Advanced Text Mining Methods

Carsten Jentsch, Kai-Robin Lange

TU Dortmund, Department of Statistics

30.11.2022

technische universität
dortmund

fakultät
statistik

Schedule and Requirements

Requirements

- Presentation in English; 30+10 minutes for Bachelor students, 45+10 minutes for Master students
- Report up to 12 pages long in English or German

Schedule

- Distribution of topics: voting for priorities until Sunday, the 4th of December
- 10 days prior to presentation: hand-in of the slides and short discussion with us
- Presentations: February 2023 (tbd)
- Reports due: 27th of March 2023

Projects

Comparing embedding models

- Word2Vec vs. GloVe vs. fastText vs. ?
- Strengths and weaknesses of each algorithm
- When to use which
- Theory- (comparing algorithms) or practice-oriented (study on a data set) or both

References

-Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching Word Vectors with Subword Information."
-Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. "Bag of Tricks for Efficient Text Classification."
-Mikolov, Tomas, Ilya Sutskever, et al. 2013. "Distributed Representations of Words and Phrases and Their Compositionality."
-Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space."
-Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation." In Empirical Methods in Natural Language Processing (EMNLP), 1532–43.

Figure 1: Two-dimensional visualization of semantic change in English using SGNS vectors.[2] **a**, The word *gay* shifted from meaning "cheerful" or "frolicsome" to referring to homosexuality. **b**, In the early 20th century *broadcast* referred to "casting out seeds"; with the rise of television and radio its meaning shifted to "transmitting signals". **c**, *Awful* underwent a process of pejoration, as it shifted from meaning "full of awe" to meaning "terrible or appalling" (Simpson et al., 1989).

Projects

Diachronic embeddings

- How to model static embeddings across time
- Basic ideas, advantages and disadvantages of some algorithms
- Theory- (comparing algorithms) or practice-oriented (study on a data set) or both

References

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. "Diachronic Word Embeddings and Semantic Shifts: A Survey." In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA: Association for Computational Linguistics, 1384–97.

Projects

The development of pre-trained language models

- Overview of the different algorithms – from the first pre-trained model to BERT
- What ideas does BERT borrow from its predecessors?

References

- https://jalammar.github.io/illustrated-bert/
-Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.”.
-Howard, Jeremy, and Sebastian Ruder. 2018. “Universal Language Model Fine-Tuning for Text Classification.”.
-Peters, Matthew E. et al. 2018. “Deep Contextualized Word Representations.”
-Vaswani, Ashish et al. 2017. “Attention Is All You Need.”.

Projects

Modern pre-trained language models

- BERT vs. RoBERTa vs. DistilBERT vs. SBERT
- What are the differences?
- Primary focus on one of the "unique"models (i.e. SBERT or DistilBERT)

References

-Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.”
-Liu, Yinhan et al. 2019. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.”
-Reimers, Nils, and Iryna Gurevych. 2019. “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks.”
-Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. “DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter.”
-Vaswani, Ashish et al. 2017. “Attention Is All You Need.”.

Projects

Adapters for pre-trained language models

- Idea of an adapter and difference to regular fine-tuning
- Transition from PET to PERFECT
- Theory- (comparing algorithms) or practice-oriented (study on a data set) or both

References

- https://adapterhub.ml/
- https://aclanthology.org/2020.emnlp-demos.7
- https://www.youtube.com/watch?v=_Z9qNT-g14U
- https://arxiv.org/abs/1902.00751

Projects

Machine translation

- Why and how is DeepL "better" than Google Translate?
- How is machine translation done and how is it evaluated? What is "the best" translation?

References

Kocmi, Tom et al. 2021. “To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation.” In Proceedings of the Sixth Conference on Machine Translation, Online: Association for Computational Linguistics, 478–94.

Reserved ⇒ Emotive	Emotive ⇒ Reserved
I liked the movie. ⇒ I cannot even describe how amazing this movie was!!	I loved every minute of the movie! ⇒ I liked the movie.
I was impressed with the results. ⇒ I was absolutely blown away with the results!!	I was shocked by the amazing results! ⇒ I was surprised by the results.
American ⇒ British	**British ⇒ American**
The elevator in my apartment isn't working. ⇒ The lift in my flat isn't working.	The lift in my flat isn't working. ⇒ The elevator in my apartment isn't working.
The senators will return to Washington next week. ⇒ The MPs will return to Westminster next week.	MPs will return to Westminster next week. ⇒ Representatives will return to Washington next week.
Polite ⇒ Rude	**Rude ⇒ Polite**
Are you positive you've understood my point? ⇒ you've never understood my point!	What the hell is wrong with your attitude? ⇒ Perhaps the question is more about your attitude.
Could you ask before using my phone? ⇒ I ask you to stop using my phone!	I could care less, go find somebody else to do this crap. ⇒ I could be wrong, but I would try to find somebody else to do this.
Formal ⇒ Informal	**Informal ⇒ Formal**
I hereby commit to never purchase anything from this institution in the future. ⇒ i gonna never buy anything from this place again.	best book ever!! ⇒ The book is highly recommended.
I couldn't figure out what the author was trying to say. ⇒ i dont know what ur trying to say.	couldnt figure out what author tryna say ⇒ The reader couldnt figure out what the author was trying to say.
Positive ⇒ Negative	**Negative ⇒ Positive**
I was pretty impressed with the results. ⇒ I was pretty disappointed with the results.	I was pretty disappointed with the results. ⇒ I was pretty impressed with the results.
I will definitely buy this brand again. ⇒ I will definitely not buy this brand again.	I definitely won't buy this brand again. ⇒ I definitely won't hesitate to buy this brand again.

Table 1: Examples of transferring along five different axes of style. The same model is used across all examples, with no additional training. Words deleted from the input are red, and words added in the output are blue. Within each category, a fixed tiny set of exemplars is chosen, and a fixed delta scale and tuning rates are used. The exemplars and settings are provided in Appendix A.2.

Projects

Language/Text style transfer

- Changing the style of a text without chaning its content
- Analyzing a survey: How is the transition done?
- Parallel vs. Non-parallel data
- Theory- (comparing algorithms) or practice-oriented (study on a data set) or both

References

Jin, Di et al. 2022. "Deep Learning for Text Style Transfer: A Survey." Computational Linguistics 48(1): 155–205.

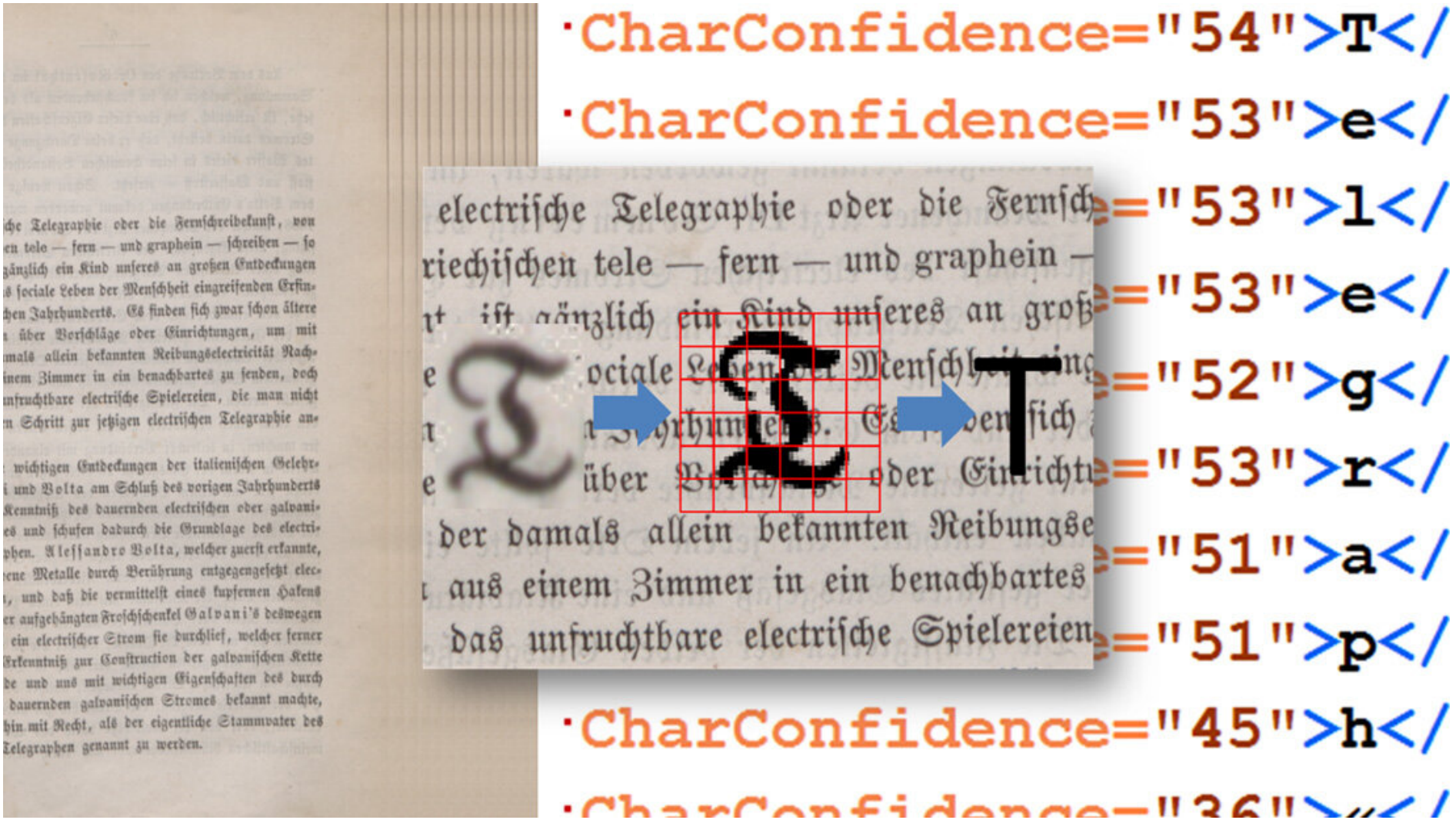
Projects

Zero-shot vs. Few-shot learning

- Important task in NLP: Getting data; how to train models with few or no data?
- What are the differences and how does it differ from "normal"training?
- Own experiment with few-shot and zero-shot learning

References

- https://joeddav.github.io/blog/2020/05/29/ZSL.html
-Brown, Tom B. et al. 2020. “Language Models Are Few-Shot Learners.”
Yin, Wenpeng, Jamaal Hay, and Dan Roth. 2019. “Benchmarking Zero-Shot Text Classification: Datasets, Evaluation and Entailment Approach.”


·CharConfidence="54">T</
·CharConfidence="53">e</
="53">l</
="53">e</
="52">g</
="53">r</
="51">a</
="51">p</
·CharConfidence="45">h</
electrische Telegraphie oder die Fernsch
riechischen tele — fern — und graphein —
ist gänzlich ein Kind unseres an groß
ociale
über
der damals allein bekannten Reibungse
aus einem Zimmer in ein benachbartes
das unfruchtbare electrische Spielereien

Projects

Optimal Character Recognition

- Important task in NLP: Getting data; historical texts often not available digitally
- OCR analyzes scanned documents to filter out the text
- What properties can scans have to optimize OCR?
- Comparing different OCRs

References

- https://tesseract-ocr.github.io/
- https://aws.amazon.com/de/textract/
- https://cloud.google.com/vision
- https://research.aimultiple.com/ocr-accuracy/

https://epaper.oeeee.com/epaper/A/html/2022-10/19/content_23987.htm#article

要闻

要闻 党的二十大特别报道 A07

二十大时光

党的二十大报告在广东省公安机关党员干部、民警、辅警中引发热烈反响

以新安全格局保障广东经济社会的高质量发展

党的二十大报告在广东省水利厅党员干部、厅直系统全体党员中引发热烈反响

聚力实施水利高质量发展蓝图，开创发展新天地

10月16日上午，中国共产党第二十次全国代表大会在京开幕，广东省水利厅领导班子成员、各处室（单位）主要负责同志，以及厅直系统全体党员，收听收看开幕会，认真聆听学习党的二十大报告。 通讯员供图

语音播报 分享到：

党的二十大报告在广东省水利厅党员干部、厅直系统全体党员中引发热烈反响

聚力实施水利高质量发展蓝图，开创发展新天地

来源：南方都市报 2022年10月19日 版次：GA07 作者：陈燕

10月16日上午，中国共产党第二十次全国代表大会在京开幕。广东省水利厅领导班子成员、各处室（单位）主要负责同志，以及厅直系统全体党员，收听收看开幕会，认真聆听学习党的二十大报告。 通讯员供图

10月16日上午，中国共产党第二十次全国代表大会在京开幕。广东省水利厅领导班子成员、各处室（单位）主要负责同志，以及厅直系统全体党员，收听收看开幕会，认真聆听学习党的二十大报告。

查看往期

上一篇

Projects

Web-Scraping

- Important task in NLP: Getting data; today's newspapers often only rely on online-publishing
- Web-scraping scans websites for their texts (e.g. news articles) and corresponding meta data
- Examplary web-scraping of certain news paper and/or topic

References

- `https://www.selenium.dev/`
- `https://beautiful-soup-4.readthedocs.io/`
-Khder, Moaiad. 2021. "Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application." International Journal of Advances in Soft Computing and its Applications 13(3): 145–68.

Midjourney AI-Generated Art Wins Colorado State Fair Prize

ERIC HAL SCHWARTZ on September 5, 2022 at 10:00 am

Projects

Text-to-Image-Generators

- Compare Dall-E, Midjourney and Stable Diffusion
- How do these image generators interpret text to generate an image?

References

-Ramesh, Aditya et al. 2021. "Zero-Shot Text-to-Image Generation."
-Rombach, Robin et al. 2021. "High-Resolution Image Synthesis with Latent Diffusion Models."
- `https://www.midjourney.com/`

Projects

Other topics might include

- Semantic Role Labeling
- Coreference Resolution
- Change point detection
- Sentiment Analysis
- Fake News/Bot detection
- GDP/stock price forecasting using texts
- ...