, all words occuring at least twice in the training data were included in the vocabulary. All words outside the vocabulary were mapped to a single out-of-vocabulary word, which was treated as normal word in training and testing. Also all numbers were mapped to a single word. Table 1 summarizes the statistics of the created sets.

Finnish data set	Words in vocabulary	OOVs (%) training set	OOVs (%) test set	Unigram perplexity	Trigram perplexity
1	20002	3.6	7.1	1494.41	310.15
2	19983	3.6	6.8	1545.30	303.14
3	19975	3.7	6.7	1560.81	299.39
4	19971	3.6	7.3	1530.40	307.83

English data set	Words in vocabulary	OOVs (%) training set	OOVs (%) test set	Unigram perplexity	Trigram perplexity
1	15532	1.2	2.9	904.17	242.55
2	15528	1.2	3.0	912.03	217.93
3	15538	1.2	2.6	885.94	212.51
4	15541	1.2	2.2	801.62	199.31

Table 1: Vocabulary statistics of the data sets.

Topics	EM-iterations 16	64	256	1024
4	14.74	22.53	23.21	23.56
16	28.05	35.70	37.76	38.82
64	40.02	45.91	46.23	46.36
128	45.09	47.49	43.70	39.47

Finnish data set 1

Topics	EM-iterations 16	64	256	1024
4	3.12	7.74	8.42	8.29
16	6.16	14.46	15.81	16.19
64	11.63	20.55	20.87	20.80
128	13.90	22.18	22.68	22.72

English data set 1

4	12.74	20.58	23.11	23.96
16	28.57	36.88	39.09	39.73
64	41.19	46.54	46.66	45.40
128	44.72	47.30	43.96	40.68

Finnish data set 2

4	3.31	8.64	9.61	9.77
16	8.02	16.81	18.29	18.91
64	13.22	22.76	23.96	24.21
128	16.17	25.04	25.49	25.40

English data set 2

4	14.97	23.12	24.14	24.46
16	26.41	34.72	37.08	38.38
64	39.23	45.30	46.04	45.24
128	45.26	47.09	43.16	41.11

Finnish data set 3

4	2.98	7.26	7.50	7.50
16	7.25	14.39	15.28	15.97
64	12.32	21.07	21.88	21.93
128	14.48	23.08	23.53	23.57

English data set 3

4	11.37	17.46	19.84	20.85
16	25.79	34.20	36.43	37.40
64	38.38	44.46	45.25	45.30
128	44.78	46.75	41.63	38.54

Finnish data set 4

4	2.13	8.10	8.69	8.71
16	5.69	12.83	14.30	14.85
64	10.69	19.30	20.07	20.62
128	14.05	21.48	21.87	22.20

English data set 4

Table 2: Comparing the unigram model with the topic model. The reported values are perplexity reduction percentages gained by using the topic model instead of the unigram model.

3.2 Training

The trigram model was trained for each data set using CMU-Cambridge toolkit [1] with Good Turing smoothing. The topic model was trained for 1, 4, 16, 64, 128 topics, and 1, 16, 64, 256, 1024 EM-iterations. Note, that using only one topic in the topic model is equivalent to the unigram model, which can be seen in the updating rules of the EM algorithm. The distributions $P(w|t)$ and $P(t|d)$ were initialized randomly before EM-iterations.

3.3 Results and Discussion

Table 2 on the preceding page shows the perplexity reductions in percentages when the unigram model is replaced with the topic model. The first interesting remark is that the topic model seems to help relatively more on Finnish data. Also, on both languages the performance of the topic model seems to improve quite dramatically as the number of topics increases, but on Finnish data, the 128-topic model starts to overfit after 64 iterations. However, on English data this does not seem to happen even with 128 topics and 1024 iterations. One could also say that running 64 iterations seems to be a reasonable choice regardless of the number of topics.

Next, the performance of the traditional trigram model was compared with the linearly combined trigram-topic model:

$$P_{combined}(w) = \lambda P_{trigram}(w) + (1 - \lambda)P_{topic}(w). \tag{6}$$

The weight of the trigram model was varied and the weights resulting in best perplexity results are reported in Table 3 on the next page, with the corresponding perplexity reductions in Table 4. Again, the topic model seems to benefit Finnish data more than English data. It is also interesting, that even though the pure topic model started to overfit on Finnish data after 64 iterations with 128 topics (see Table 2), the performance of the combined trigram-topic model continues to increase up to 1024 iterations.

4 Conclusion

The topic model introduced in [2] was compared to the traditional unigram and trigram models on real-world Finnish and English news data. Various number of topics and EM-iterations were tried, and while the topic model increased performance on both languages, the Finnish data seemed to benefit more. However, this might be partly caused by the fact, that the English articles were longer, and further experiments on other corpora would be needed to find out if the topic model is really more suitable for Finnish than for English.

Also, the results on the English data were not as good as the results reported by Gildea and Hofmann in [2]. They got 38% reductions when replacing the unigram model with the topic model, and 16% reductions when the topic model was combined with the trigram model, while in this work the reductions were around 25% and 8%, respectively. Of course, the different data set is a factor, but another reason might be that the annealing used by Gildea and Hofmann was not used in this work. Also, a linear interpolation was used to combine the topic model with the trigram model instead of the more computationally expensive unigram rescaling used in [2].

In future it would be interesting to implement the annealing, and see how much it affects the results on these data sets. Also it might be interesting to try different initializations for the EM-algorithm, since the initialization might have severe effect on the convergence speed, and the perplexity results.

Topics	EM-iterations 1	16	64	256	1024
1	0.88				
4		0.80	0.76	0.74	0.74
16		0.74	0.70	0.68	0.68
64		0.70	0.68	0.66	0.66
128		0.68	0.68	0.68	0.68

Finnish data set 1

Topics	EM-iterations 1	16	64	256	1024
1	0.90				
4		0.88	0.86	0.86	0.86
16		0.86	0.82	0.82	0.82
64		0.84	0.80	0.80	0.80
128		0.84	0.80	0.80	0.80

English data set 1

1	0.88				
4		0.82	0.76	0.76	0.76
16		0.74	0.70	0.68	0.68
64		0.70	0.68	0.68	0.68
128		0.70	0.68	0.70	0.70

Finnish data set 2

1	0.90				
4		0.90	0.86	0.86	0.86
16		0.88	0.84	0.82	0.82
64		0.86	0.82	0.80	0.80
128		0.84	0.80	0.80	0.80

English data set 2

1	0.90				
4		0.80	0.76	0.76	0.76
16		0.76	0.72	0.70	0.70
64		0.72	0.68	0.68	0.68
128		0.70	0.68	0.70	0.70

Finnish data set 3

1	0.92				
4		0.90	0.88	0.88	0.88
16		0.88	0.84	0.84	0.84
64		0.86	0.82	0.82	0.82
128		0.84	0.82	0.82	0.82

English data set 3

1	0.88				
4		0.82	0.78	0.78	0.76
16		0.74	0.72	0.70	0.70
64		0.70	0.68	0.68	0.68
128		0.68	0.68	0.70	0.70

Finnish data set 4

1	0.92				
4		0.90	0.88	0.88	0.88
16		0.90	0.86	0.86	0.86
64		0.88	0.84	0.84	0.84
128		0.86	0.82	0.82	0.82

English data set 4

Table 3: The best values for λ (the weight of the trigram model) in the linearly combined trigram-topic model (Eq. 6).

Topics	EM-iterations 1	16	64	256	1024
1	1.31				
4		4.13	6.43	6.61	6.70
16		8.60	12.11	13.18	13.79
64		14.62	19.35	20.52	21.44
128		17.48	21.66	22.58	22.76

Finnish data set 1

Topics	EM-iterations 1	16	64	256	1024
1	1.56				
4		1.90	2.77	2.94	2.90
16		2.37	4.41	4.88	4.98
64		3.50	6.33	6.56	6.56
128		4.30	7.12	7.35	7.39

English data set 1

1	1.10				
4		3.24	5.38	6.22	6.48
16		8.27	12.53	13.66	14.01
64		14.39	18.84	20.06	20.62
128		16.66	20.97	21.92	22.32

Finnish data set 2

1	1.15				
4		1.54	2.49	2.74	2.82
16		2.26	4.46	4.87	5.09
64		3.48	6.31	6.77	6.92
128		4.33	7.47	7.81	7.85

English data set 2

1	0.98				
4		3.61	6.09	6.38	6.44
16		7.27	11.13	12.24	12.89
64		13.26	17.84	19.46	20.02
128		17.08	21.45	22.14	22.69

Finnish data set 3

1	1.20				
4		1.50	2.26	2.34	2.32
16		2.10	3.78	4.07	4.29
64		3.59	6.13	6.45	6.53
128		4.17	6.85	7.04	7.13

English data set 3

1	1.14				
4		3.14	4.75	5.37	5.66
16		7.53	11.24	12.34	12.87
64		13.26	18.31	19.61	20.48
128		17.14	21.28	22.13	22.51

Finnish data set 4

1	1.10				
4		1.30	2.06	2.19	2.20
16		1.71	3.08	3.56	3.76
64		2.68	5.11	5.32	5.44
128		3.42	5.87	6.12	6.25

English data set 4

Table 4: Comparing the trigram model with the linearly combined trigram-topic model. The reported values are perplexity reduction percentages gained by using the combined trigram-topic model instead of the traditional trigram model.

References

[1] P. Clarkson and R. Rosenfeld. Statistical language modeling using the CMU-Cambridge toolkit. In *Proc. Eurospeech '97*, pages 2707–2710, Rhodes, Greece, 1997.

[2] D. Gildea and T. Hofmann. Topic-based language models using EM. In *Proceedings of the 6th European Conference on Speech Communication and Technology*, pages 2167–2170, Budapest, Hungary, 1999.

[3] T. Hofmann. Probabilistic latent semantic indexing. In *Research and Development in Information Retrieval*, pages 50–57, 1999.

[4] K. Koskenniemi. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production.* PhD thesis, University of Helsinki, Department of General Linguistics, 1983.