Word Sense Alignment and Disambiguation for Historical Encyclopedias

Thora Hagen[1], Fotis Jannidis[1], and Andreas Witt[2]

[1]University of Würzburg
[2]University of Cologne & IDS Mannheim

1 Introduction

Encyclopedias occupy an important and unique role in the preservation and production of knowledge; however most of this historical knowledge has not been digitally recovered yet. A graph network based on multiple historical encyclopedias, which also represents the diachronic information, makes it possible to access the state of knowledge at a specific time but also to explore the semantic changes over time. The main challenge in the creation of the graph is the alignment of the articles. Encyclopedias in general are seldom structured uniformly or follow universal guidelines, even when the editions are from the same editor (e.g. Brockhaus) as opposed to dictionaries for example. This does not only apply to certain design choices but also to the general style and flow of the text, which complicates the extraction of structured knowledge. This includes the scope of definitions of lemmas and how additional senses of one lemma are provided; whether those definitions might be included in the same article, given a separate section in the same article or given their own dedicated article entirely. More challenges include orthographic variation and different conventions in language use for referring to the same concepts over time, i.e. archaic synonyms (for example German Neuyork for New York, or Senfkohl for Rucola ('Arugula').

This paper will address the challenge of creating a knowledge graph from a corpus of historical encyclopedias with a special focus on word sense alignment (WSA) and disambiguation (WSD). More precisely, we examine WSA and WSD approaches based on article similarity to link messy historical data, utilizing Wikipedia as a ground-truth component – as the lack of a critical overlap in content paired with the amount of variation between and within the encyclopedias does not allow for choosing a "baseline" encyclopedia to align the others to. Additionally, we are comparing the disambiguation performance of conservative methods like the Lesk algorithm to more recent approaches, i.e. using language models to disambiguate senses.

2 Related Work

Word sense alignment and word sense disambiguation take up two of the most crucial steps for building a knowledge graph resource, in the sense of finding the concepts or subjects that make up the graph (e.g. Synsets in WordNet (Miller, 1998)). In general,

three different kinds of automatic WSA can be distinguished: text-to-text (as in this case), graph-to-graph and text-to-graph. Alignment procedures are thus either based on text or graph similarity measures (Matuschek, 2015). In 1986, one of the first approaches to WSD was introduced by Lesk (1986): the amount of overlap between pairs of contexts (in words) decides which sense is the correct one. Since then, many more approaches related to Lesk's algorithm have been tested and applied (see for example Pais et al. (2020)). For the German language especially, despite it's age, the Lesk algorithm still stands as one of the most essential WSD methods (Henrich and Hinrichs, 2012).

In recent years language models such as BERT (Devlin et al., 2019) have been used for WSD (Loureiro et al., 2021), where either fine-tuning (adapting the model to a specific task) or feature extraction (generating word embeddings) methods can be applied. Language models based on neural networks, in particular on the transformer architecture (Vaswani et al., 2017), have revolutionized natural language processing in recent years. In its pre-training step, a model "learns" a language by learning to predict masked words in large amounts of text data (masked language modeling). Similarly to training word embeddings, the "sense" of a word is derived from its neighboring word context.

SBERT (Reimers et al., 2019) in particular is a pre-trained transformer model that is specialized in efficiently generating embeddings for sentences and paragraphs rather than for single words only. It derives it's efficiency through a siamese network structure and it's effectiveness from new objective functions from the field of semantic textual similarity during pre-training; e.g. given a triplet of focus sentence, positive sentence, and negative sentence, adapt the network in a way that the distance between the focus and the positive sentence is smaller. Multiple pre-trained models have been published by the authors[1] for other researchers to use.

TSDAE (Wang et al., 2021)is another network designed to capture the meaning of sentences. It introduces a sequential denoising auto-encoder to pre-trained transformers: the training objective becomes reconstructing previously corrupted input (from swapping or deleting words for example). Additionally, the attention from the decoder has limited access to the output from the encoder, namely only the sentence vector and not the token vectors, forcing the model to "concentrate" on creating meaningful sentence embeddings.

3 Methods

The corpus for our experiment consists of six German encyclopedias[2] of the long 19^{th} century, as presented in table 1. To align the encyclopedia articles, we chose to integrate Wikipedia as an intermediate step, as due to the content and structural variance within the encyclopedia corpus, none of the works can represent a basis for WSA or WSD. It prevents superficial alignment based on simple stylistic similarities, as for example would be conceivable for encyclopedias by the same editors or encyclopedias from roughly the same time period (e.g. Brockhaus 1837 and Herloßsohn 1834 as opposed to Brockhaus 1911). By using Wikipedia as the largest online knowledge database, which now contains not only objects but also lexemes due to the integration

[1] https://www.sbert.net/docs/pretrained_models.html

[2] These encyclopedias are part of larger set of historical reference works which were all converted to TEI (Hagen et al., 2020). They can be downloaded at http://dx.doi.org/10.5281/zenodo.4039569.

Table 1: Overview of the encyclopedias.

Editor	Year	Description	# Entries	# Tokens
Brockhaus	1809-1811	General lexicon targeted at the general population.	6,960	1,186,000
Brockhaus	1837-1841	General lexicon briefly covering everyday subjects in a strictly non-scientific way, while focusing on illustrations.	7,049	2,604,000
Brockhaus	1911	General, pocketbook edition lexicon.	82,780	2,434,000
Herder	1854-1857	General lexicon with short explanations of various topics.	39,755	2,256,000
Herloßsohn	1834-1838	General encyclopedia explicitly targeted at middle-class women interested in education.	7,099	1,461,000
Meyer	1905-1909	Comprehensive general lexicon targeted at the general population.	156,264	17,437,000

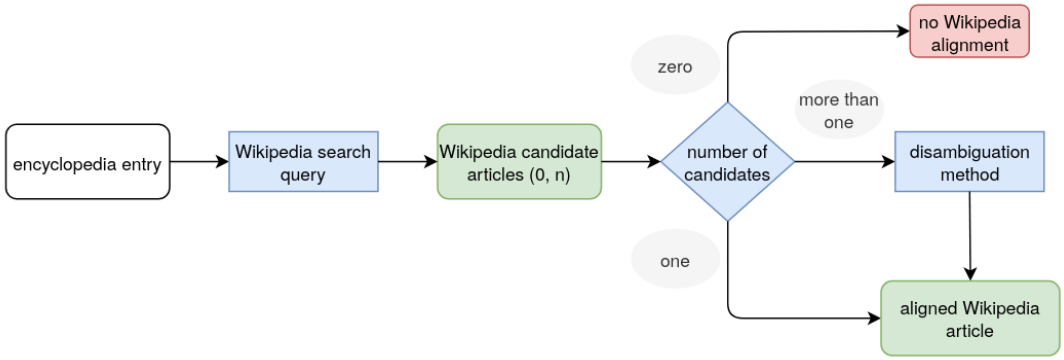

Figure 1: Proposed WSA and WSD workflow.

of Wiktionary, we hope to obtain the largest possible content coverage for our encyclopedias. Additionally, Wikipedia can find concepts via multiple different, partly also archaic, labels through Wikidata. An alignment of encyclopedia articles occurs when the articles get the same Wikipedia page assigned.

The first step of the alignment problem is to generate n Wikipedia page candidates for each article across all encyclopedias, which we do by looking up the headword of the articles in Wikipedia via the MediaWiki API. Beforehand, we performed orthographic normalisation of the encyclopedias using CAB (Jurish, 2008) to ensure that spelling variants have no impact on the Wikipedia search. The search usually returns the best possible match for the search term, and in certain cases, Wikipedia disambiguation pages. However, we want to explicitly make use of those disambiguation pages, which are often passed over by the judgement of the Wikipedia search algorithm. Therefore for every search term, we first try to look for a disambiguation page of the term specifically before simply looking up the term. The result of this step is a list of candidate Wikipedia pages as provided by the disambiguation page, a single Wikipedia page, or no page for every headword. One candidate equals one Wikipedia page summary (i.e. the text part which comes before the table of contents). A schematic depiction of the workflow is shown in figure 3.

We then test three different WSD methods: the Lesk algorithm to provide a baseline,

Table 2: Comparision of TSDAE, SBERT and Lesk algorithm performances (F1 scores). The column n≥2 presents the WSD performance only, and the last two columns further break down the WSD results.

method	overall	n≥2	10>n≥2	n≥10
TSDAE	0.63	0.38	0.47	0.29
SBERT	0.62	0.34	0.49	0.19
Lesk	0.61	0.27	0.39	0.17

SBERT, and TSDAE. To evaluate the performance of all three methods, we first manually annotated a set of encyclopedia articles with their corresponding Wikipedia page. Currently our evaluation dataset contains about 670 articles. The candidate generation process is deterministic, i.e. all three methods are evaluated on the same candidates for each encyclopedia article.

With SBERT, we chose the multilingual model (Reimers and Gurevych, 2020) for our experiment. For TSDAE, we trained a model[3] with 100,000 randomly selected sentences from Wikipedia on $\text{GBERT}_{\text{base}}$[4]. For each encyclopedia article and it's candidate list of Wikipedia summaries, we first derive the sentence embeddings from either SBERT or TSDAE, and then predict the closest candidate via the cosine distances between all candidate embeddings and the encyclopedia article embedding.

4 Evaluation and Discussion

The results (presented in table 2) show that TSDAE is the best performing method overall. The difference is especially noticeable for encyclopedia entries with more than 10 Wikipedia candidates, where TSDAE seems to be better at identifying the correct candidate in comparison to SBERT and Lesk. Using language models instead of Lesk in general seems to predominantly boost cases with 2 to 9 candidates.

To further dissect the results, we automatically classified our test data with one of four classes of encyclopedia articles based on their content: biographies (persons), locations, concrete objects and abstract concepts. We trained a simple bag-of-words SVM classifier with a different, labeled set of encyclopedia articles from the same corpus and manually corrected any errors for our test instances. Presented in table 3 are the results of the TSDAE model, listed separately by the class of the encyclopedia articles. To ensure an fair comparison between the different subsets of the test data, we excluded all test instances which do not have a Wikipedia page as a ground truth label assigned (i.e. true negatives) and where disambiguation would be necessary. Currently, the amount of those true negatives directly impacts the performance on the subset negatively, as we did not set a threshold for matching the closest candidate with an encyclopedia article (yet). As the resulting number of test instances for 10>n≥2 and n≥10 due to the split are quite low (median of 11), the F1 scores for these cases should be seen as an indication rather than concrete evidence.

The findings suggest that the method mostly struggles with named entities, for persons especially when 10 or more candidates are given. The performance does not necessarily seem to correlate with candidate recall however. It would be reasonable

[3] Code from https://www.sbert.net/examples/unsupervised_learning/TSDAE/README.html

[4] https://huggingface.co/deepset/gbert-base

Table 3: TSDAE performances (F1 scores) on different article categories. Test instances, where disambiguation would be necessary and which have no Wikipedia page assigned, are omitted in this experiment. The last column shows the median number of generated candidates for ambiguous terms (n≥2).

category (% of test set)	overall	n≥2	10>n≥2	n≥10	median number of candidates
person (15%)	0.47	0.56	0.82	0.29	14.0
abstract (34%)	0.66	0.69	1.00	0.50	16.5
location (23%)	0.58	0.27	0.27	0.27	9.5
object (27%)	0.70	0.56	0.50	0.62	9.5
all	0.66	0.51	0.64	0.39	10.5

to assume that a higher candidate recall would lead to a lower performance, as the algorithm has more candidates to decide between and the pure chance of picking an incorrect one is higher. But abstract concepts in particular, with the highest F1 score for ambiguous terms, also have the highest candidate recall. Intuitively, we can instead assume that the senses of texts about persons or locations are more similar amongst each other, and thus harder to disambiguate.

There are a few steps that can still be taken to further improve the results of the different disambiguation methods in particular. To improve the precision for named entities, the method could be adapted to focus on elements of both texts that are more likely to be different between the candidates, e.g. dates and other named entities. For biographies especially, dates of birth and places of birth are very likely to be given and are usually (also for humans) a reliable indicator of who's who (see the work by Ardanuy et al. (2016)).

Secondly, a threshold for (not) matching the nearest candidate with its encyclopedia article could be determined. The algorithm currently always matches the nearest Wikipedia page for ambiguous articles. As there are multiple instances in the test dataset, where no Wikipedia page exists but still multiple Wikipedia pages are found automatically, a distance threshold can help assigning a true negative label and could potentially increase the overall performance.

Additionally, the language models used here are currently not trained on the encyclopedic data. Continued pre-training for TSDAE on a couple of sentences from the encyclopedias for example could additionally boost the performance. Also, the input for the search query can still be optimized to increase the recall as well as precision, for example by extracting and using lemma synonyms from the encyclopedia article instead, if the encyclopedia formatting allows for a simple automatic extraction. Finally, we would like to increase our current test dataset size to further look into the difference between various subsets of the data and produce more stable results on that end.

The results show that aligning and disambiguating historical data is an extremely challenging task. However, the end goal of attending to this specific alignment problem can be an equally rewarding result: a lemma list that links concepts across time as a first step in creating a temporal knowledge graph from historical encyclopedias. As the lemmas are still connected to their respective encyclopedia articles, this list might already help researchers investigate semantic shifts of concepts during the 19^{th} and 20^{th} century.

Acknowledgements

Thanks to the German Text Archive (Deutsches Textarchiv), especially Bryan Jurish, who generously performed the orthographic normalization, lemmatization and tagging of all encyclopedia texts using CAB (Jurish, 2008).

This work is part of the "EncycNet" project, funded by the German Research Foundation (DFG).

References

Ardanuy, M. C., M. van den Bos, and C. Sporleder (2016). You shall know people by the company they keep: person name disambiguation for social network construction. In Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 63–73.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics.

Hagen, T., E. Ketzan, F. Jannidis, and A. Witt (2020). Twenty-two Historical Encyclopedias Encoded in TEI: a New Resource for the Digital Humanities. In Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Online, pp. 112–120. International Committee on Computational Linguistics.

Henrich, V. and E. Hinrichs (2012). A Comparative Evaluation of Word Sense Disambiguation Algorithms for German. In LREC, pp. 576–583.

Jurish, B. (2008). Finding canonical forms for historical German text. In Text Resources and Lexical Knowledge: selected papers from the 9th Conference on Natural Language Processing (KONVENS 2008), pp. 27–38. De Gruyter.

Lesk, M. (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC '86, New York, NY, USA, pp. 24–26. Association for Computing Machinery.

Loureiro, D., K. Rezaee, M. T. Pilehvar, and J. Camacho-Collados (2021). Analysis and Evaluation of Language Models for Word Sense Disambiguation. Computational Linguistics, 1–55.

Matuschek, M. (2015). Word Sense Alignment of Lexical Resources. Ph. D. thesis, Technische Universität Darmstadt.

Miller, G. A. (1998). WordNet: An electronic lexical database. MIT Press.

Pais, V., D. Tufiș, and R. Ion (2020). MWSA task at GlobaLex 2020: RACAI's word sense alignment system using a similarity measurement of dictionary definitions. In Proceedings of the 2020 Globalex Workshop on Linked Lexicography, Marseille, France, pp. 69–75. European Language Resources Association.

Reimers, N. and I. Gurevych (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Reimers, N., I. Gurevych, N. Reimers, I. Gurevych, N. Thakur, N. Reimers, J. Daxenberger, I. Gurevych, N. Reimers, I. Gurevych, et al. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. In Advances in neural information processing systems, pp. 5998–6008.

Wang, K., N. Reimers, and I. Gurevych (2021). TSDAE: Using Transformer-based Sequential Denoising Auto-Encoderfor Unsupervised Sentence Embedding Learning. arXiv preprint arXiv:2104.06979.