General Online Research Conference
GOR 18
28 February to 2 March 2018, TH Köln – University of Applied Sciences, Cologne, Germany

Christopher Harms, SKOPOS GmbH & Co. KG
Sebastian Schmidt, SKOPOS GmbH & Co. KG

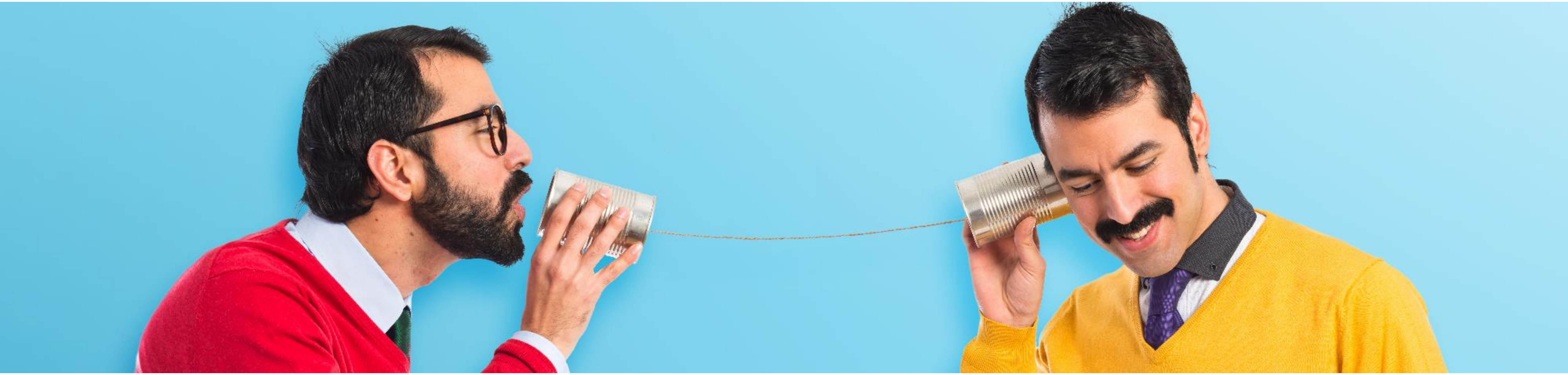
Learning From All Answers: Embedding-based Topic Modelling for Open-Ended Questions

Contact: christopher.harms@skopos.de

Suggested citation: Harms, Christopher, & Schmidt, Sebastian. 2018. "Learning From All Answers: Embedding-based Topic Modelling for Open-Ended Questions.." General Online Research (GOR) Conference, Cologne.

#LearningFromAllAnswers

GOR 2018 | Cologne | March 1, 2018



Learning From All Answers:

Embedding-based Topic Modelling for Open-Ended Questions

Christopher Harms, Consultant Research & Development
Sebastian Schmidt, Director Research & Development

Initial Scenario

What can we do to improve our service for you?

Initial Scenario

How to extract information from open-ended questions?

- **Word Cloud**
- **Qualitative summary**
- **Code plan**
  - Manual coding
  - Automatic coding through supervised learning

- **Can we improve this through unsupervised Machine Learning?**

Methods Overview

- **Naïve Keyword Extraction**
- **Latent Dirichlet Allocation** (LDA; Blei et al., 2003)
- **Embedding-based Topic-Modelling** (ETM, Qiang et al., 2016)

Methods Overview

Naïve Keyword Extraction

- Nouns indicate topics
- Extraction through a pre-trained POS tagger (e.g. spaCy)
- Catch different forms of same word: Lemmatization or Stemming
- Word Cloud of resulting terms, highlighting relative frequency

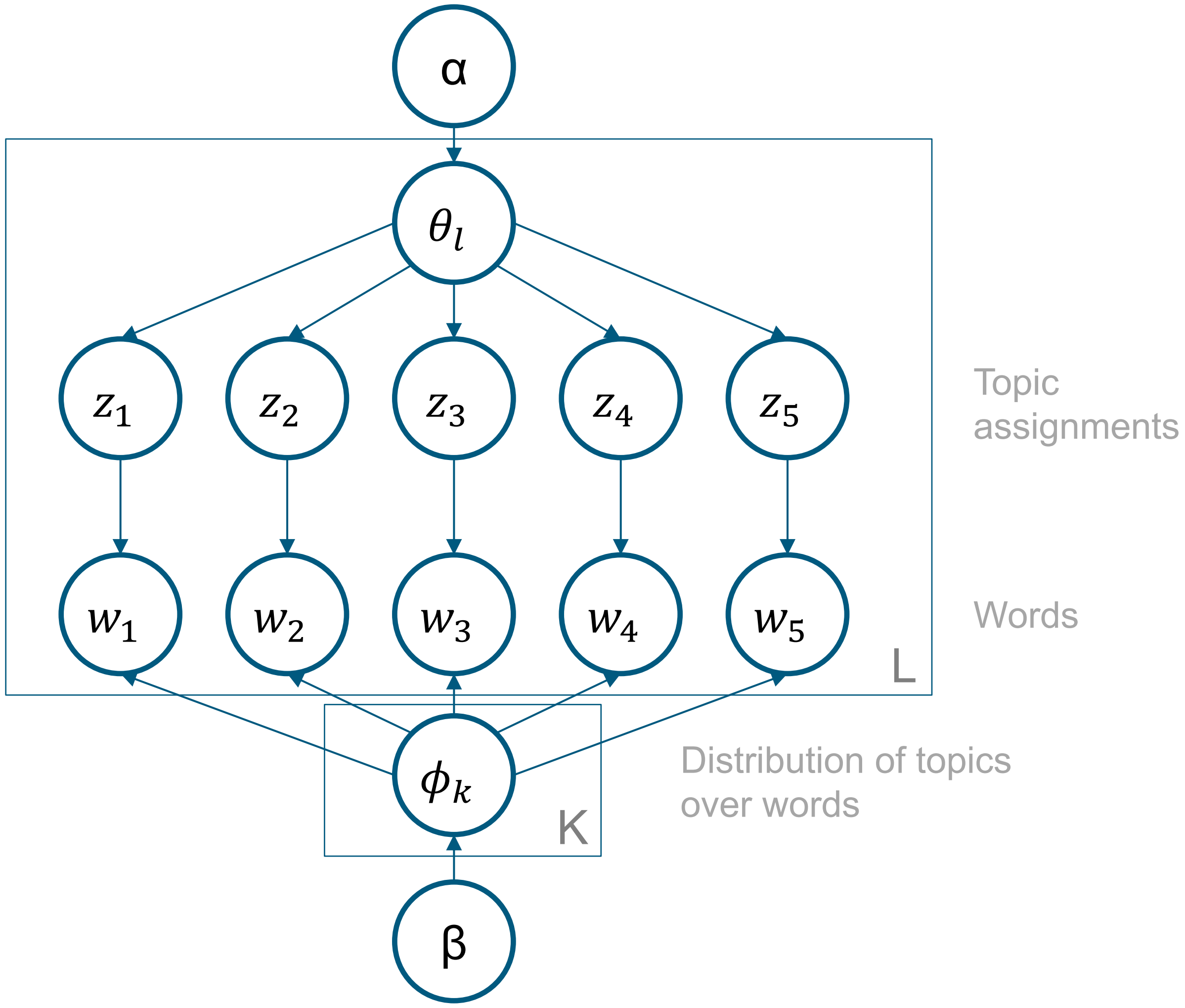
Working from home for me means freedom and independence. I can just go for a walk when there is sunny weather and I need a break.

- Home
- Freedom
- Independ~~ence~~
- Walk
- Weather
- Break

SKOPOS

Latent Dirichlet Allocation

- Bayesian generative probabilistic model
- Each topic is a probability distribution over words
- Inference: Find the relationship between words and topics for a given corpus



Latent Dirichlet Allocation

Benefits

- Co-occurring words are grouped into a topic
- Readily available programming packages (e.g. gensim)

Disadvantages

- Number of topics has to be chosen *a priori*
- Large corpus needed for reasonable results
- No knowledge about relationship between different words (e.g. “buffet” and “restaurant”)

Word Embeddings

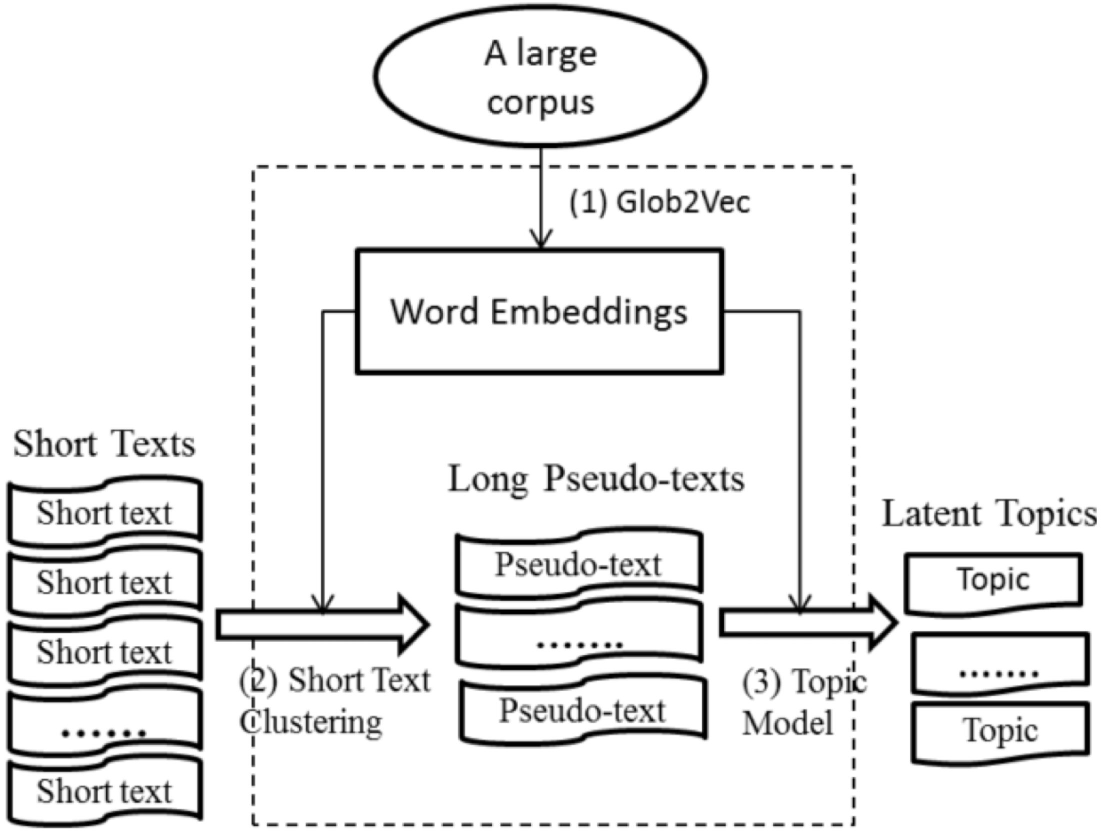
```
king - man + woman = queen

breakfast + lunch = brunch
```

- Embeddings contain information about word relationships
- Trained on a very large corpus of texts
- Each word becomes a multidimensional vector

Embedding-based Topic Modelling

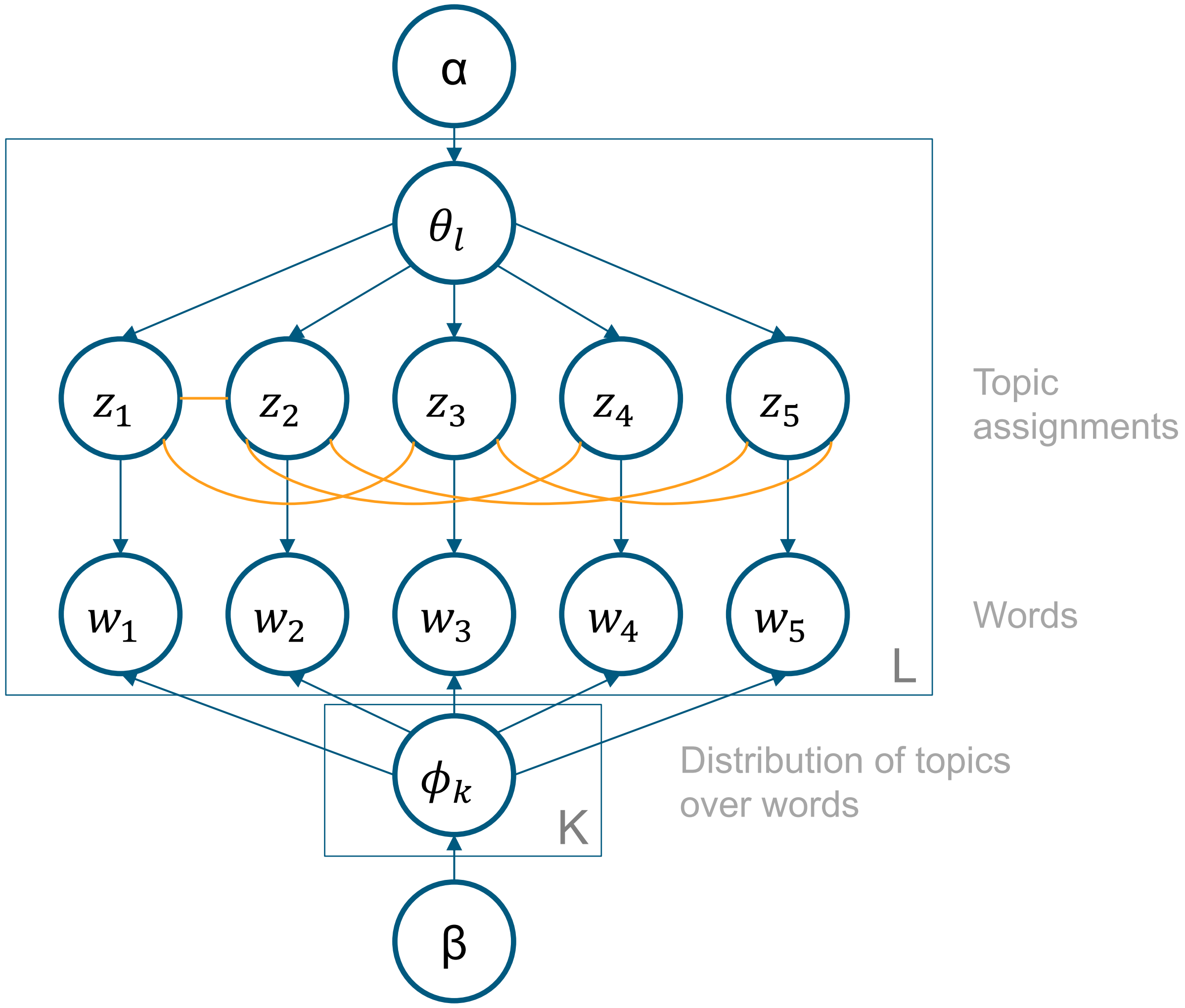
- Extension of the LDA model
  1. Aggregate short texts into pseudo-documents
  2. Assign similar words more likely to the same topic
- Word embeddings are used for similarity of documents and words



Methods Overview

Embedding-based Topic Modelling

- Undirected edge between topics for similar words (binary potential): **Similar words should be more likely belong to the same topic**
- Graphical model is a Markov Random Field (MRF-LDA, Xie et al., 2015)
- Weight for binary potential, if 0 model reduces to LDA



Embedding-based Topic Modelling

Benefits

- Knowledge of word relationships is incorporated (pre-trained embeddings)
- k-Means improves Topic Modelling of short texts

Disadvantages

- Number of pseudo-texts and topics has to be chosen *a priori*
- Computationally expensive
- Requires a large corpus for reasonable results
- No prepared software packages available

Proof-of-Concept

Datasets

Twitter (Sentiment140)

- 10.000 tweets in English language
- Purely observational

Survey Responses

- 10.000 survey responses in German language
- Responses to three different questions concerning travel

SKOPOS

Results: Resulting Topics with Top5 Words (excerpt)

LDA

Topic #1	Topic #2	Topic #3
hope	twitter	morning
better	phone	good
sick	use	cold
feeling	site	snow
feel	tweets	car

Topic #1	Topic #2	Topic #3
gut	super	immer
geklappt	einfach	zufrieden
organisiert	nein	buchen
gefallen	unkompliziert	gerne
reise	schnell	reisen

ETM

Topic #1	Topic #2	Topic #3
new	sad	sleep
cold	house	time
better	watching	night
damn	night	hours
need	thank	bed

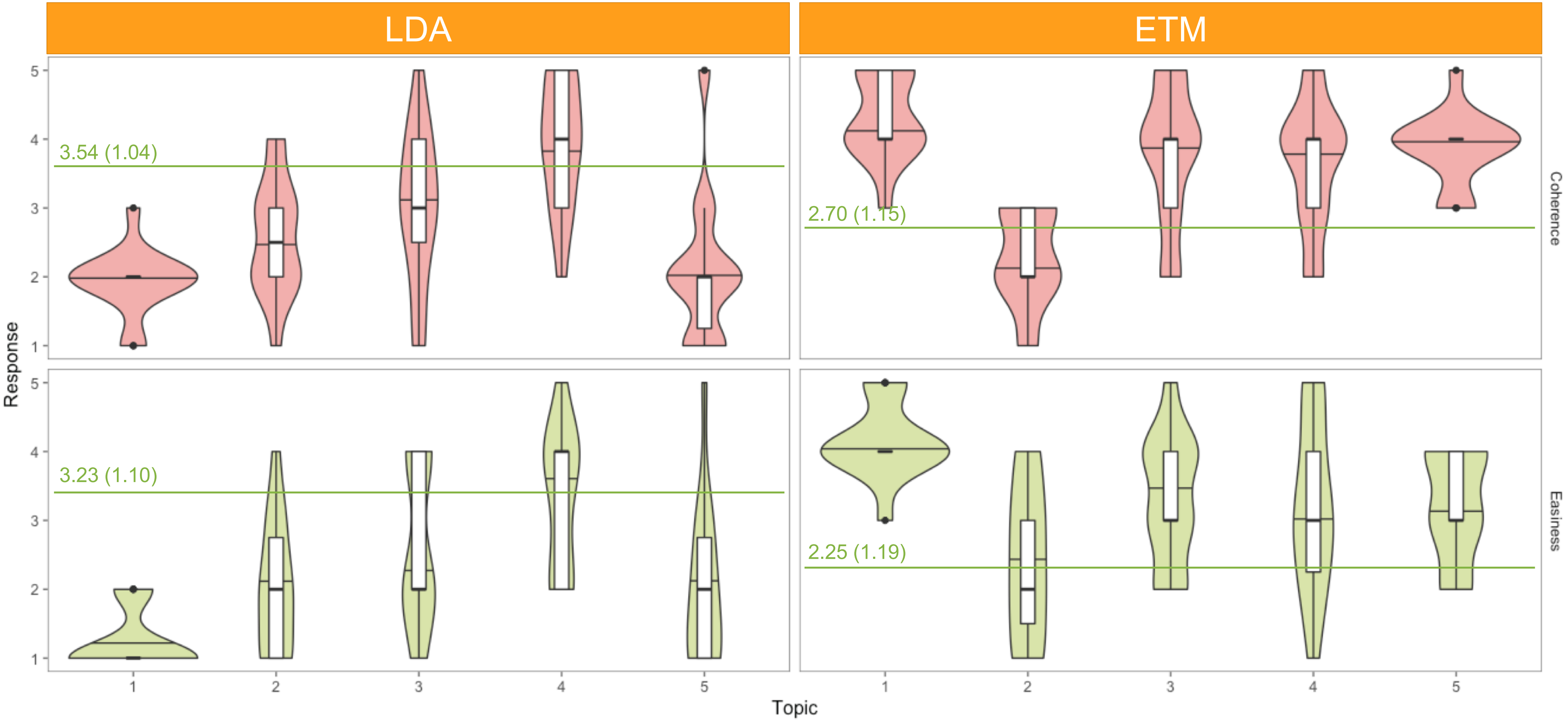
Topic #1	Topic #2	Topic #3
super	geklappt	service
einfach	reibungslos	organisation
stimmt	vielen	hotel
tolle	dank	hotels
funktioniert	perfekt	information

Expert Review

- Classical Machine Learning metrics not informative for real research projects
- Question of interest for us: Can our (human) colleagues work with the results provided by the algorithms?

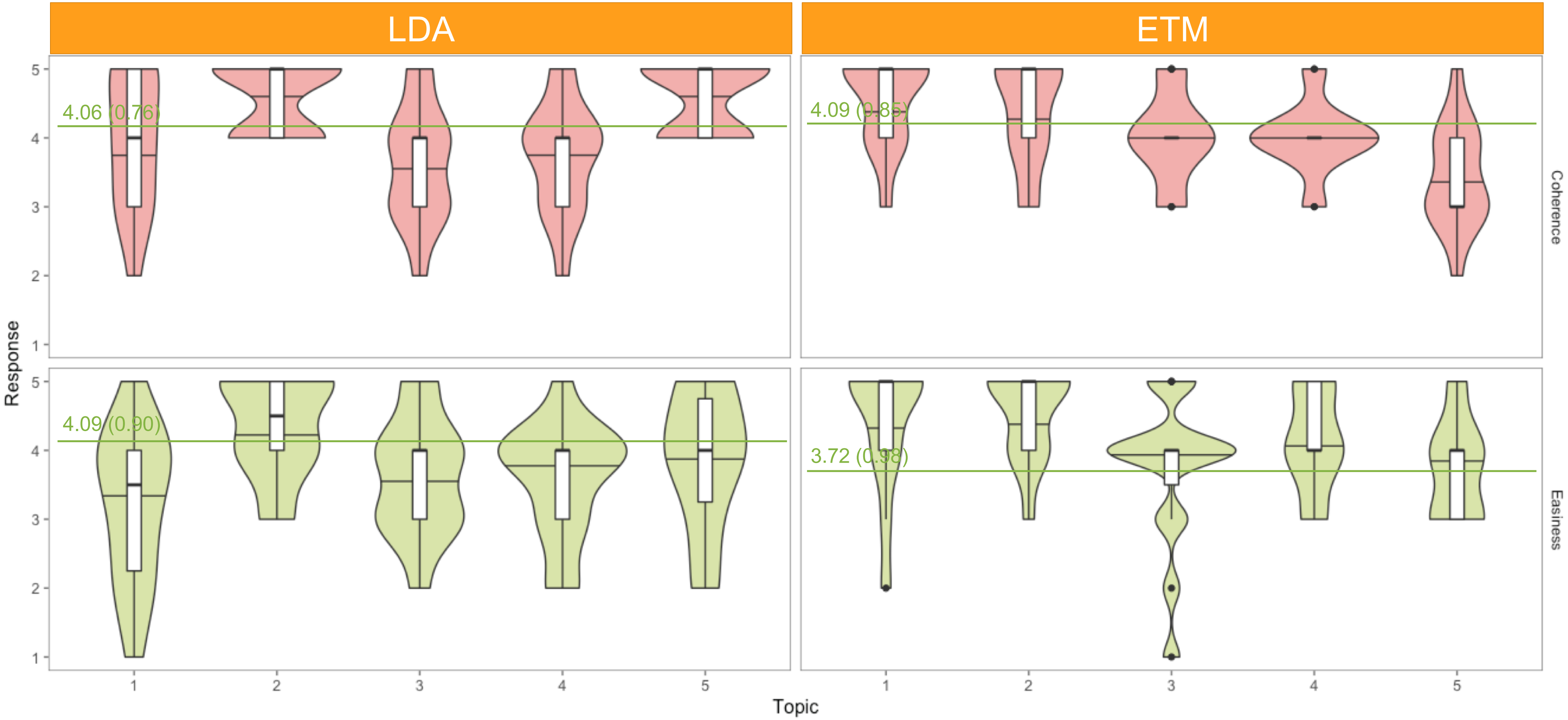
- **Are resulting topics coherent?**
  That is, can words associated with a topic indeed be grouped into a sensible topic?

Results: Expert Review (English Dataset)



SKOPOS

Results: Expert Review (German Dataset)



Summary

- **English:** LDA results more coherent than ETM results
- **German:** ETM and LDA rated equally coherent

- **But:** Highly dependent on topic selection

Summary

Our Learnings

- Proof of Concept – needs further development
- Fine-tuning of hyper-parameters and techniques required
- Pre-trained word vectors provide valuable information
- Lots of data required for best results (> 1,000 responses)
- Metric for usefulness in real-world environment?

Thank You For Your Attention!

Further questions? Let‘s talk!

Christopher Harms
Consultant Research & Development

christopher.harms@skopos.de
@chrisharms

Sebastian Schmidt
Director Research & Development

sebastian.schmidt@skopos.de

SKOPOS
market research

References

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022.
- Qiang, J., Chen, P., Wang, T., & Wu, X. (2016). Topic Modeling over Short Texts by Incorporating Word Embeddings. *CEUR Workshop Proceedings, 1828*, 53–59. Retrieved from http://arxiv.org/abs/1609.08496
- Xie, P., Yang, D., & Xing, E. P. (2015). Incorporating Word Correlation Knowledge into Topic Modeling. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)* (pp. 725–734). Retrieved from http://www.cs.cmu.edu/~pengtaox/papers/naacl15_mrflda.pdf