DomRec: Investigating Domain-centric Recommendation and Analysis of Entity Linking Methods

Anonymous ACL submission

Abstract

Detecting textual mentions and linking them to corresponding entities in a knowledge base is an essential task performed by a variety of existing entity linking approaches. This paper investigates the relationship between domains and annotator system performance. To this end, we employ a predictive approach arguing performance based on domain using learned topic vectors and machine learning models. We train our models on 6 common datasets for 12 state-of-the-art annotators. By analysing domain-specific characteristics across domains and methods, we demonstrate that no single technique excels across all domains, and that performance may be enhanced by selecting the most suitable system for each context. Our findings underline the importance of domain awareness in the development and deployment of text-processing systems, providing a pathway for more adaptable and robust methodologies. We release and open source all generated data, code and findings on our repository[1] and on Zenodo[2].

1 Introduction

Entity Linking (EL) serves as a fundamental task in natural language processing, aiming to associate entity mentions in text with corresponding entries in a given knowledge base. Despite significant advances in EL, the performance of these systems can vary substantially across different textual domains, presenting a challenge for their deployment in diverse applications, such as knowledge enrichment, semantic search, question answering and overall enhancing information retrieval. EL approaches often lack the flexibility required to excel across varying domains – to our knowledge, a commonly shared assumption (Ding et al., 2024; Shavarani and Sarkar, 2023; João et al., 2020), but never explicitly proven. The development of domain-specific knowledge graphs as well as the assumed lack of effectiveness from general-purpose EL approaches over time resulted in the need for domain-specific EL techniques emerging (Shen et al., 2015). Particularly noteworthy is the area of *biomedical entity linking* with special-purpose automated annotation approaches targeting its specific challenges (French and McInnes, 2023); other areas for domain-specific annotations include products, entertainment, finance and tourism (Shen et al., 2015). Extant entity linking research has focused on creating and identifying coherent contexts within documents in order to successfully disambiguate candidate entities (Zu et al., 2024; Ayoola et al., 2022; Christmann et al., 2022; van Hulst et al., 2020; Nanni and Fabo, 2016; Flati and Navigli, 2014; Han and Sun, 2012). However, to the best of our knowledge, no research has attempted to identify a deeper link between domains and system performance – a lack we specifically address in this paper:

We investigate the effectiveness of text-processing techniques and their performance uniformity across domains, and whether some systems exhibit high performance in certain domains despite failing to do so in the general one. Hence, to look into the matter in further detail, we developed a prediction-based approach inferring a system for a given domain by learning from text-based topic vectors.

Further, we systematically analyse the relationships between text domains and system performance on mentions, providing insights into how different systems can be tailored or selected for specific domains. Our findings reveal that the choice of method benefits from domain-dependent decision-making, with the potential to enhance accuracy in practical applications.

In this paper, we specifically address the following research questions:

[1] https://anonymous.4open.science/r/domrec-6805/

[2] https://doi.org/10.5281/zenodo.14498260

RQ1 Is there a link between domain and system performance?

RQ2 Is document domain a sufficient information source to identify the best-performing method?

In attempting to qualitatively respond to the above questions, our developed contributions in this paper are as follows:

1. Domain-specific analysis of state-of-the-art datasets and Named Entity Recognition and Disambiguation (NERD) methods. We show that domain constitutes a significant and non-random predictor.

2. Model architecture for domain-sensitive recommendation of entity linking approaches.

3. Development, evaluation and release of annotated data from 12 entity linking systems for 6 data sets (`AIDA CoNLL-YAGO`, `RSS-500`, `Reuters-128`, `News-100`, `KORE50`, `MedMentions`); embeddings; topics; code; approach and metadata.

Thus, we present our methodology based on topic modelling, followed by an evaluation of its effectiveness across identified domains. We further discuss implications of our findings for the broader field of entity linking, emphasizing the importance of domain awareness in the development and deployment of NERD approaches.

2 Related Work

The relationship between textual domains and the performance of Natural Language Processing (NLP) systems has garnered considerable attention in recent years with large language models taking centre stage. In this section, we draw the links between our research exploring the domain-to-linker relationship and the various approaches developed to enhance EL performance across diverse domains. For the sake of identifying a variety of domains, topics and contexts, we make use of topic modelling techniques, allowing for the unsupervised detection and grouping of related and mentioned texts and phrases. In our research, we experimented with two state-of-the-art topic modelling techniques. One of which was Top2Vec (Angelov and Inkpen, 2024), a method learning topics directly from latent document representations by recognising dense regions within a given embedding space. Based on dense regions, it extracts groupings of most representative words given in order to define meaningful topics. Another approach we employ for our experiments is BERTopic (Grootendorst, 2022), a topic model utilising BERT (Devlin et al., 2018) embeddings combined with clustering techniques to find meaningful topics. To the best of our knowledge, state-of-the-art research in the domain of entity linker recommendation is scarce. In (João et al., 2020), the authors attempt to leverage systems' individual strengths on a mention to mention basis, recommending a particular linking technique. They acknowledge the assumed effect of domains, but did not investigate its impact. Additionally, the authors only utilised 3 entity linking systems (including Babelfy and TagMe - both systems also included in this paper) and evaluate on 3 datasets. Noullet et al. present a framework in (Noullet et al., 2021) with a baseline linker recommendation module. Their approach uses a support vector machine model, presenting it as a stepping stone to the broader research audience.

In (Flati and Navigli, 2014), authors introduce concepts from word sense disambiguation to entity linking and in combination with dense subgraph heuristics aim to create a consistent and high-coherence context, yielding qualitative disambiguation results. With CLOCQ (Christmann et al., 2022), Christmann et al. improve upon existing approaches by working four levels of signals into their ranking algorithm. They introduce word-level scores for matching and relatedness, but further also include text-wide coherence and connectivity for disambiguation results along with dynamic candidate set size considerations. DBpediaSpotlight (Mendes et al., 2011) utilises a four-stage pipeline including spotting through an extended set of label lexicalizations identified and part-of-speech tagging mechanisms, a candidate selection step and an entity disambiguation step utilising vector space model representations with heuristics including customised inverse candidate frequency metrics.

Regarding annotated datasets, `AIDA-CoNLL-YAGO` (Hoffart et al., 2011) links entities to the YAGO, Wikipedia or Freebase Knowledge Base (KB), providing a Named-Entity Recognition (NER), Entity Disambiguation (ED) and EL dataset. KORE50^{DYWC}(Noullet et al., 2020) particularly contains less frequent and hard-to-disambiguate mentions of entities, making up a gold-level standard entity linking dataset, which links to various knowledge graphs or bases: DBpedia, YAGO, Wikidata and Crunchbase. Due to its small size, it mainly functions for evaluation purposes in related research. Further, with the N3 collection (Röder et al., 2014), authors introduce a collection made up of 3 data sets: `News-100`, `Reuters-128` and `RSS-500`. News-100 is a dataset made up of 100 German news articles. Reuters-128 includes a subset of articles from the Reuters-21578[3] dataset, initially created for text categorization. Whereas RSS-500 is a corpus created from 1,457 RSS feeds as initially released by (Goldhahn et al., 2012) and contains a wide range of topics ranging from politics, business and science from major global news outlets. Another dataset investigated in this paper is `MedMentions` (Mohan and Li, 2019). It is derived from the MEDLINE and PubMed corpus, linked to the UMLS knowledge base and constitutes a large-scale dataset for specialised biomedical entity linking.

3 Methodology

This paper utilises a predictive approach to analyse the domain-dependency of entity linking systems by employing a *naive* input representation for supervised machine learning models. Furthermore, this paper also experiments with more meaningful input representations for the purpose of annotator recommendation and analyses thereof. To this end we designed experiments and trained supervised learning models on commonly used datasets. We annotate these datasets with extant EL workflows, compute metrics for models to learn towards and adapt these results into a series of *best-vs.-all* datasets. Thus, each datapoint is labelled with its best-performing annotator. We employ systems accessing DBpedia (or Wikipedia) information to ensure knowledge base-conforming comparability for entities and spans.

Our experiments cover 3 different input representations to analyse input signal type significance for our employed learning methods: contextualised document embeddings, 1-hot encodings and a combination of topic and document embeddings.

[3]https://www.daviddlewis.com/resources/testcollections/reuters21578/

For our document embedding-related experiments, we generate *contextualised* document embeddings with the help of BERT (Devlin et al., 2018). These latent representations are known for integrating a wealth of contextual knowledge, potentially already including implicit domain information. Employing these embeddings serves the purpose of setting a baseline in regards to information provided to the models, as their latent representations transmit a depth of information even without explicit domain information.

In contrast, our 1-hot encoding representation maps a given document to one of 35 automatically detected topics. This representation serves as a check for sufficiency of information solely based on naive and highly restrictive topic information. The simplicity of the representation is meant to ascertain the presence of domain-based bias. Finally, we designed an experiment combining topic and document embedding information with the latter aspect being processed via dimensionality reduction to ensure equal initial feature weights, verifying whether explicit topic or domain information may help latent document representations further improve prediction results.

We generate annotation data based on 12 different systems for 6 data sets[4] with help of the linking framework described in (Noullet et al., 2021), adhering to data generation in pre-existing and interoperable formats.

In the following, we describe designed experimental setups for our multiclass classification task along with techniques necessary for the completion thereof.

3.1 Dataset Creation

We chose the following 12 systems Babelfy (Flati and Navigli, 2014), CLOCQ (Christmann et al., 2022), DBpediaSpotlight (Mendes et al., 2011), Falcon 2.0 (Sakor et al., 2020), OpenTapioca (Delpeuch, 2020), ReFinED (Ayoola et al., 2022), Radboud Entity Linker (REL) (van Hulst et al., 2020), ReLiK (Orlando et al., 2024), spaCy (Explosion, 2024), SpEL (Shavarani and Sarkar, 2023), TagMe (Piccinno and Ferragina, 2014), and TextRazor (TextRazor Ltd., 2024) for our dataset creation. Our choice of methods was motivated by the state-of-the-art performance, stability, widespread use in existing research to

[4]https://anonymous.4open.science/r/domrec-6805/

Table 1: Identified, annotated & grouped Topics and their abbreviations.

Topic	Abbr.	Subtopics
Medical Research	MED	
Pol. Conflict News	POL	Chin. Sociop., Pol. Elections, Kurd. Pol., Is.-Pal. Relations, Conflict & Pol. Violence
Fin. Market Trends	FINMA	Commodity Trading Dyn., Fin. Perf. Metrics, Fin. Market Insights
Gov. & Administration	GOV	
Sports Analysis	ANALYSIS	Cricket Perf. Metrics, Soccer Leagues and Comp., Int. Socc. Comp., Socc. and Player Profiles, Football League Anal.
Game Strat. & Players	PLAYERS	Sports Coach. & Mgmt, Baseball Inning Details, Football & Players, Baseball & Players
Corp. Market Insights	CORP	Corp. Collab., Stock Market Insights, Corp. Announcements
News & Celebrities	CELEBNEWS	Notable Athl. & Celeb., Research and Reports, News Outlets & Reporting
World Champ.	CHAMP	Tennis Tournaments and Champ., Athletic Achievements & Champ.
Sports Event Roundup	EVENT	Tennis Tournament Highlights, MLB Teams & Matchups, Sports Highlights and M.
League Matches	MATCHES	Sports League Standings, MLB Team Rivalries, Soccer Leagues and M.
German Language	GRMN	Misc. German Phrases, German Language Constructs

increase research benefit, compatibility, and up-to-dateness of results. Further, in order to maximise comparability, be able to analyse and create recommendations based on annotator performance, we employed 6 commonly-used datasets spanning a variety of domains (`AIDA-CoNLL-YAGO` (Hoffart et al., 2011), `MedMentions` (Mohan and Li, 2019), `RSS-500` (Röder et al., 2014), `Reuters-128` (Röder et al., 2014), `News-100` (Röder et al., 2014), `KORE50` (Noullet et al., 2020)).

In a data preparation step, we annotate all datasets with our repertoire of systems using the entity linking framework presented in (Noullet et al., 2021) to this end. Each annotator's results are ranked by computing F1 scores for each document's mentions. Taking these scores into account, we assign to each document the label of one or (in case of equal scores) *multiple* top-ranked annotators. Each attributed label represents one system per input signal our models learn to predict. We do not apply tie-breaking, instead duplicating data points with best-performing systems' labels to allow for a degree of flexibility within our data and trained models.

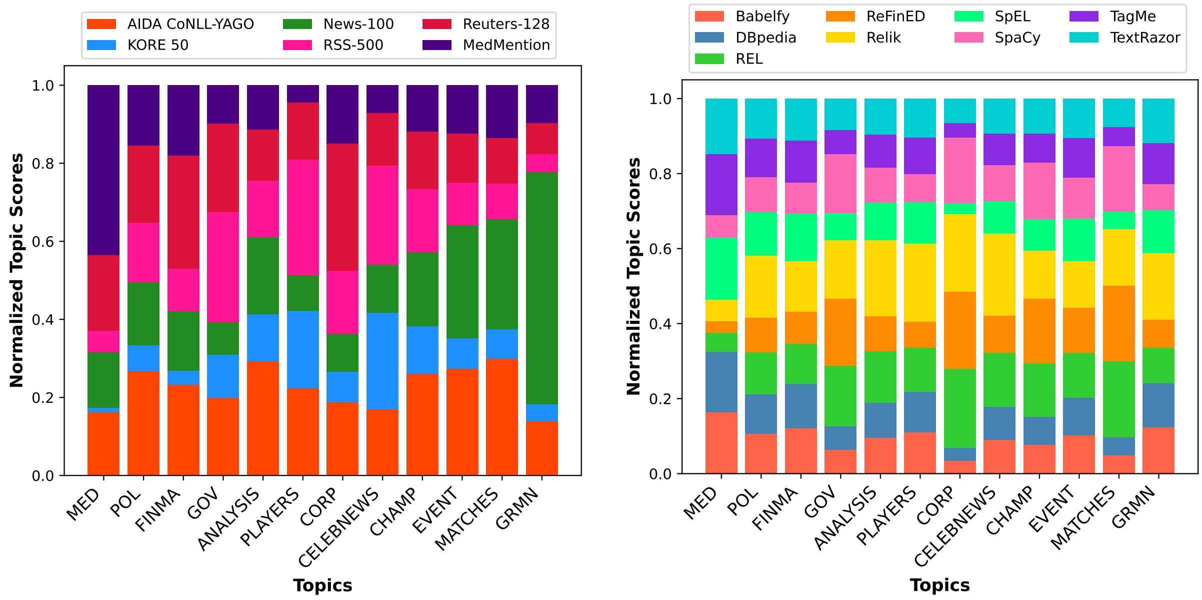
3.2 Document Embeddings

Utilising BERT (Devlin et al., 2018), we generate contextualised sentence and document embeddings each mapped to one or more annotator labels. With these we investigate in a simple yet powerful fashion the potential of a highly specific input representation to the annotation method reaching the best-performing F1 annotation score. We employ the `bert-base-cased` case-sensitive version of BERT trained on the English language corpus made up of English language Wikipedia[5] and Toronto BookCorpus (Zhu et al., 2015).

3.3 Topic Model

Applying topic modelling techniques, we discover abstract topics occurring in a collection of unstructured text documents. Extracting topics enables better understanding of the dataset by identifying underlying themes and implicit structures within the data in an explicit fashion. For our experiments, we employed two state-of-the-art topic modelling techniques, namely Top2Vec (Angelov and Inkpen, 2024) and BERTopic (Grootendorst, 2022). Our experiments yielded similar results with negligible differences with both employed topic mod-

[5] `https://en.wikipedia.org/wiki/English_Wikipedia`



(a) By normalised **dataset** topic scores for each topic

(b) By normalised **system** topic scores for each topic

Figure 1: Proportions of Dataset and System topic scores for the whole dataset

elling techniques (see Jupyter Notebooks[6,7,8] on our GitHub page for qualitative performance comparisons). Therefore, we chose to use Top2Vec, the current state-of-the-art method in this field. All provided experimental results and visualisations in this paper were performed using Top2Vec and `universal-sentence-encoder` embeddings.

As we consider it valuable to investigate the mapping function of *identified topic* to method class label, we explore the effect through use of explicit topics encoded as 1-hot vectors - each indexed position representing a respective topic. Being a radical oversimplification of the recommendation problem, this allows us to detect the degree of skewness incurred by annotation methods based on domain and whether domain information by itself is sufficient meaningful recommendations. Reaching a relatively good performance despite the simplification of a topic-to-class interpretation would therefore imply a potential gap to be exploited in qualitative result optimization endeavours.

For this line of experiments, we automatically extract topics within all of our investigated datasets in an unsupervised fashion. In Table 1, we list all 35 identified topics through Top2Vec via hierarchical density-based clustering. We further apply the topic model's integrated hierarchical topic reduction technique, reducing the number of topics to 12 grouped topics to avoid overcrowding for the sake of meaningful visualisation and figure simplification. Each abstract topic is labeled through use of a state-of-the-art large language model[9] based on topic documents' common textual features. Upon grouping of subtopics into parent topics, each parent topic's label is adjusted to match its encompassing members' contents and assigned an abbreviation for simplified reference. We note that the identified topics match our employed datasets' source data.

Further, in Fig. 1a and Fig. 1b we visualise the topic distribution for each dataset and EL methods, respectively. We design experiments utilising both document-specific topic vectors, as well as 1-hot encoding representations thereof, document embeddings and combinations thereof, among others. Document-specific topic vectors approximate document embedding representations with dimensions equal to the number of topics. In contrast, our 1-hot encoding representation is designed to radically define exactly one *main* topic per document. Please note that this is intended to be a highly limited signal with the purpose of identifying topic relevance for the linker recommendation task in mind.

[6]https://anonymous.4open.science/r/domrec-6805/bertopicVStop2vec.ipynb

[7]https://anonymous.4open.science/r/domrec-6805/evaluation_bertopic.ipynb

[8]https://anonymous.4open.science/r/domrec-6805/evaluation_top2vec.ipynb

[9]https://huggingface.co/docs/transformers/model_doc/llama3

3.4 Dimensionality Reduction

In order to allow for topic and embedding vectors to have similar potential for generalization, we reduce document embedding vector dimensionality for our experiments that jointly utilise topic and document embedding signals in the learning process. We herewith mean to balance the effects of dimensional imbalance between topic and document vectors on the learning process of our employed machine learning methods. Thus, we apply dimensionality reduction on high-dimensional document vectors using *Principal Component Analysis* (PCA). As the name indicates, this technique identifies so-called principal components, along which the variation is highest, and projects the data onto these components. Through this process, some of expressivity of contextualised document embeddings is lost, but we consider this loss negligible in order to allow for a more balanced feature size between latent topic and document representations.

4 Results

We used multiple input representations and trained a variety of supervised machine learning models allowing us to predict an appropriate linking methodology for each. These models further allow us to analyse the data from different aspects due to their underlying assumptions and architectures. Due to wanting to cover multiple domains within our training set, we evaluate our models on the combined datasets with a 70%/30% train-to-test split ratio. In Table 2 we include weighted scores for F1, precision and recall, as well as Recall@2 and Recall@3 for all input representations. Weighted metrics were computed to account for label imbalance from differing amounts of datapoints per domain. Overall for all representations, Support Vector Machine (SVM) or Multilayer Perceptron (MLP) perform best, trading between first and second places in most cases. Unsurprisingly, Random Forest (RF) models perform well on easily categorizable input features as displayed in our 1-hot encoding and combined representation experiments. We note that our experiments utilising both document embeddings as stand-alone signals and in combination with topic vectors only diverge minimally despite the latter yielding slightly better results, particularly for MLP.

Using document embedding vectors as predictors yielded some of the highest precision, recall and F1 scores, representing the most informative latent representation of our data. Further, this proves the link expressed as intuition in prior research between a document's content and an expected top-performing method (label) due to every employed machine learning model being able to successfully predict target labels. Our trained random forest model achieved an F1 score of 40.44% despite intuitively being ill-suited to classifying within an embedding space, yet substantially better than random guesses as illustrated by dummy classifiers (most frequent: 9.16%, uniform distribution: 13.37%). Also, *most frequent* implies a system that was the top-performing one for the most input documents. Our best F1 prediction performance was produced by a MLP model (45.08%) in large parts due to a 4.4% improvement (44.28%) over SVM (39.86%) in precision despite being ranked second in recall (46.86%) to our SVM (48.81%). This further exemplifies the context-sensitive nature of our employed document embeddings, containing information on a word as well as contextual levels.

While models based on our 1-hot encoding representation intentionally only possess a very limited range of input signals, all trained models seem to relatively easily adapt to the simple data structure, reaching similar if not identical results as is the case with RF, SVM and MLP. We note that all scores, particularly recall (43.92%–44.11%) scores are above ground truth-based baseline results for *most frequent* (23.8%) and *uniform* distributions (11.62%) in every case.

As our recall values for SVM spike from 48.71% to 70.44% for recall@2 in a combined setting, the 21.73% difference indicate a certain degree of tie-breaking ambiguity within the prediction. It seems as though our recommendation regardless of model used is hampered from having to choose one from among multiple ideal systems within a context, causing confusion. This could be an indication that multiple systems have similar detection results, making it inherently difficult for a model to choose the right one. Reaching meaningful results despite for the more limiting metrics highlights the importance of domain importance even further.

Additionally to recommender evaluation metrics, F1 scores for recommendation result performance on the annotation task (Table 2, *right of* ||) in relation to the *Oracle* are displayed: analogously following recommender metric trends, document embeddings (F1: 0.3418), as well as combined topic & document embeddings (F1: 0.3413) perform best. Results show that no chosen system has been conceived to function on all chosen datasets nor domains. In particular, MedMentions is a large single-domain dataset and its mentions of medical entities are hard for any general-domain annotator to correctly annotate, worsening annotator results.

4.1 Domain-specificity

In a second part of our evaluation, we focus on analysing underlying topic distributions across different datasets (Fig. 1a) and annotators (Fig. 1b). We process each dataset through our topic model including the topic rankings and their corresponding scores, reflecting their importance for a given input. A larger relative bar indicates a prevalence of this topic in the case of datasets and a more frequent top performance in the case of annotators for a given topic. For instance, one can see that the dataset News-100 contributes to the topic *German Language* (GRMN) as can be expected due to its makeup consisting of German news articles. Similarly, MedMentions mainly contributes to the domain of *Medical Research* (MED), an expected outcome considering its biomedical domain-specific nature. In Fig. 1b, system strengths and weaknesses can be observed, e.g.: SpEL, TagMe, DBpedia and Babelfy are shown to be unsuited to the `CORP` domain while performing well in the `MED` domain.

While Fig. 1 presents an overview of our dataset, Fig. 2 shows the distribution across topics for our ground truth (Fig. 2a), document embeddings (Fig. 2b), 1-hot encoding (Fig. 2c) as well as combined topic & document embeddings (Fig. 2d) when predicted with a MLP. Despite evaluation metrics not changing substantially between document embeddings and our combined approaches, it is noticeable that certain domains undergo substantial shifts. For instance, while for Fig. 2b ReFinED is not predicted at all for MED despite ground truth ideally requiring for it to, both 1-hot topic representation alone, as well as the combined experiments (Fig. 2d) include it again – approaching the ideal distribution. Further, spaCy never reaches best-performing results for the `MED` domain in our ground truth and is correctly never recommended in said domain for the naive 1-hot (Fig. 2c) experiments, in contrast to the contextualised domain models. In our naive approach of pure topic-based linker recommendation (1-Hot), one notices that some (Babelfy, DBpedia Spotlight, SpEL) of the usually present systems have disappeared entirely. This implies that datapoints previously predicted as one of these are absorbed by one or multiple of the other methods. Upon analysis of confusion matrices, we have discovered that our model has a higher likelihood of misclassifying Babelfy and DBpedia Spotlight for TagMe primarily and for ReLiK next. Further, we see that SpEL is mainly absorbed by TagMe which can be observed nicely when comparing the ground truth data with document embeddings-based models. As such, it stands to reason that due to their absence in the naive models, predictions ideally classified towards these methods, would be partially absorbed by TagMe and ReLiK. This phenomenon can be observed for instance by comparing Fig. 2a and Fig. 2c: in `MED`, TagMe goes from a relatively equal share with TextRazor towards clearly dominating the domain. From looking at our data visualizations, the ambiguity between these may be due to them having relatively similar results within varying domains and alternating for the top-ranked position. Moreover, interestingly TextRazor disappears completely from its weakest domain (`CORP`) from the embedding to the combined experiments, accurately representing desired ground truth data predictions.

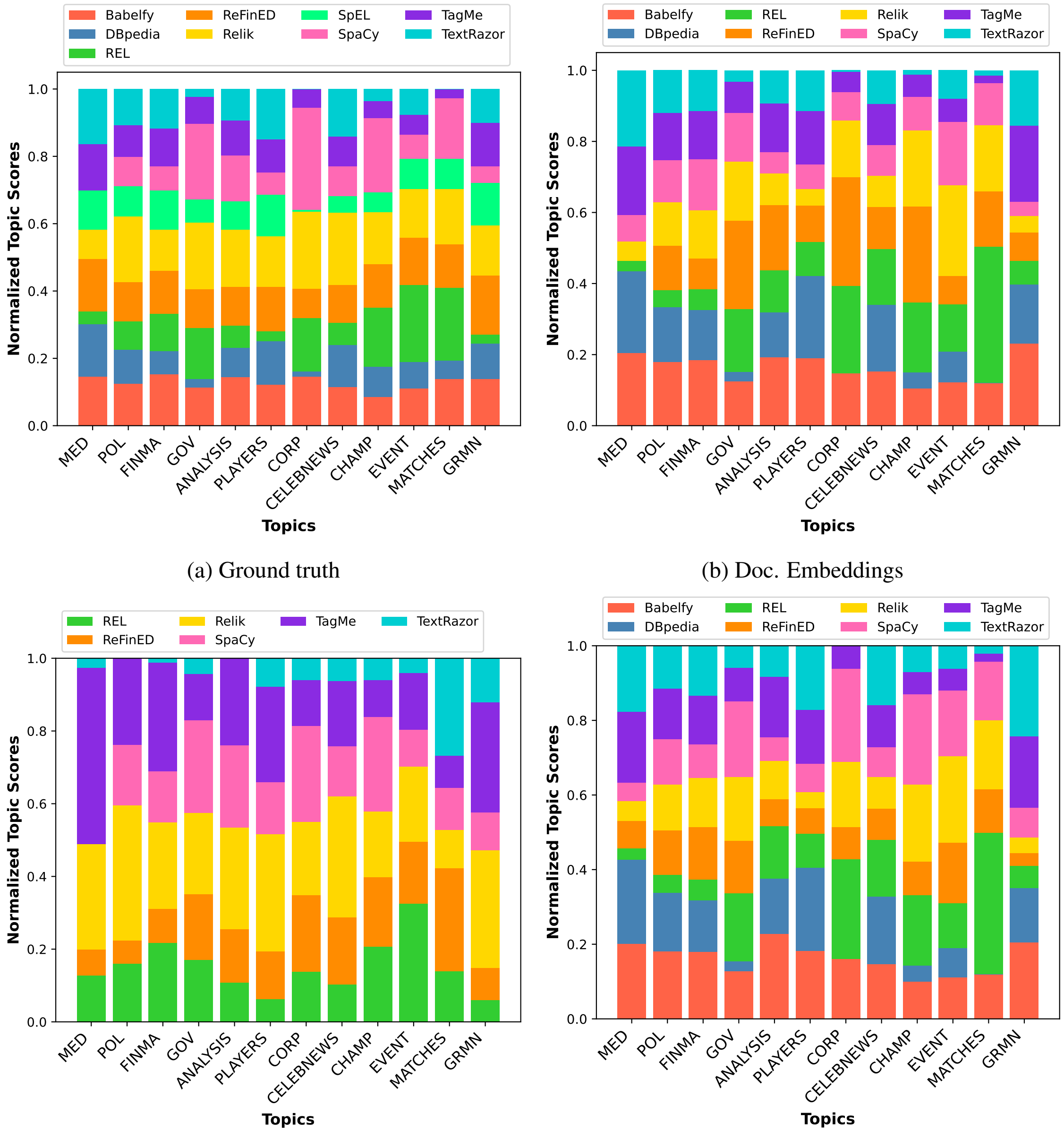
5 Conclusion

In this paper, we show that despite naive assumptions regarding domain representations, a link between topic and optimal choice of system can be witnessed throughout domains. Further, this assumption holds true despite existing techniques seemingly reigning supreme for given datasets. This could imply a large potential uplift when employing the appropriate techniques, possibly even through combination of multiple technologies to generalise across domains. Our analyses show that some methods perform similarly well to each other in a given domain, potentially creating tie-breaking issues when it comes to recommendation as indicated by large jumps in performance between recall@1 and recall@2 metrics. Moreover, we provide all findings and data in standard machine-readable formats and upload them to freely accessible platforms.

Finally, we have discovered that utilising highly naive signals to a recommendation, ambiguous results are swallowed up by one or more prediction labels, hinting at a degree of system result overlap within given domains.

Table 2: Model Evaluation weighted metrics (left of ||) and system task evaluation (right of ||).

Representation (Dataset)	Model	F1	Precision	Recall	Recall@2	Recall@3	F1 (Task)	Relative F1
Ground Truth	Oracle	1.0	1.0	1.0	1.0	1.0	0.3748	100%
	Dummy (MF ⟶ Best)	0.0916	0.0567	0.2381	0.3013	0.3856	0.2953	78.7%
	Dummy (Uniform)	0.1337	0.1842	0.1162	0.2118	0.3185	0.1817	48.4%
Doc. Embeddings	Random Forest	0.4044	0.4034	0.4563	0.6740	0.8315	0.3360	89.6%
	SVM	0.4254	0.3986	**0.4881**	**0.7044**	**0.8543**	**0.3418**	91.2%
	k-NN	0.4255	0.4154	0.4402	0.6470	0.7879	0.3353	89.4%
	MLP	**0.4508**	**0.4428**	0.4686	0.6907	0.8391	0.3393	90.5%
1-Hot Encoding (Topic)	Random Forest	0.3077	0.2645	0.4407	0.3177	0.4357	0.3219	85.9%
	SVM	**0.3198**	**0.3191**	0.4392	**0.3218**	0.4567	0.3248	86.7%
	k-NN	0.2501	0.2233	0.3444	0.3049	0.4267	**0.3266**	87.1%
	MLP	0.3079	0.2518	**0.4411**	**0.3218**	**0.4583**	0.3240	86.4%
Topic & Document Embeddings	Random Forest	0.4209	0.4157	0.4625	0.6622	0.8083	0.3344	89.2%
	SVM	0.4249	0.3983	**0.4871**	**0.7044**	**0.8519**	**0.3413**	91.1%
	k-NN	0.4207	0.4118	0.4331	0.6475	0.7922	0.3349	89.3%
	MLP	**0.4448**	**0.4500**	0.4577	0.6802	0.8382	0.3401	90.7%



(a) Ground truth

(b) Doc. Embeddings

(c) 1-Hot Encoding (Naive)

(d) Topic & Document Emb.

Figure 2: Topic distribution for Systems (Test data).

6 Limitations

Our approach generates a *best-vs-all* type of dataset ranking on a document-level based on F1 scores. Despite our choice of focusing F1, there are valid arguments to account for precision or recall instead. Further, the design of *best-vs-all* comes with some disadvantages regarding recommendation, such as not considering second-placed systems. These could be consistently barely below the best-ranking annotation system, but still overtake all other ones on average. Due to the nature of the problem we are trying to solve, it is likely for there to be duplicate best systems for a given document. As such, we generate multiple labels for the same datapoint to a non-insignificant amount, generalising, but also potentially confusing our model due to the similitude of the input signals expecting varying outputs and forcibly dragging recommendation results down. Further, despite having put considerable effort into our choice of systems, it would be great to have more specialised domain-specific linking approaches to make use of – something we intend to look into in the future – to have an in-depth discussion on domain-specific predictions and effective ways of exploiting domain information for the benefit of annotation quality and robustness.

References

Dimo Angelov and Diana Inkpen. 2024. Topic modeling: Contextual token embeddings are all you need. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 13528–13539. Association for Computational Linguistics.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum. 2022. Question entity and relation linking to knowledge bases via CLOCQ (full paper). In *Joint Proceedings of SemREC 2022 and SMART 2022 co-located with 21st International Semantic Web Conference (ISWC 2022), Hybrid event, Hangzhou, China, October 24-27, 2022*, volume 3337 of *CEUR Workshop Proceedings*, pages 33–47. CEUR-WS.org.

Antonin Delpeuch. 2020. OpenTapioca: Lightweight Entity Linking for Wikidata. In *Proceedings of the 1st Wikidata Workshop co-located with the 19th International Semantic Web Conference*, volume 2773 of *Wikidata'20*. CEUR-WS.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Yifan Ding, Amrit Poudel, Qingkai Zeng, Tim Weninger, Balaji Veeramani, and Sanmitra Bhattacharya. 2024. Entgpt: Linking generative large language models with knowledge bases. *CoRR*, abs/2402.06738.

Explosion. 2024. spaCy, Industrial-Strength Natural Language Processing.

Tiziano Flati and Roberto Navigli. 2014. Three Birds (in the LLOD Cloud) with One Stone: BabelNet, Babelfy and the Wikipedia Bitaxonomy. In *Proceedings of the Posters and Demos Track of 10th International Conference on Semantic Systems*, volume 1224 of *SEMANTiCS'14*, pages 10–13. CEUR-WS.org.

Evan French and Bridget T. McInnes. 2023. An overview of biomedical entity linking throughout the years. *J. Biomed. Informatics*, 137:104252.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 759–765. European Language Resources Association (ELRA).

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR*, abs/2203.05794.

Xianpei Han and Le Sun. 2012. An entity-topic model for entity linking. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, July 12-14, 2012, Jeju Island, Korea*, pages 105–115. ACL.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 782–792. ACL.

Renato Stoffalette João, Pavlos Fafalios, and Stefan Dietze. 2020. Better together: An ensemble learner for combining the results of ready-made entity linking systems. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, SAC '20, page 851–858, New York, NY, USA. Association for Computing Machinery.

Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia Spotlight: Shedding Light on the Web of Documents. In *Proceedings the 7th International Conference on Semantic Systems*, I-SEMANTICS'11, pages 1–8. ACM.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with UMLS concepts. In *1st Conference on Automated Knowledge Base Construction, AKBC 2019, Amherst, MA, USA, May 20-22, 2019*.

Federico Nanni and Pablo Ruiz Fabo. 2016. Entities as topic labels: improving topic interpretability and evaluability combining entity linking and labeled LDA. In *11th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2016, Krakow, Poland, July 11-16, 2016, Conference Abstracts*, pages 632–635. Alliance of Digital Humanities Organizations (ADHO).

Kristian Noullet, Rico Mix, and Michael Färber. 2020. KORE 50^{DYWC}: An Evaluation Data Set for Entity Linking Based on DBpedia, YAGO, Wikidata, and Crunchbase. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 2389–2395. European Language Resources Association.

Kristian Noullet, Samuel Printz, and Michael Färber. 2021. Clit: Combining linking techniques for everyone. In *The Semantic Web: ESWC 2021 Satellite Events - Virtual Event, June 6-10, 2021, Revised Selected Papers*, volume 12739 of *Lecture Notes in Computer Science*, pages 88–92. Springer.

Riccardo Orlando, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. Relik: Retrieve and link, fast and accurate entity linking and relation extraction on an academic budget. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 14114–14132. Association for Computational Linguistics.

Francesco Piccinno and Paolo Ferragina. 2014. From tagme to WAT: a new entity annotator. In *ERD'14, Proceedings of the First ACM International Workshop on Entity Recognition & Disambiguation, July 11, 2014, Gold Coast, Queensland, Australia*, pages 55–62. ACM.

Michael Röder, Ricardo Usbeck, Sebastian Hellmann, Daniel Gerber, and Andreas Both. 2014. N^3 - A collection of datasets for named entity recognition and disambiguation in the NLP interchange format. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 3529–3533. European Language Resources Association (ELRA).

Ahmad Sakor, Kuldeep Singh, Anery Patel, and Maria-Esther Vidal. 2020. Falcon 2.0: An entity and relation linking tool over wikidata. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 3141–3148. ACM.

Hassan Shavarani and Anoop Sarkar. 2023. Spel: Structured prediction for entity linking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11123–11137. Association for Computational Linguistics.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.*, 27(2):443–460.

TextRazor Ltd. 2024. TextRazor, Extract Meaning from your Text.

Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. 2020. Rel: An entity linker standing on the shoulders of giants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20. ACM.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

Lizheng Zu, Lin Lin, Song Fu, Jie Liu, Shiwei Suo, Wenhui He, Jinlei Wu, and Yancheng Lv. 2024. Pathel: A novel collective entity linking method based on relationship paths in heterogeneous information networks. *Inf. Syst.*, 126:102433.