Friedrich-Schiller-Universität Jena
Institut für Informatik
Studiengang Informatik, B.Sc.

Transferring Relevance Judgments with Pairwise Preferences

Bachelorarbeit

Fabian Hofer
geb. am: 22.08.2002 in Eisenach

Matrikelnummer 199111

1. Gutachter: Prof. Dr. Matthias Hagen
2. Gutachter: Maik Fröbe

Datum der Abgabe: 13. März 2025

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Jena 07743, Germany, 13. März 2025

. . . . . .
Fabian Hofer

Zusammenfassung

This thesis will presents a methods for transferring relevance judgments from one dataset to another, which is based on the idea that the relevance of a document is determined by the similarity of its content to the content of other documents. The goal is using large language models to learn a representation of the content of documents and then uses this representation to predict the relevance of documents in a new dataset based on the ealier judgments.

Inhaltsverzeichnis

1 Introduction **1**

2 Related Work **4**

3 Methodology **7**
3.1 Datasets 9
3.2 Document Segmentation 10
3.2.1 Document Selection 10
3.2.2 Segmentation with spaCy 11
3.3 Passage Scoring 12
3.4 Candidate Selection 13
3.4.1 Document Retrieval from Target Dataset 13
3.4.2 Postprocessing of Selected Target Documents 14
3.4.3 Composing Final Candidates 15
3.5 Pairwise Preferences 16

4 Evaluation **19**
4.1 Inter-annotator Agreement 19
4.2 Rank Correlation Passage Scores 20
4.3 Rank Correlation 20
4.3.1 Spearman's Rank Correlation Coefficient 21
4.3.2 Kendall's Tau Rank Correlation Coefficient 21
4.3.3 Pearson's Correlation Coefficient 22
4.3.4 Greedy Correlation Coefficient 22
4.3.5 Unit testing Correlation Coefficients 22
4.3.6 Correlation Coefficient of Passage Scores 22
4.3.7 Candidate Selection 22

5 Conclusion & Further Work **25**

A My First Appendix **26**

Literaturverzeichnis **27**

Kapitel 1

Evaluation

This chapter focuses on the evaluation of the transfer pipeline. First, the evaluation methods used are introduced. Rank correlation is employed to compare the original relevance assessments with the relevance scores inferred during the pipeline process. Next, the individual steps of the pipeline are analyzed and evaluated to examine intermediate results and identify potential weaknesses. At the end, the final phase of the transfer pipeline is assessed by evaluating the original relevance assessments of the retrieval tasks with the inferred relevance scores produced by the pipeline.

1.1 Inter-annotator Agreement

In order to guarantee the validity of the inferred relevance assignments and the suitability of the dataset for training purposes, it is essential to evaluate the accuracy of the assigned labels, particularly in the context of natural language processing applications. This evaluation is typically performed by calculating the Inter-Annotator Agreement (IAA) between the annotators who labelled the dataset.

The Inter-Annotator Agreement is a statistical measure that quantifies the consistency between the annotations provided by multiple annotators in a collaborative annotation project. It quantifies the level of agreement or disagreement between annotators when labelling the same dataset, thereby providing insight into the objectivity and quality of the annotations. A high IAA indicates the presence of reliable and transparent annotation guidelines, whereas a low IAA may be indicative of task ambiguity or inconsistencies in annotator interpretation.

There are a variety of methods for calculating IAA, each with its strengths and weaknesses. In this thesis, I will utilise the Cohen's Kappa coefficient, which is a statistical measure that can be used to assess the degree of agreement

between annotators when categorising data. In this research, each document query pair will be assigned a relevance score of 0, 1, or 2. A score of 0 indicates that the document is not relevant to the query, a score of 1 indicates that the document is somewhat relevant, and a score of 2 indicates that the document is highly relevant.

- Score of 0, 1 or 2 correct?
- Formular
- IAA sources
  - Medium
  - Messverfahren zum Inter-annotator-agreement (IAA)
  - fortext
- cohen kappa sources
  - scikit-learn
  - Medium

1.2 Rank Correlation Passage Scores

The relevance scores of the passages were then used to assign relevance labels to the passages. To do this, I used to open source tool autoqrels. The tool can be used to automatically assign relevance labels to passages based on the relevance scores of the passages.

- Why was this step done?
- How were the relevance labels assigned?
- Usage of autoqrels
- What is a relevance label?

1.3 Rank Correlation

To assess the relationship between the relevance labels and the calculated scores, a correlation analysis was conducted. Relevance labels represent an ordinal scale with values of 0, 1, and 2, denoting increasing levels of relevance, while calculated scores range from -1 to 1, representing the model's prediction

for each document's relevance with respect to the query. Given the distinct nature of these data types, it was necessary to select a correlation measure that could accommodate both ordinal and continuous data without assuming a linear relationship.

1.3.1 Spearman's Rank Correlation Coefficient

Among the common correlation measures—Pearson's, Spearman's, and Kendall's—the Spearman's rank correlation coefficient was selected as the most appropriate measure for this analysis. Spearman's correlation is particularly suited to this context for several reasons:

1. Ordinal Nature of Relevance Labels: The relevance labels are ordinal, meaning they indicate an ordered relationship ($0 < 1 < 2$), but the intervals between values may not represent equal differences in relevance. Spearman's correlation is designed for ranked or ordinal data, making it ideal for assessing relationships where the exact distance between values is less meaningful than the order.

2. Monotonic Relationship Requirement: Spearman's correlation assesses whether there is a monotonic relationship between two variables rather than a strict linear relationship. This is important given that the relevance labels and calculated scores may not vary linearly but are expected to follow a general trend (e.g., higher relevance labels should be associated with higher calculated scores).

3. Robustness to Outliers and Non-Normality: Unlike Pearson's correlation, Spearman's correlation does not assume normally distributed data or homoscedasticity, which is suitable given the categorical nature of relevance labels and potential non-normal distribution of the calculated scores.

- +1: perfect monotonic agreement between rankings
- above $\tilde{0}.6$: strong monotonic relationship
- 0: no monotonic relationship
- -1: perfect monotonic decrease

1.3.2 Kendall's Tau Rank Correlation Coefficient

Kendall's Tau measures the ordinal association between two variables by comparing the number of concordant and discordant pairs of data. It's especially

useful for small datasets and ordinal data, offering a more nuanced view of monotonic relationships. The Kendall coefficient ranges from -1 to +1, with higher absolute values indicating stronger associations.

- +1: perfect agreement between rankings
- above $\tilde{0}.5$: strong positive association
- 0: no association between rankings
- -1: perfect disagreement between rankings

1.3.3 Pearson's Correlation Coefficient

Pearson's Correlation measures the linear relationship between two continuous variables. It assumes that the data is normally distributed and is sensitive to outliers. The result, called the Pearson correlation coefficient (r), ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no linear correlation.

- +1: perfect positive linear relationship
- above $\tilde{0}.7$: strong positive linear relationship
- 0: no linear relationship
- -1: perfect negative linear relationship

1.3.4 Greedy Correlation Coefficient

- What does intended mean?
- What is the greedy version of the correlation coefficients?
- How does it work?

To ensure that the correlation analysis is working as intended a greedy version of the correlation coefficients was implemented. This version of the correlation coefficients will first determine the best ...

1.3.5 Unit testing Correlation Coefficients

1.3.6 Correlation Coefficient of Passage Scores

1.3.7 Candidate Selection

Tabelle 1.1: Unit testing correlation methods for reference scores $[0.2, 0.7, 0.5]$ and three label sets: first two with expected correlation of 1, and last one with lower correlation. Methods include Spearman's ρ, Kendall's τ, and Pearson's r.

Comparative Set	**Default**			**Greedy**		
	τ	ρ	r	τ	ρ	r
$[0, 2, 1]$	1.00	0.99	1.00	1.00	1.00	1.00
$[0, 1, 1]$	0.82	0.92	0.87	1.00	1.00	1.00
$[0, 0, 1]$	0.00	0.11	0.00	0.50	0.50	0.50

Tabelle 1.2: 5-fold cross-validation of the passage scores.

Retrieval Model	**Default**			**Greedy**		
	τ	ρ	r	τ	ρ	r
BM25	0.4968	0.5937	0.5151	0.5463	0.5833	0.5468
DFR_BM25	0.5003	0.5986	0.5246	0.5553	0.5920	0.5585
DFIZ	0.4807	0.5716	0.4944	0.5065	0.5389	0.5062
DLH	0.5320	0.6345	0.5647	0.5851	0.6245	0.5935
DHP	0.4617	0.5466	0.4616	0.4926	0.5217	0.4877
DirichletLM	0.4361	0.5083	0.3955	0.4624	0.4881	0.4444
Hiemstra_LM	0.5738	0.6796	0.6043	0.6488	0.6930	0.6596
LGD	0.4975	0.5929	0.5152	0.5458	0.5807	0.5443
PL2	0.4949	0.5914	0.5149	0.5377	0.5719	0.5429
TF_IDF	0.5011	0.5984	0.5222	0.5464	0.5820	0.5464

Tabelle 1.3: Results for different strategies of candidate selection.

Dataset	**Naive**			**Nearest Neighbor**			**Union**		
	Precision	Recall	Documents	Precision	Recall	Documents	Precision	Recall	Documents
argsme21	0.0175	0.9586	100,000	0.0559	1.0000	38,807	0.0146	1.0000	123582

Anhang A

My First Appendix

This was just missing.

Literaturverzeichnis

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022. doi: 10.48550/ARXIV.2210.11416. URL `https://doi.org/10.48550/arXiv.2210.11416`.

Cyril W. Cleverdon. The significance of the cranfield tests on index languages. In Abraham Bookstein, Yves Chiaramella, Gerard Salton, and Vijay V. Raghavan, editors, *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Chicago, Illinois, USA, October 13-16, 1991 (Special Issue of the SIGIR Forum)*, pages 3–12. ACM, 1991. doi: 10.1145/122860.122861. URL `https://doi.org/10.1145/122860.122861`.

Maik Fröbe, Janek Bevendorff, Lukas Gienapp, Michael Völske, Benno Stein, Martin Potthast, and Matthias Hagen. Copycat: Near-duplicates within and between the clueweb and the common crawl. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2398–2404. ACM, 2021. doi: 10.1145/3404835.3463246. URL `https://doi.org/10.1145/3404835.3463246`.

Maik Fröbe, Lukas Gienapp, Martin Potthast, and Matthias Hagen. Bootstrapped ndcg estimation in the presence of unjudged documents. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Procee-*

dings, Part I, volume 13980 of *Lecture Notes in Computer Science*, pages 313–329. Springer, 2023. doi: 10.1007/978-3-031-28244-7_20. URL https://doi.org/10.1007/978-3-031-28244-7_20.

Lukas Gienapp, Maik Fröbe, Matthias Hagen, and Martin Potthast. Sparse pairwise re-ranking with pre-trained transformers. In Fabio Crestani, Gabriella Pasi, and Éric Gaussier, editors, *ICTIR '22: The 2022 ACM SIGIR International Conference on the Theory of Information Retrieval, Madrid, Spain, July 11 - 12, 2022*, pages 72–80. ACM, 2022. doi: 10.1145/3539813.3545140. URL https://doi.org/10.1145/3539813.3545140.

Sean MacAvaney and Luca Soldaini. One-shot labeling for automatic relevance estimation. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2230–2235. ACM, 2023. doi: 10.1145/3539618.3592032. URL https://doi.org/10.1145/3539618.3592032.

Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. Simplified data wrangling with ir_datasets. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2429–2436. ACM, 2021. doi: 10.1145/3404835.3463254. URL https://doi.org/10.1145/3404835.3463254.

Joel Mackenzie, Matthias Petri, and Alistair Moffat. A sensitivity analysis of the MSMARCO passage collection. *CoRR*, abs/2112.03396, 2021. URL https://arxiv.org/abs/2112.03396.

Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, 2008. doi: 10.1145/1416950.1416952. URL https://doi.org/10.1145/1416950.1416952.

Ronak Pradeep, Rodrigo Frassetto Nogueira, and Jimmy Lin. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *CoRR*, abs/2101.05667, 2021. URL https://arxiv.org/abs/2101.05667.

Mark Sanderson. Test collection based evaluation of information retrieval systems. *Found. Trends Inf. Retr.*, 4(4):247–375, 2010. doi: 10.1561/1500000009. URL https://doi.org/10.1561/1500000009.

Ellen M. Voorhees. The philosophy of information retrieval evaluation. In Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck, editors, *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001, Revised Papers*, volume 2406 of *Lecture Notes in Computer Science*, pages 355–370. Springer, 2001. doi: 10.1007/3-540-45691-0_34. URL https://doi.org/10.1007/3-540-45691-0_34.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 38–47. ACM, 2024. doi: 10.1145/3626772.3657813. URL https://doi.org/10.1145/3626772.3657813.