TECHNICAL UNIVERSITY OF MUNICH

TUM Data Innovation Lab

Cross lingual Semantic Search

18.02.2019

Authors:	Liyan Jiang, Phuong Mai, Dmytro Rybalko
Mentors:	Dr.-Ing. Andreas Schoknecht, TWT GmbH Science & Innovation
Co-Mentor:	Michael Rauchensteiner (Department of Mathematics)
Project Lead:	Dr. Ricardo Acevedo Cabra (Department of Mathematics)
Supervisor:	Prof. Dr. Massimo Fornasier (Department of Mathematics)

Table of contents

Introduction to Information Retrieval

Question answering pipeline

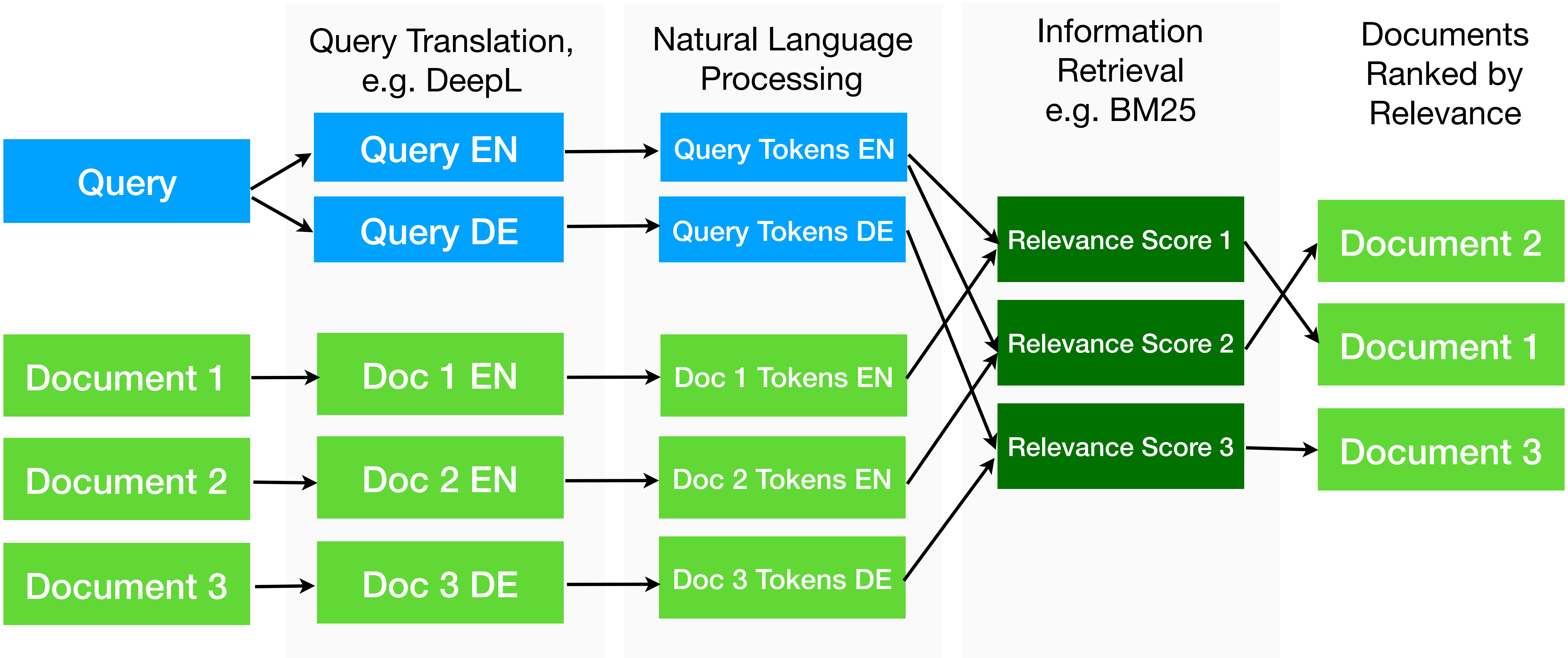

Figure 1: Project Pipeline (own figure)

Cranfield dataset

- Developed by Cleverdone et al. (College of Aeronautics at Cranfield)
- Publicly available

- 1400 abstracts of academic papers
- 255 queries
- Gold-standard

- Exclusively in English
- Translated to German (Google translate)

Cranfield query-document example

Question	how is the heat transfer downstream of the mass transfer region effected by mass transfer at the nose of a blunted cone
Relevant document	experimental investigation of the aerodynamics of a wing in a slipstream . an experimental study of a wing in a propeller slipstream was made in order to determine the spanwise distribution of the lift increase due to slipstream at different angles of attack of the wing and at different free stream to slipstream velocity ratios . the results were intended in part as an evaluation basis for different theoretical treatments of this problem . the comparative span loading curves, together with supporting evidence, showed that a substantial part of the lift increment produced by the slipstream was due to a /destalling/ or boundary-layer-control effect . the integrated remaining lift increment, after subtracting this destalling lift, was found to agree well with a potential flow theory . an empirical evaluation of the destalling effects was made for the specific configuration of the experiment .

Introduction to metrics

Precision

$$Precision = \frac{|\{\textbf{relevant documents} \cap \textbf{retrieved docuemnts}\}|}{\{\textbf{retrieved documents}\}}$$

Mean Average Precision (MAP)

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$$

Discounted Cumulated Gain (DCG)

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)} \qquad nDCG_p = \frac{DCG_p}{IDCG_p}$$

Mean Reciprocal Rank (MRR)

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i}$$

Status of Balabel [1]

- Master Thesis in 2018 at TWT GmbH Science & Innovation
- Implementation of models for semantic retrieval with retrofitting techniques
  - Boolean Model
  - TF-IDF $$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D)$$
  - Dual Embedding space Model (DESM)
  - Combine models
- Results on Cranfield dataset

	MAP	MRR
Thesis	0.131	0.396

Preprocessing

Preprocessing Process

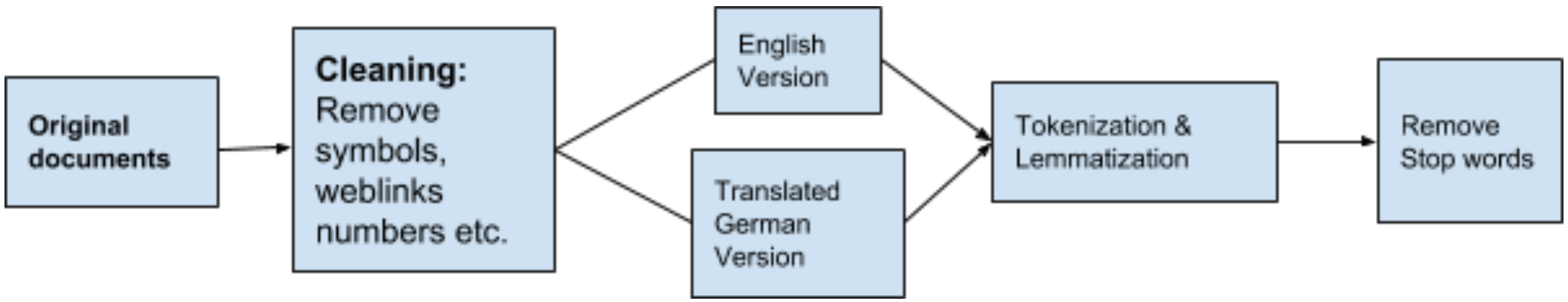

Figure 2: Preprocessing Pipeline (own figure)

Tokenisation

- Chunk text into small pieces of tokens

Stop words removal

- Some words are overwhelmingly common in text. E.g. “the”, “a”, “and”, etc.
- Of little semantic significance to Information Retrieval tasks

Lemmatisation

- Lemma, the morphological analysis of words. E.g. Lemma “go” has word forms “went”, “goes” and “gone”
- Map word forms to lemma

Information Retrieval Methods

Topic Model

Introduction to LDA topic model

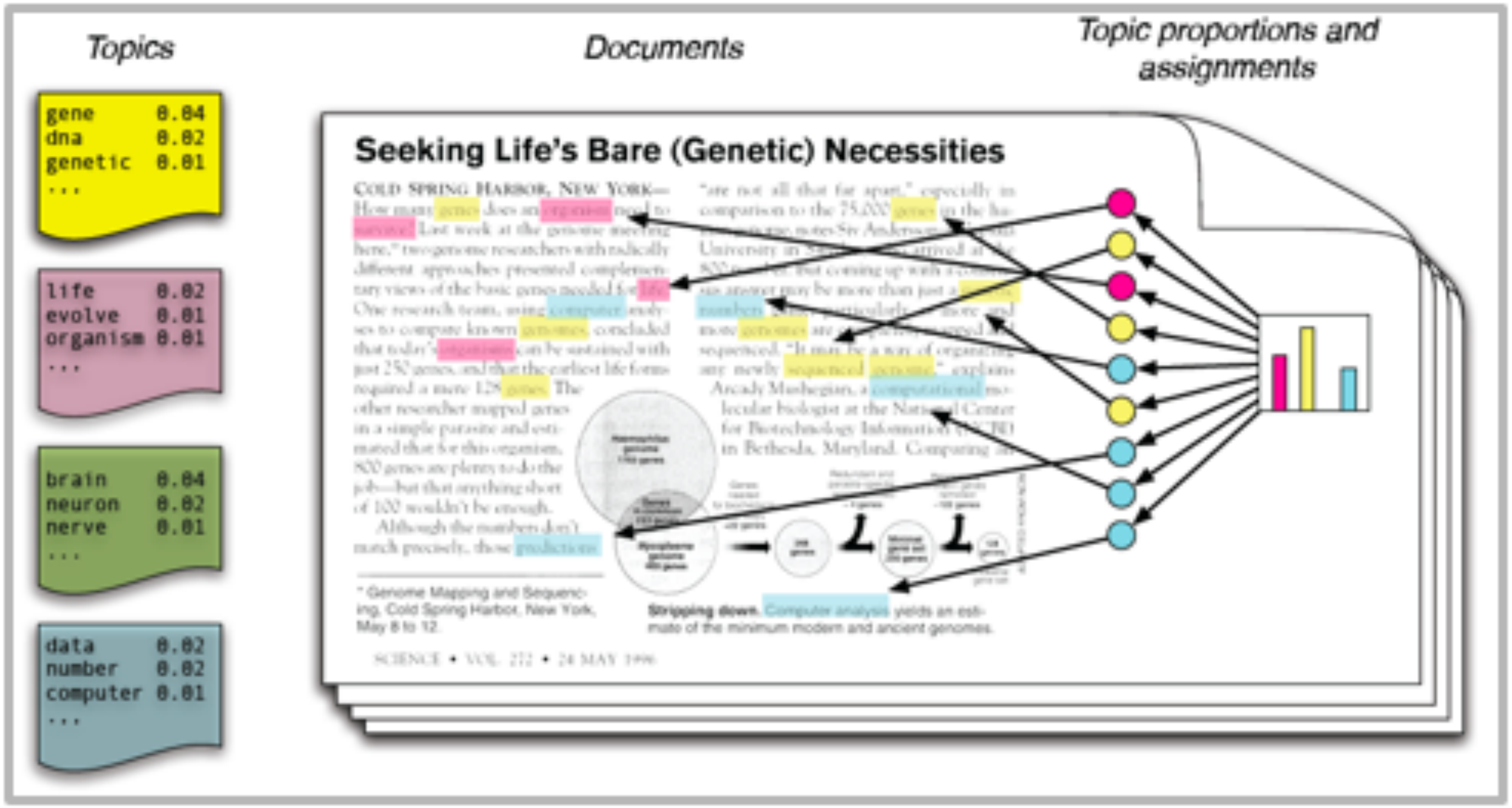

Figure 3: LDA Topic Model. Adapted from [7]

- **Assumption**:
  - Each document is a bag of words.
  - Each topic is a probability distribution of words
- **Output**: document-topic matrix and topic-word matrix

Hyperparameter Tuning

- Hyperparamter: number of topics
- Training:
    - Look into two metrics:
        - **Perplexity**: shows how well this model can predict a sample; the lower the better.
        - **Coherence**: Based on human-interpretability; the higher the better.
- Using grid search to find optimal number of topics.

Hyperparameter Tuning

- What we do:
    - Plot the coherence score and perplexity w.r.t different number of topics
    - We choose 12 as a trade-off.

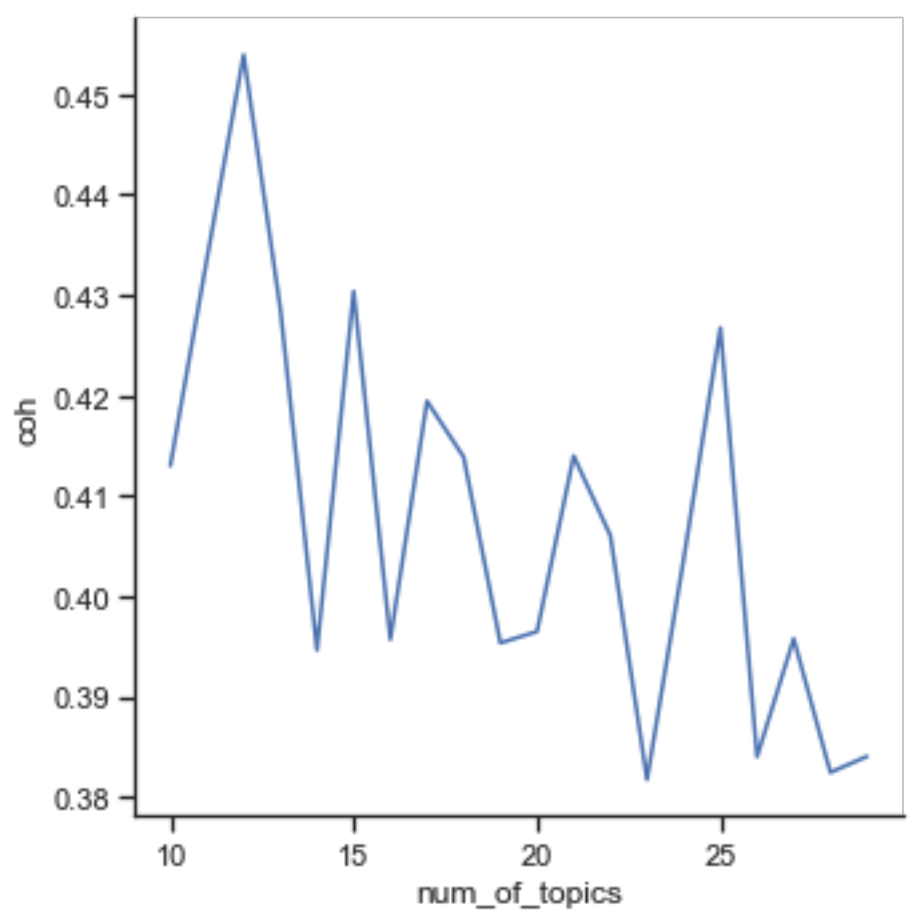

Figure 5: line plot coherence and number of topics (own figure)

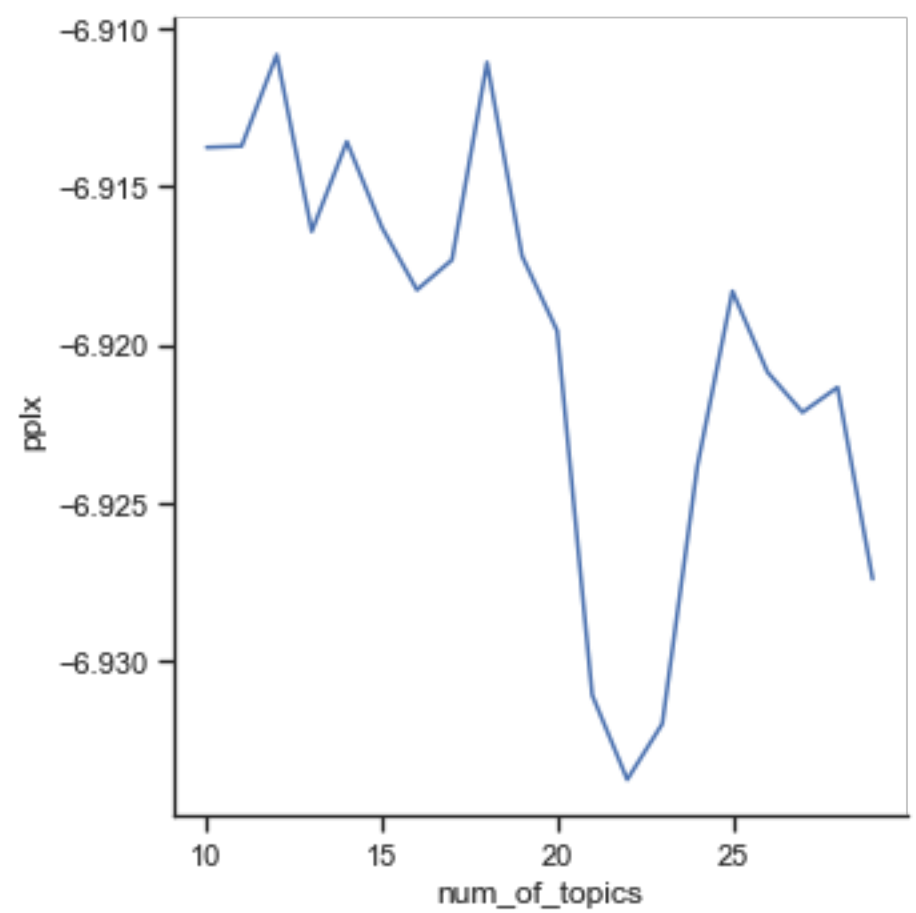

Figure 6: line plot perplexity and number of topics (own figure)

Visualization

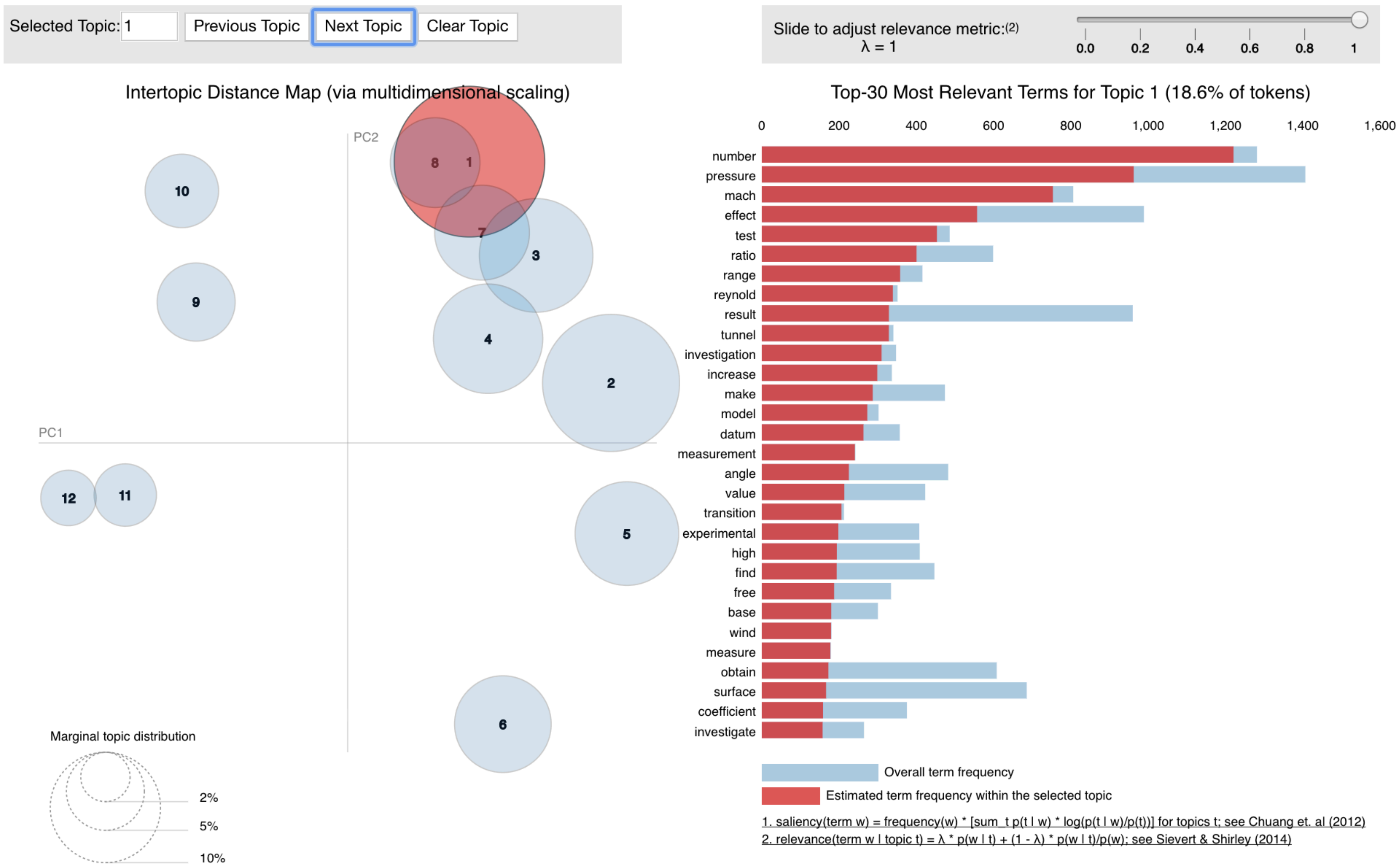

Figure 7: LDA Topic model visualisation (own figure)

Divergence Measure

- Find similar documents based on document topic frequency
  - Jensen Shannon Divergence

$$\mathrm{JSD}(P \parallel Q) = \frac{1}{2} D(P \parallel M) + \frac{1}{2} D(Q \parallel M)$$

where $M = \frac{1}{2}(P + Q)$

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right).$$

  - Distance measure: Euclidean distance
  - Take the inverse of the JSD/Euclidean as the relevance score.

Pipeline

- Demo: Find relevant documents using LDA Topic Model

For each document:

what problems of heat conduction in composite slabs have been solved so far .

- Preprocessing

```
['problem', 'heat', 'conduction', 'composite', 'slab', 'solve']
```

- Term frequency

```
[(35, 1), (130, 1), (131, 1), (135, 1), (142, 1), (402, 1)]
```

- Topic frequency

```
[0.59063751, 0.03963743, 0.02765178, 0.01664344, 0.05313112,
0.03534573, 0.01707318, 0.04674571, 0.04881173, 0.02815013,
0.07957382, 0.01659842]
```

Process in batch: (Both queries and documents)

225x12

1400x12

JSD OR EUD

225x1400

argsort()[,:15]

225x15

Query topic frequency matrix
225 queries
12 topics

Document topic frequency matrix
1400 documents
12 topics

Relevance scores for each
(query, document) pairs

15 most relevant documents
For each query

Figure 8: Process of finding relevant documents (own figure)

Results

- trec_eval metrics
- Model: LDA Topic model (12 topics)

Method/Metrics	MAP	MRR	p_5	p_10	ndcg_5	ndcg_10
LDA_JSD	0.5097	0.5518	0.304	0.2004	0.5333	0.5863
LDA_EUD	0.4768	0.4931	0.2942	0.1991	0.494	0.5568
Thesis	0.131	0.396	/	/	/	/

Best results

Conclusion

- We achieved satisfying results
  - Way beyond the baseline model in all metrics
  - Document topic frequency is good enough to capture the semantic features
- But
  - Texts are not long enough
  - Try more sophisticated hyper-parameter tuning

Exact Matching Methods

State-of-the-art IR

- TF-IDF: count-based retrieval

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D)$$

- Exact Matching Baselines perform comparable to semantic matching Baselines!

GOV2 collection

Model Type	Model Name	Topic titles MAP	Topic titles nDCG@20	Topic titles P@20	Topic descriptions MAP	Topic descriptions nDCG@20	Topic descriptions P@20
Exact Matching Baselines	QL	0.295^{-}	0.409^{-}	0.510^{-}	0.249^{-}	0.371^{-}	0.470^{-}
	BM25	0.295	0.421	0.523	0.256^{-}	0.394	0.483
	SDM	0.319^{+}	0.441^{+}	0.549^{+}	0.275	0.411	0.512^{+}
Semantic Matching Baselines	RM3	0.301	0.395^{-}	0.512	0.263^{-}	0.372^{-}	0.476
	LM+WE-VS	0.295^{-}	0.408^{-}	0.509^{-}	0.254^{-}	0.382^{-}	0.474^{-}
	WE-GLM	0.299^{-}	0.411^{-}	0.513	0.253^{-}	0.384^{-}	0.478
Our Approach	NWT	0.304	0.422	0.524	0.274	0.404	0.492

Figure 9: Comparing exact and semantic matching baselines for GOV2 document collection. Adapted from [5].

Terrier

Open source search engine that implements state-of-the-art methods

- Widely used in practice: BM25

$$score(D, Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{TF(q_i, D) \cdot (k_1 + 1)}{TF(q_i, D) + k_1 (1 - b + b \frac{|D|}{avgdl})}$$

- with k_1 and b are free parameters. Usually, k_1= [1.2, 2.0], b = 0.75.
- Our project: 16 models from Terrier: PL2, Hiemstra LM, DLH, ..

Results of English dataset

- Results of best methods

Model/Metrics	MAP	MRR	P@5	P@10	NDCG@5	NDCG@10
DLH	0.4529	0.4773	0.2853	0.1964	0.4774	0.5409
Hiemstra LM	0.4518	0.4796	0.2836	0.1964	0.4751	0.5412
BM25	0.4445	0.466	0.2836	0.1937	0.4706	0.5354
PL2	0.4544	0.4812	0.2871	0.196	0.4825	0.5415
Thesis	0.131	0.396	/	/	/	/

Best results

Results of German dataset

- Results of best methods

Model/Metrics	MAP	MRR	P@5	P@10	NDCG@5	NDCG@10
Lemur TF-IDF	0.4476	0.4873	0.2827	0.1964	0.4717	0.5362
BM25	0.4353	0.4623	0.2863	0.196	0.4604	0.526
PL2	0.4514	0.4857	0.2889	0.1982	0.4784	0.5420
Best of English	0.4529	0.4773	0.2853	0.1964	0.4774	0.5409

Best results among German dataset

Phrase detection by Mikolov et al. [4]

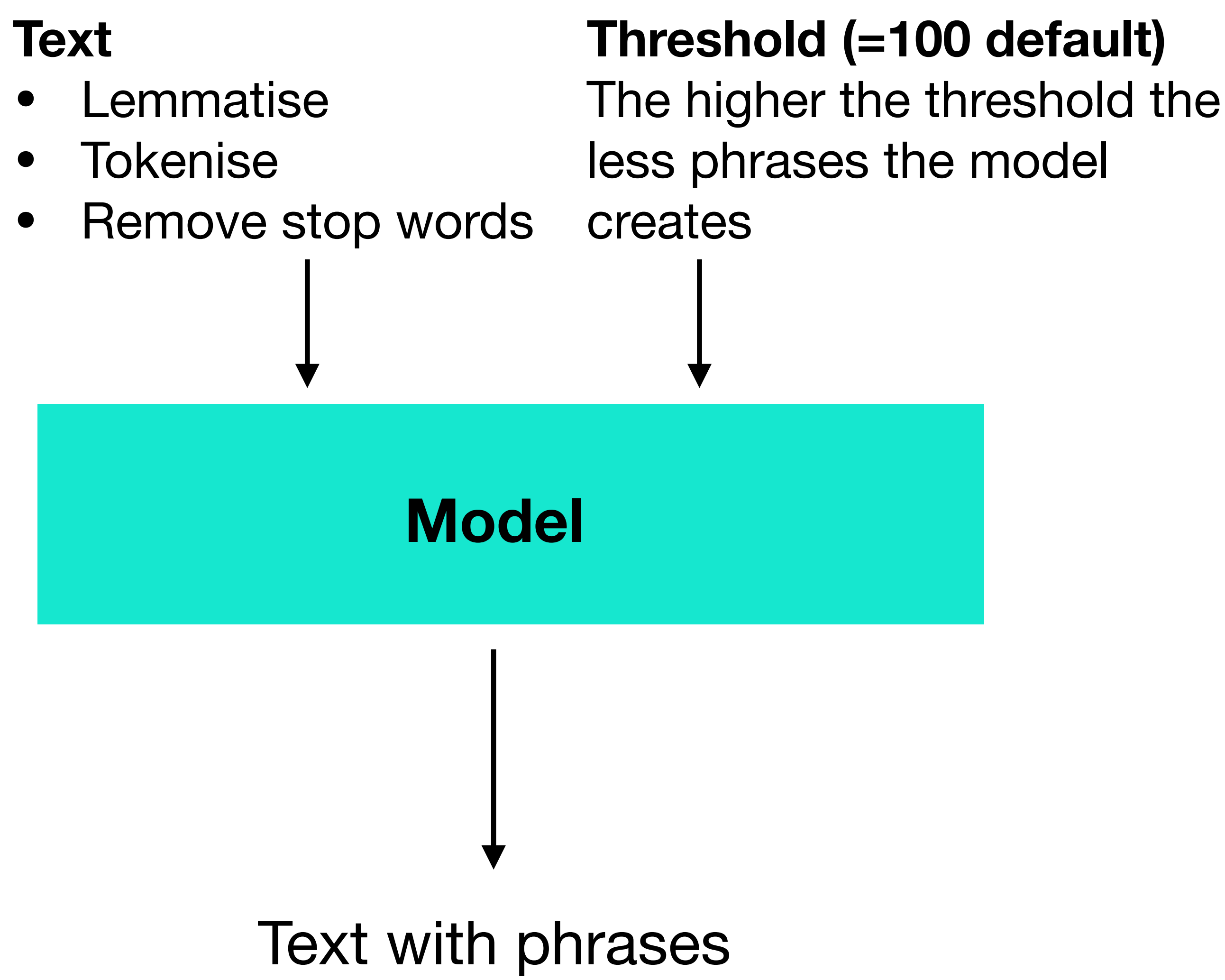

Figure 10: Word2vec's phrase detection architecture (own figure)

how is the heat transfer downstream of the mass transfer region effected..

heat_transfer downstream mass transfer region effect

Phrase detection English

- Unfortunately, only one increase in precision

Model/Metrics	MAP	MRR	P@5	P@10	NDCG@5	NDCG@10
DLH	0.4392	0.4628	0.2853	0.1964	0.4676	0.5307
Hiemstra LM	0.4457	0.4759	0.2836	0.1956	0.4721	0.5363
PL2	0.4458	0.4666	0.2898	0.1942	0.4761	0.5328

Better than in original results

Phrase detection German

- Small increase in values
- BM25 outperforms all methods

Model/Metrics	MAP	MRR	P@5	P@10	NDCG@5	NDCG@10
Lemur TF-IDF	0.455	0.4855	0.2844	0.1956	0.4772	0.54
BM25	0.4607	0.4888	0.2818	0.1956	0.479	0.544
PL2	0.4599	0.4858	0.28	0.1969	0.4763	0.5438

Better than in original results

Conclusion

- Exact Match Baselines are strong!
  - Outperformed Balabels results [1]
- Phrase detection does not improve significantly
  - English: No improvement
  - German: BM25 performed the best
  - Small Cranfield dataset: Only bigram detection possible

Neural Ranking Model using Adversarial Learning

Neural Network Approach

Can we use deep learning to predict relevance?

- **Yes:** there are a lot of NN architectures for relevance prediction

- **No:** Cranfield is too small, can not be used for training

Neural Network Approach

Solution: **Transfer learning**

Train NN on large open-source information retrieval datasets and apply trained model to Cranfield dataset

Problem: **Dataset bias**

Neural Network Approach

Possible solution:

NN learns features automatically from raw text. Try to force it to learn as dataset invariant features as possible.

Cross Domain Regularization using Adversarial Learning

Model Architecture

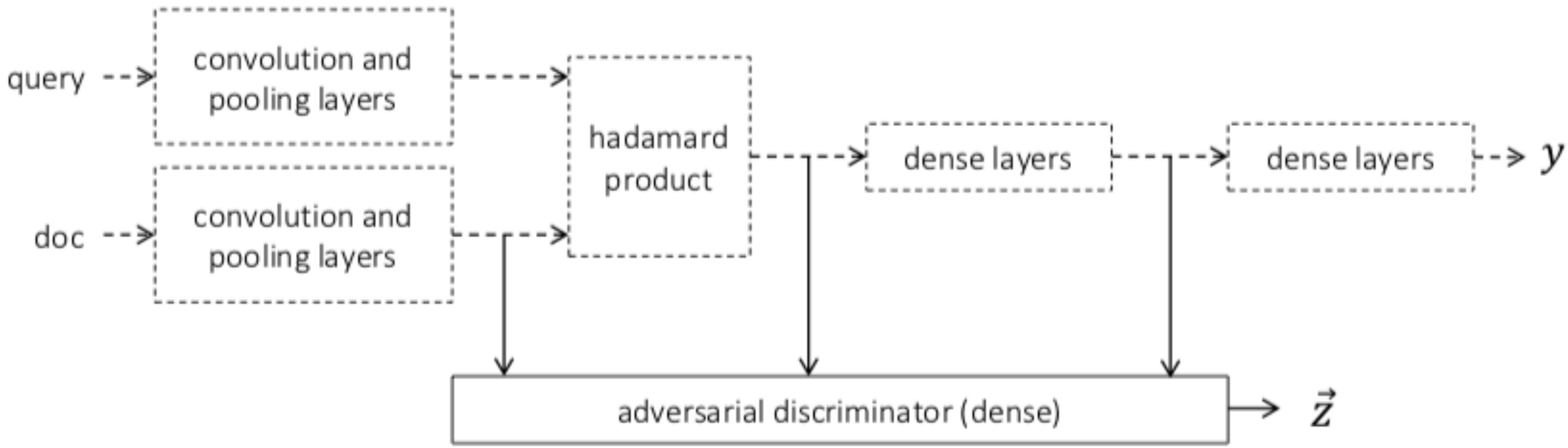

Figure 11: Duet model with adversarial discriminator. Source [2]

- Use reversal gradient layer of adversarial discriminator to force model to learn dataset independent features

Datasets

WebAP

- A crawl of .gov sites
- Number of questions: 82
- Number of answers: 8,027

insuranceQA

- Vocabulary size: 69,580
- Number of questions: 16,889
- Number of answers: 27,413

Yahoo L4

- Forum for Questions and Answers of different topics: Sports, Politics, Home&Garden ..
- Number of questions: 142,627
- Number of answers: 819,604 (filtered)

Training

1. **Separate Training**

- Duet Model
- Train on WebAP/insuranceQA/Yahoo L4
- Evaluation on Cranfield

2. **Adversarial Training**

- Duet and Adversarial Model
- Train on WebAP+insuranceQA+Yahoo L4
- Evaluation on Cranfield

NN Results

Training dataset/ Metrics	MAP	MRR	P@5	P@10	NDCG@5	NDCG@10
WebAP - 1 Epoch	0.4047	0.4259	0.2587	0.1920	0.4191	0.4994
insuranceQA - Ep 1	0.4062	0.4411	0.2622	0.1929	0.4259	0.5043
Yahoo L4 - Ep 1	0.4198	0.4564	0.2649	0.1889	0.439	0.5099
Duet adversarial network	0.4403	0.4831	0.2658	0.1951	0.4537	0.5308

Best results

Final results

Baseline/ Metrics	IR Techniques	MAP	MRR	P@5	P@10	NDCG@5	NDCG@10
Topic Modelling	LDA	0.5097	0.5518	0.304	0.2004	0.5333	0.5863
Exact Matching	DLH	0.4529	0.4773	0.2853	0.1964	0.4774	0.5409
	Hiemstra LM	0.4518	0.4796	0.2836	0.1964	0.4751	0.5412
	PL2	0.4544	0.4812	0.2871	0.196	0.4825	0.5415
Neural network approach	Duet adversarial network	0.4403	0.4831	0.2658	0.1951	0.4537	0.5308
Thesis	Baseline model	0.131	0.396	/	/	/	/

Best results

Conclusion

Conclusion

- We tested three approaches: Topic Modelling, Exact Matching, Neural Networks
- All of them outperformed Balabel's results [1]
- Exact Matching and Topic Modelling are cheap, fast and perform as good as neural network approach

Sources

[1] Balabel, M. (2018) CLEISST: a Cross-lingual Engine for Informed Semantic Search in the Technical Domain. Master thesis, Universität Stuttgart, Germany. Institut für Maschinelle Sprachverarbeitung.

[2] Cohen, Daniel, et.al. "Cross Domain Regularization for Neural Ranking Models using Adversarial Learning", *SIGIR*,2018

[3] Guo, Jiafeng, et al. "Semantic matching by non-linear word transportation for information retrieval." *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016.

[4] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*. 2013.

[5] Guo, J., Fan, Y., Ai, Q. and Croft, W.B., 2016, October. Semantic matching by non-linear word transportation for information retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 701-710). ACM.

Backup slides

Terrier Models

- BB2
- BM25
- DFR_BM25
- DLH
- DLH13
- DPH
- DFRee
- Hiemstra_LM
- DirichletLM
- IFB2
- In_expB2
- In_expC2
- InL2
- LemurTF_IDF
- LGD
- PL2
- TF_IDF
- Please refer to http://terrier.org/docs/v3.5/configure_retrieval.html for further information

Trec eval

- Evaluation of true query results and terrier query results
- 4/25 Metrics
  - **MAP**: Mean average precision
  - **MRR**: Mean reciprocal rank
  - **P@5**: Precision for 5 retrieved documents
  - **P@10**: Precision for 10 retrieved documents
  - **NDCG@5**: Normalised discounted cumulative gain for 5 retrieved documents
  - **NDCG@10**: Normalised discounted cumulative gain for 10 retrieved documents

Document Clustering

- Idea: using document topic frequency matrix to cluster the documents.
- Recap: what is document topic frequency

NxD

N: number of documents
D: number of topics in LDA model
Value in each cell: (R_ij)
the probability of document i being assigned to topic j

Example:
R[2,:] = [0.08261366, 0.02481725, 0.37290511, 0.,
0.03365476, 0.02093827, 0. , 0.27946776,
0.05703923, 0.01500396, 0.08671068, 0.]

- Two clustering appraches:
  - Take the most frequent topic
  - KMeans clustering
    - Set number of topics to 12 (optimal)
    - Using sklearn.cluster.KMeans to generate cluster
    - Taking the topic frequency as features
    - Generate 225 clusters, corresponding to the number of queries.

Document Clustering

- Results:

	MAP	MRR	P@5	P@10	NDCG@	NDCG@1
most frequent	0.0139	0.0277	0.0053	0.0116	0.0071	0.0206
Means	0.0014	0.0059	0.0027	0.0013	0.0031	0.0003

- Reasons
  - Intuitively: too general
  - Entropy goes down
    - Entropy: $H = -\sum_{i=0}^{n} freq(t_i) \log freq(t_i)$
      - before: *[0.08261366, 0.02481725, 0.37290511, 0., 0.03365476, 0.02093827, 0. , 0.27946776,0.05703923, 0.01500396, 0.08671068, 0.]* => **Entropy = 1.6553507**
      - After: [0,0,1,0,0,0,0,0,0,0,0,0] => **Entropy = 0**
    - Which means that the model carries less information

Document distribution

Example sentences from different documents in **Topic 1**

“wassermann gave analytic solutions for the temperature in a double layer slab, with a triangular heat rate input at one face, insulated at the other, and with no thermal resistance at the interface”

“this type of heating rate may occur, for example, during aerodynamic heating”

“it was desired to estimate the eddy viscosity in axisymmetric, compressible wakes”

“it is concluded that the heat transfer through the equilibrium stagnation point boundary layer can be computed accurately by a simple correlation formula”

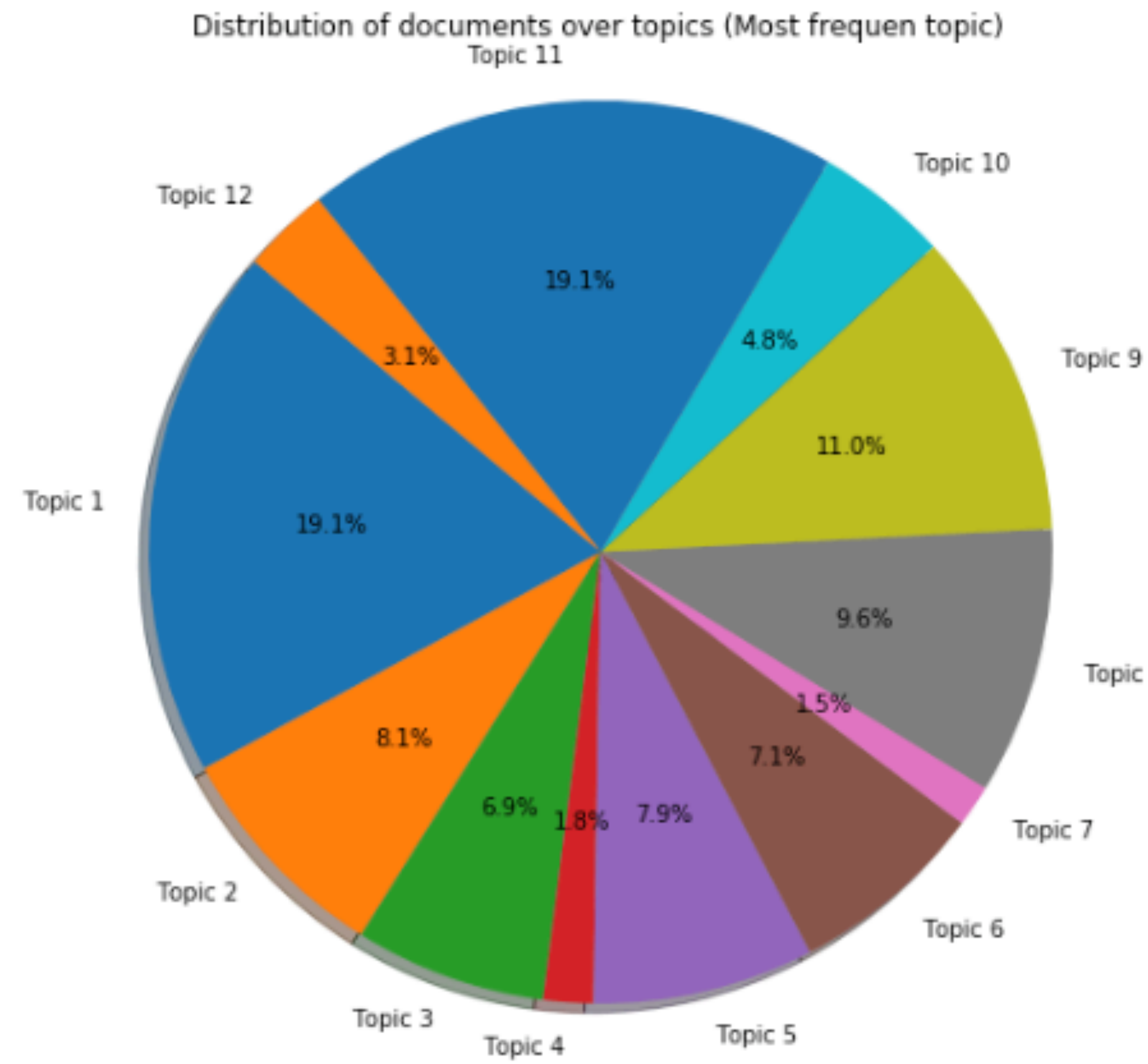

Figure 1: Document clustering result (Most frequent topic) (own figure)

Hierarchical Dirichlet Process model (HDP)

- Extension of LDA topic model
- Unsupervised topic model
- Extracted 150 topics from the Cranfield dataset
- Results: not better than LDA topic model

	MAP	**MRR**	**P@5**	**P@6**	**NDCG@5**	**NDCG@10**
HDP	0.4096	0.4142	0.2963	0.1938	0.4255	0.5013
LDA_JSD	0.4741	0.4934	0.2942	0.1996	0.4918	0.5554

Best results

HDP generated 150 topics

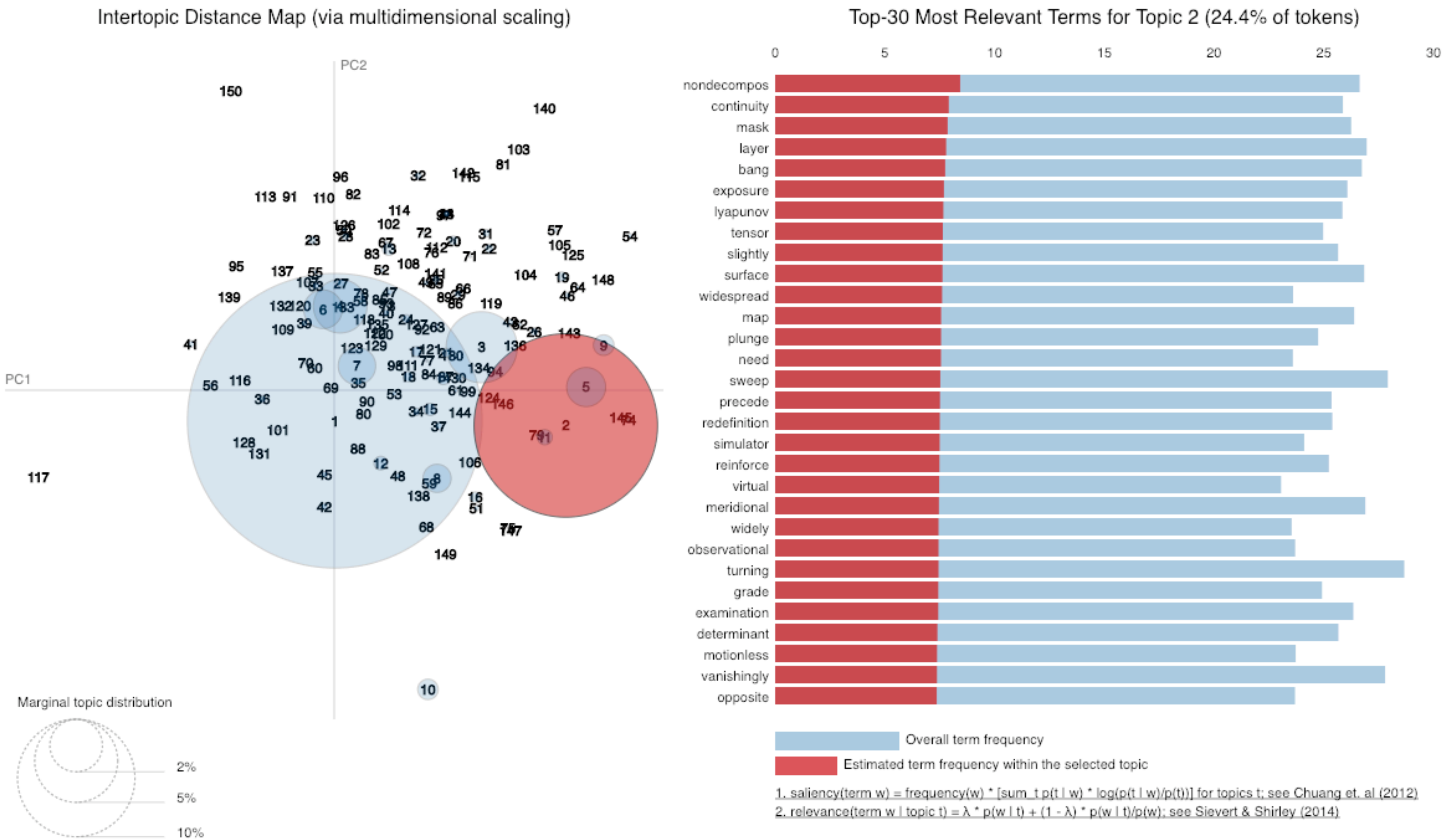

Figure 6: HDP visualisation (own figure)

Dataset I: WebAP

- A crawl of .gov sites
- Number of questions: 82
- Number of answers: 8,027
- Average length of a passage: 45 words

Question	Describe the history of the U.S. oil industry
Answer	The oil industry in Alaska, due to its dynamic nature and significant economic impacts, has been the source of much discussion. The industry has been involved in an unprecedented amount of legislation, lawsuits, and continued business negotiations with the State. Part of the reason for this intense interest is the magnitude of both the industry's workforce and related payroll. The Department of Labor's (DOL) 1995 Nonresidents

Dataset II: InsuranceQA

- Insurance documents
- Number of questions: 16,889
- Number of answers: 27,413
- Vocabulary size: 69,580

Question	medicare-insurance What Does Medicare IME Stand For? 16696
Answer	According to the Centers for Medicare and Medicaid Services website, cms.gov, IME stands for Indirect Medical Education and is in regards to payment calculation adjustments for a Medicare discharge of higher cost patients receiving care from teaching hospitals relative to non-teaching hospitals. I would recommend contacting CMS to get more information about IME
Irrelevant answer	Unless something has changed recently with their testing protocol, no State Farm does not test for THC.

Dataset III: Yahoo L4

- Forum for Questions and Answers of different topics: Sports, Politics, Home&Garden ..
- Number of questions: 142,627
- Number of answers: 819,604 (filtered)

Question	How to clean window screens?
Best answer	Nylon covered sponges are great for cleaning window screens
Other answers	I usually take the screen out and lay it on the ground. I use the bathroom cleaner (scrubbing bubbles) then use the hose to wash it off.

Comparing predictions

Rank	BB2 doc_id	True relevance	Duet doc_id	True relevance
1	13	4	462	4
2	486	-1	184	2
3	56	3	30	3
4	142	4	66	3
5	184	2	12	3