

WikiUMLS: Aligning UMLS to Wikipedia via Cross-lingual Neural Ranking

Afshin Rahimi
School of ITEE
The University of Queensland
Brisbane, Australia
a.rahimi@uq.edu.au

Timothy Baldwin **Karin Verspoor**
School of Computing and Information Systems
The University of Melbourne
Melbourne, Australia
tbaldwin@unimelb.edu.au
karin.verspoor@unimelb.edu.au

Abstract

We present our work on aligning the Unified Medical Language System (UMLS) to Wikipedia, to facilitate manual alignment of the two resources. We propose a cross-lingual neural reranking model to match a UMLS concept with a Wikipedia page, which achieves a recall@1 of 72%, a substantial improvement of 20% over word- and char-level BM25, enabling manual alignment with minimal effort. We release our resources, including ranked Wikipedia pages for 700k UMLS concepts, and WikiUMLS, a dataset for training and evaluation of alignment models between UMLS and Wikipedia collected from Wikidata. This will provide easier access to Wikipedia for health professionals, patients, and NLP systems, including in multilingual settings.¹

1 Introduction

The Unified Medical Language System (UMLS)² is a controlled vocabulary resource, enabling standardisation of biomedical terminology and interoperability of electronic health systems across the world (Bodenreider, 2004). UMLS has good coverage in only a handful of languages such as English, impeding its uptake in health systems in different language settings (Markó et al., 2006). In addition, concept definitions in UMLS are either missing or very short, limiting its use as a medical encyclopedia. Wikipedia, on the other hand, is a crowd-sourced encyclopedia that is a primary source of online medical knowledge for practitioners, students, and the general public (Heilman et al., 2011; Heilman and West, 2015; Shafee et al., 2017; Murray, 2019). Wikipedia hosts an increasing number of health-related articles (Matheson-Monnet and Matheson, 2017). Wikipedia articles exist in 100s of languages, far exceeding the multilingual support in UMLS. Together, these characteristics point to the value of Wikipedia for patient-facing health information.

Our goal in this work is to align UMLS concepts to their corresponding Wikidata and through that to their available Wikipedia concepts, to expand the language support for UMLS terminology with little effort. This will have a direct impact on patients worldwide by facilitating adoption of UMLS (including clinical terminology SNOMED-CT (Donnelly, 2006) or Medical Subject Headings, MeSH (Aronson and Lang, 2010)) in international healthcare systems, and also supporting medical information seeking needs of patients with varying linguistic backgrounds. For example, a patient whose native language is not English might receive a discharge summary in English with mentions of symptoms, diagnoses, or medications in formal clinical language. This language may not match Wikipedia titles exactly, making search hard (and noisy). Aligning UMLS and Wikipedia facilitates such information seeking, makes medical information accessible in a patient’s native language, and supports the need for patient access to information in consumer-friendlier terms (Zeng et al., 2001; Smith and Stavri, 2005). Most common medical terms are covered in Wikipedia; for instance, it is estimated that 80% of SNOMED-CT terms have a corresponding Wikipedia page (Ngo et al., 2019).

Our contributions are as follows: (1) we use the multilingual resources in UMLS and Wikidata in a neural ranking model to retrieve Wikipedia articles given a UMLS concept, and achieve a recall@1

¹Resources available at <https://github.com/afshinrahimi/wikiuMLS>
This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>.

²<https://www.nlm.nih.gov/research/umls>

of 72%, representing a 20% increase over a BM25 model; (2) we release our new WikiUMLS dataset (collected from Wikidata) for training and evaluation of UMLS-to-Wikipedia alignment models; and (3) we release the output of our cross-lingual ranking model as a large-scale silver-standard alignment of UMLS and Wikipedia, facilitating new applications that can make use of the alignment.

2 Background

UMLS combines multiple vocabularies into a unified coding system by mapping terms referring to the same concept to a single concept ID (CUI). A single CUI can have several aliases in different languages from various vocabularies. A large proportion of UMLS concepts (roughly 100k) have descriptions that are very short (mostly one sentence), and so are not adequate for dissemination of knowledge or information access. Medical entity pages in Wikipedia have longer descriptions, are connected to other entities through hyperlinks, and are enriched by the Wikidata (Vrandečić and Krötzsch, 2014) knowledge graph. Wikipedia is rich in content and multilinguality, as anyone can contribute to or revise an article. For example, the `diabetes` Wikipedia page exists in more than 134 languages, compared to only a dozen languages in UMLS.

Knowledge-base alignment: There are two main approaches for aligning knowledge bases such as Wikipedia and UMLS: (1) embedding-based alignment, and (2) string and semantic matching. In embedding-based methods, entity embeddings are learnt from text co-occurrence statistics or knowledge-graph (KG) relations separately, and are then aligned using a seed alignment dictionary (Mikolov et al., 2013; Chen et al., 2017), or adversarial learning (Qu et al., 2019). However, the text-based methods suffer from non-comparable corpora (Ormazabal et al., 2019), and the KG methods have not been tried on Wikidata because of its large scale and entity variety.

String and semantic matching methods are based on similarity between the entity names or descriptions in the two knowledge-bases (KBs). Each entity in KB_1 is used as a query against all the entities in KB_2 using a scoring method such as cosine similarity or BM25. Entities are represented by bag of words or bag of character n -grams. The result of these cheap methods can be further improved by supervised (neural) ranking models (Guo et al., 2016; Rao et al., 2019; Wang et al., 2018). Particularly when query and candidate documents are represented by an encoder such as BERT (Devlin et al., 2019), pre-trained on massive amounts of text data, neural rerankers perform substantially better than IR-based methods (Nogueira and Cho, 2019; Akkalyoncu Yilmaz et al., 2019). Our methods fit within this group.

3 Method

Given a UMLS concept c_i represented by query $q_i = \{t_i^1, \dots, t_i^N\}$, where t_i^n is an alias term for c_i in UMLS, we use the English Wikipedia augmented with multilingual Wikidata as a document collection $D = \{d_1, \dots, d_{|D|}\}$, to retrieve page d_j matching concept c_i . Each page is represented by its title, text (only for candidate generation), and multilingual aliases from Wikidata. We follow a two-stage retrieval procedure: (1) candidate generation, where an IR method (e.g. BM25) is used to retrieve related documents; and (2) reranking of the top k candidates via a learn-to-rank method (Liu, 2009).

3.1 Candidate Generation

We index Wikipedia collection D using `Lucene`, and build query q_i from UMLS to retrieve the top $k=64$ relevant pages. We use a Boolean disjunction between all alias terms in UMLS, and search in the title, text, and multilingual aliases fields in D . BM25 relies on exact term matches, and small variations can result in a mismatch. As a result, we also experimented with a character n -gram method (`TFIDFchar`) successfully used in Murty et al. (2018) for candidate generation in medical entity linking. We build a bag of character n -grams ($n \in [1, 5]$) weighted by `TF-IDF` within term boundaries, and use cosine similarity between q_i and each $d \in D$ (excluding page text) to generate the top $k=64$ candidates.

3.2 Reranking

We formulate the reranking task as passage pair binary classification (Nogueira and Cho, 2019), where the first passage is q_i for concept c_i from UMLS, and the second passage is the set of Wikipedia alias

UMLS Concept	UMLS Query	Gold Wiki Candidate
CUI: C0017168	GERD, Acid Reflux, oesofagusaan-doening, ...	disorder of the esophagus, enfermedad esofagal, ...

Table 1: The multilingual UMLS aliases are used as a query against all Wikipedia entities to generate k candidates. The candidates are then reranked using a neural model.

names for each of top $k=64$ documents ranked by candidate generation (Wikipedia text is excluded for reranking). The goal is to predict if a pair is a match or not, by minimising the following objective:

$$L(q_i, \text{cand}_i) = -\log(f(q_i, d^+)) - \sum_{d^- \in \text{cand}_i^-} \log(1 - f(q_i, d^-))$$

where (q_i, d^+) is the matching UMLS-Wiki pair, and cand_i^- is the set of remaining negative candidates generated by BM25. Function f is the passage pair encoder, for which we use BERT’s <CLS> token encoding (Devlin et al., 2019) and a linear projection. We also experiment with BiOBERT (Lee et al., 2019) because it is pre-trained on medical literature, which is a better domain fit for our task.

The major shortcoming of BERT and BiOBERT is that they don’t encode multilingual aliases effectively, particularly if the scripts are different. Aliases of every concept in UMLS and Wikipedia are available on average in 1.4 and 11.2 different languages, respectively. This multilingual data can be utilised for triangulation of the matching pair. For example, *GERD* is a term used in UMLS for Gastroesophageal reflux disease, and is also a given name in Germanic languages. If it is used for retrieval, many Wikipedia pages related to people with that name will rank high. However, if alias names of *GERD* in other languages (e.g. Japanese) are used in the query and documents, the disambiguation becomes easier. To encode multilingual alias concept names we require a model that embeds tokens of different languages into the same embedding space. To this end, we also experiment with multilingual BERT (BERT_{multi}).

3.3 Data

WikiUMLS is a UMLS to Wikipedia aligned dataset we create in this work. It consists of about 17.8k UMLS concepts that are manually linked to their matching Wikipedia page by Wikipedia content contributors. All the UMLS concepts in WikiUMLS are from Medical Subject Headings vocabulary (MeSH) (Lipscomb, 2000). A WikiUMLS record is a tuple of (UMLS CUI, UMLS concept alias set, Wikipedia page title, Wikidata concept alias set). We use a UMLS concept’s alias set as query, and compare it with every Wikidata entity’s alias set to retrieve the matching Wikipedia page. The Wikipedia aliases and their links to UMLS are taken from their corresponding record in Wikidata, a collaborative knowledge-base that is tightly connected to Wikipedia. An entity in Wikidata has multilingual aliases,³ and is linked to both a Wikipedia page, and possibly a UMLS concept.⁴ We split the matching pairs into roughly 10k, 2k, and 5.8k for the training, validation, and test data sets, respectively. There are about 3 million remaining UMLS concepts from various vocabularies that are not aligned to Wikipedia, which we hope to align in this work.

3.4 Evaluation Methodology

We use the aliases of a concept in UMLS as a query, and the aliases for entities in Wikipedia as the document collection. A document is relevant to a query if the pair are manually aligned in WikiUMLS (a matching UMLS-Wikipedia pair). We evaluate the models by measuring recall at different positions of their ranking (recall@ k). In this paper, recall and accuracy are equivalent because we assume each UMLS concept has a unique matching Wikipedia page. Because our proposed retrieval method has two stages (candidate generation followed by neural reranking), we use normalised recall (N. recall@ k) to evaluate the neural reranking stage independent of candidate generation stage by ignoring queries for which the candidate generator didn’t retrieve a relevant document.

³For the purposes of this work, we focus on the 24 languages currently supported in UMLS: en, es, it, nl, fr, pt, de, cs, ru, zh, ja, hu, tr, ko, no, et, sv, pl, fi, el, lv, da, eu, he.

⁴A sample Wikidata record: <https://wikidata.org/wiki/Q12206>.

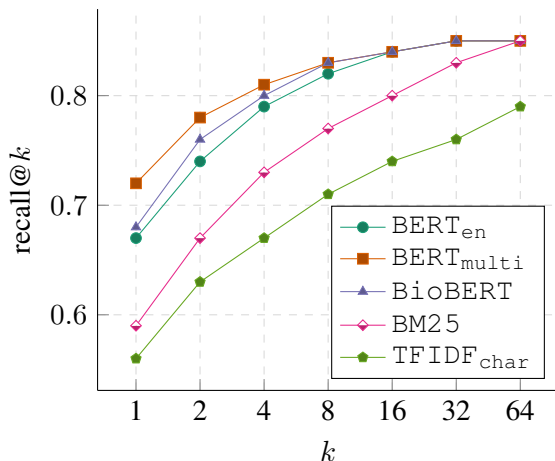


Figure 1: recall@ k for neural reranking models (BERT_{multi}, BERT_{en}, BioBERT), and word- (BM25) and character-level (TFIDF_{char}) models.

Method	Recall			N. Recall	
	@64	@4	@1	@4	@1
BM25	85	73	59	—	—
TFIDF _{char}	79	67	56	—	—
BERT _{en}	85	79	67	92	78
BioBERT	85	80	68	94	79
BERT _{multi}	85	81	72	95	84

Table 2: Performance of the vector space models (BM25, TFIDF_{char}), and neural ranking models (BERT_{en}, BioBERT, BERT_{multi}) over Wikipedia page retrieval given a UMLS concept. Normalised recall (N. recall@ k) is calculated by excluding queries for which the candidate generator (BM25) didn’t retrieve the relevant document.

4 Results

The performance of the candidate generation methods for word- (BM25) and character-level (TFIDF_{char}) retrieval, and also the reranking methods (BERT, BioBERT, and BERT_{multi}) are shown in Figure 1. In all top k positions, BM25 outperforms TFIDF_{char}, which is surprising given that TFIDF_{char} was reported to achieve strong performance for entity linking in Murty et al. (2018).⁵ As shown in Table 1, BM25 is able to retrieve the correct Wikipedia page at $k=64$ with 85% accuracy (compared to 75% for TFIDF_{char}), an upper-bound for the performance of the reranking models.

For the reranking models, BioBERT performs only slightly better than BERT (68% vs. 67% recall@1), although it has been pretrained on large amounts of biomedical literature. BERT_{multi} performs better than both BERT and BioBERT, achieving a recall@1 of 72%, given the cross-lingual nature of the task. Compared to classic vector space approaches (e.g. BM25), BERT_{multi} shows an improvement of 22%, and 10% for recall@1 and recall@10, respectively. We also report normalised recall at $k=1$ and $k=4$, by excluding the test instances for which BM25 doesn’t retrieve the gold candidate. Here, BERT_{multi} achieves normalised recall at $k=1$ and $k=4$ of 84% and 95%, respectively. This indicates that BERT_{multi} is highly successful at ranking in the case that the correct document is retrieved.

5 Conclusions

We proposed passage pair ranking models based on pretrained contextual encodings for aligning UMLS and Wikipedia, to help bridge between health information systems, and empower consumers with understanding of their health condition. We developed a dataset, WikiUMLS, for training and testing alignment models between the two knowledge-bases, and proposed neural reranking models that substantially outperform BM25. We showed that the use of multilingual aliases in BERT_{multi} substantially improves recall@1 compared to BioBERT (72 vs. 68).

The use of subword information such as BPE (Sennrich et al., 2016) as used in XLM (Conneau and Lample, 2019) might improve performance, which we leave for future work. Utilising the relationships between concepts in UMLS and Wikipedia (through Wikidata) to align the two knowledge graphs is also an interesting future direction. We also intend to release a large Wikipedia-based Entity Linking (EL) dataset by using the top-ranked Wikipedia pages for UMLS queries, to be used in state-of-the-art EL models such as zeshel (Logeswaran et al., 2019).

⁵The recall@100 of TFIDF_{char} for entity linking candidate generation in Murty et al. (2018) is 61.09%, much higher than the results for our re-implementation over the same dataset, possibly due to differences in implementation or the candidate set.

Acknowledgements

This work was funded by the Australian Research Council through the ARC Training Centre in Cognitive Computing for Medical Technologies (project number ICI70200030), and completed while the first author was a post-doctoral researcher at The University of Melbourne.

References

- Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3488–3494.
- Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.
- Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270.
- Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI'17*, pages 1511–1517.
- Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pages 4171–4186.
- Kevin Donnelly. 2006. SNOMED-CT: The advanced terminology and coding system for ehealth. *Studies in Health Technology and Informatics*, 121:279.
- Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016*, pages 55–64.
- James M Heilman and Andrew G West. 2015. Wikipedia and medicine: Quantifying readership, editors, and the significance of natural language. *J Med Internet Res*, 17(3):e62.
- James M Heilman, Eckhard Kemmann, Michael Bonert, Anwesh Chatterjee, Brent Ragar, Graham M Beards, David J Iberri, Matthew Harvey, Brendan Thomas, Wouter Stomp, Michael F Martone, Daniel J Lodge, Andrea Vondracek, Jacob F de Wolff, Casimir Liber, Samir C Grover, Tim J Vickers, Bertalan Meskó, and Michaël R Laurent. 2011. Wikipedia: a key tool for global public health promotion. *Journal of Medical Internet Research*, 13(1):e14.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 09.
- C. E. Lipscomb. 2000. Medical Subject Headings (MeSH). *Bull Med Libr Assoc*, 88(3):265–266.
- Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331.
- Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460.
- Kornél Markó, Robert Baud, Pierre Zweigenbaum, Lars Borin, Magnus Merkel, and Stefan Schulz. 2006. Towards a multilingual medical lexicon. In *AMIA Annual Symposium Proceedings*, volume 2006, pages 534–538.
- Catherine Matheson-Monnet and David J. Matheson. 2017. Wikipedia as informal self-education for clinical decision-making in medical practice. *Open Medicine Journal*, 4(1):15–25.

- Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.
- Heather Murray. 2019. More than 2 billion pairs of eyeballs: Why aren't you sharing medical knowledge on Wikipedia? *BMJ Evidence-Based Medicine*, 24(3):90–91.
- Shikhar Murty, Patrick Verga, Luke Vilnis, Irena Radovanovic, and Andrew McCallum. 2018. Hierarchical losses and new resources for fine-grained entity typing and linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 97–109.
- Duy Hoa Ngo, Donna Truran, Madonna Kemp, Michael Lawley, and Alejandro Metke-Jimenez. 2019. Can Wikipedia be used to derive an open clinical terminology? In *Digital Health: Changing the Way Healthcare is Conceptualised and Delivered: Selected Papers from the 27th Australian National Health Informatics Conference (HIC 2019)*, volume 266, page 136.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.
- Aitor Ormazabal, Mikel Artetxe, Gorka Labaka, Aitor Soroa, and Eneko Agirre. 2019. Analyzing the limitations of cross-lingual word embedding mappings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4990–4995.
- Meng Qu, Jian Tang, and Yoshua Bengio. 2019. Weakly-supervised knowledge graph alignment with adversarial learning. *CoRR*, abs/1907.03179.
- Jinfeng Rao, Linqing Liu, Yi Tay, Wei Yang, Peng Shi, and Jimmy Lin. 2019. Bridging the gap between relevance matching and semantic matching for short text similarity modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5369–5380.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.
- Thomas Shafee, Gwinyai Masukume, Lisa Kipersztok, Diptanshu Das, Mikael Häggström, and James Heilman. 2017. Evolution of Wikipedia's medical content: past, present and future. *Journal of Epidemiology & Community Health*, 71(11):1122–1129.
- Catherine Arnott Smith and P Zoë Stavri. 2005. Consumer health vocabulary. In *Consumer Health Informatics*, pages 122–128. Springer.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.
- Lucy Lu Wang, Chandra Bhagavatula, Mark Neumann, Kyle Lo, Chris Wilhelm, and Waleed Ammar. 2018. Ontology alignment in the biomedical domain using entity definitions and context. In *Proceedings of the BioNLP 2018 Workshop*, pages 47–55.
- Qing Zeng, Sandra Kogan, Nachman Ash, and Robert A. Greenes. 2001. Patient and clinician vocabulary: How different are they? In *MEDINFO 2001 - Proceedings of the 10th World Congress on Medical Informatics*, pages 399–403.