SciConceptMiner: A system for large-scale scientific concept discovery

Zhihong Shen **Chieh-Han Wu** **Li Ma** **Chien-Pang Chen** **Kuansan Wang**
Microsoft Research
Redmond, WA, USA
{zhihosh,chiewu,v-lima3,v-chienc,kuansanw}@microsoft.com

Abstract

Scientific knowledge is evolving at an unprecedented rate of speed, with new concepts constantly being introduced from millions of academic articles published every month. In this paper, we introduce a self-supervised end-to-end system, ***SciConceptMiner***, for the automatic capture of emerging scientific concepts from both independent knowledge sources (semi-structured data) and academic publications (unstructured documents). First, we adopt a BERT-based sequence labeling model to predict candidate concept phrases with self-supervision data. Then, we incorporate rich Web content for synonym detection and concept selection via a web search API. This two-stage approach achieves highly accurate (94.7%) concept identification with more than 740K scientific concepts. These concepts are deployed in the *Microsoft Academic*[1] production system and are the backbone for its semantic search capability.

1 Introduction

Scientific knowledge has been expanded at an exponential rate over the past decades and the fast-growing volume of academic literature accentuates a pressing need for automated capture of fine-grained emerging concepts. Statistical topic models (Blei, 2012), such as *latent Dirichlet allocation* (LDA) (Blei et al., 2003), have been well-recognized for automatically extracting the topic structure of large document collections for past decades. However, it has two main limitations to prevent it from being widely applied in a modern large-scale document collection.

First, it is the scalability issue on the number of topics an LDA can model. The latest development (Chen et al., 2018) can process 131M documents with 28B tokens efficiently, however, it only extracts 1,722 topics. With the fast-growing body of scholarly communications, a comprehensive manually controlled vocabulary like *Medical Subject Headings*(MeSH) (Lowe and Barnett, 1994) contains tens of thousands of subjects (concepts) mostly in the bio-med domain; and an automated scientific knowledge exploration system such as *Microsoft Academic Graph* (MAG) (Shen et al., 2018) has hundreds of thousands of topics across all academic disciplines. A topic modeling system that is scalable not only to the size of documents but also to the number of topics is imperative.

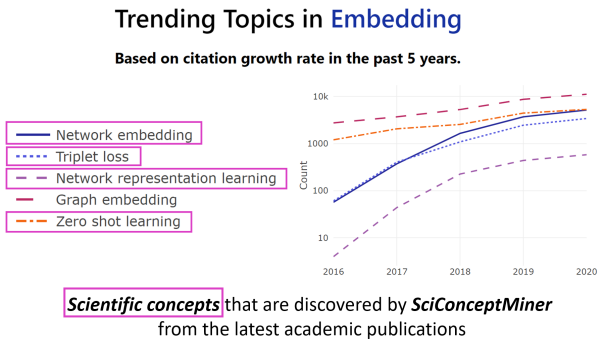

Figure 1: Trending Topics under concept *Embedding*.

Second, the result of an LDA model is a list of frequency-based terms that form a topic. It requires manual efforts to annotate such lists to generate a human-readable theme or topic name. An automatic process of identifying topic themes with authoritative names and meaningful descriptions is desired to reduce costly human interventions.

In this paper, we introduce a self-supervised end-to-end system, ***SciConceptMiner***, for automatically discovering scientific concepts from both semi-structured independent knowledge sources and unstructured academic documents. It first obtains a list of concept candidates, either from external knowledge repositories such as *Wikipedia* (Völkel et al., 2006; Vrandečić and Krötzsch, 2014) and *Unified Medical Language System* (UMLS) (Bodenreider, 2004), or directly

[1] https://academic.microsoft.com/

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 48–54, August 1st - August 6th, 2021.

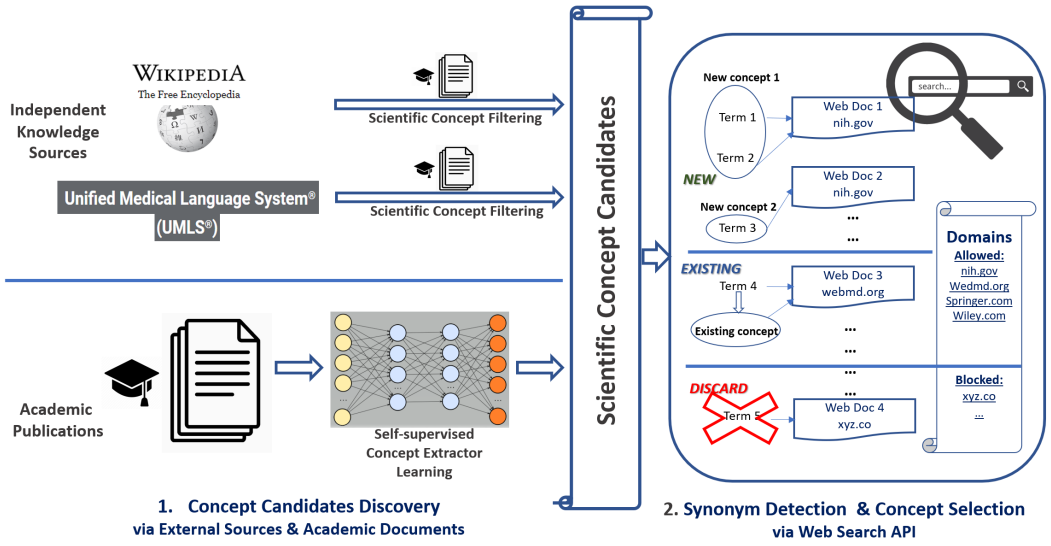

Figure 2: An overview of the ***SciConceptMiner*** system.

mining concepts from a collection of academic documents. Such concept lists are large and noisy. They are in the scale of millions and dominated by invalid or duplicate terms. We then send these candidates as queries to a search engine API and leverage rich Web content to identify legitimate concepts, cluster synonyms, and discard improper terms. The search API is also used to retrieve high-quality concept descriptions.

One example is shown in Figure 1.[2] Four out of five trending topics (*network embedding*, *triplet loss*, *network representation learning*, and *zero shot learning*) under *embedding* are extracted by our automatic concept extractor model trained on CS corpus. It demonstrates that our designed model can effectively capture the emerging trending topics from the latest scientific articles.

The **SciConceptMiner** has been deployed to identify concepts from millions of scholarly communications in *Microsoft Academic Graph* (MAG) (Sinha et al., 2015; Wang et al., 2019, 2020). The MAG with the full list of 740K scientific concepts can be freely accessed via the *Microsoft Academic*[3] search website and *MAG* data set[4].

2 System Description

As shown in Figure 2, the **SciConceptMiner** system has two stages: the first is the concept candidates discovery from various data sources; the second is synonym detection and concept clustering via a Web search API.

In the concept candidates discovery stage, we first integrate the semi-structured independent knowledge sources, *Wikipedia* and *UMLS*, into the system. Such an existing concept list in the system with associated documents enables us to train a concept extractor learning model with self-supervision. We design a BERT-based sequence labeling model to make a binary prediction on whether a word or phrase in a sentence is a scientific concept or not. This proposed model is trained on self-supervised data generated from existing concepts (from *Wikipedia* and *UMLS*) tagged to a collection of academic documents. We do the concept inference with the trained model to generate concept candidates for the next stage.

Concept candidates, as the input to the second stage, are either from external knowledge sources or inferred from academic documents. Both sources have high noisy ratios with different natures. The independent source such as *Wikipedia* has high-quality entities (well-defined names and descriptions, rare duplication, and rich links and relationships with each other) but type noisy (many other types of entities than academic concepts). The *UMLS* candidates and the inferred candidates from an unstructured corpus have more irrelevant phrases and concept synonyms. With the help of a search engine API to retrieve top N documents by using concept candidates as queries, we analyze the returning web pages and associated URL domain information collectively. This process would iden-

[2] This is a snapshot captured in March 2021 for *Embedding* concept at *Microsoft Academic* production system: https://academic.microsoft.com/topics/41608201.

[3] https://academic.microsoft.com/

[4] https://docs.microsoft.com/en-us/academic-services/graph/

tify around 3-5% of candidates from the first stage as proper scientific concepts with consistently high accuracy (94-95% based on sample results) across all data sources, with over 740K concepts in total.

2.1 Concept Candidate Discovery

2.1.1 Semi-structured Independent Knowledge Sources

There are many independent knowledge sources, either manually curated or automatically created or a hybrid of both. Among them, the most notable ones are Wikipedia, WikiData[5], DBpedia[6], and Yago[7] in general domains and MeSH[8], UMLS[9] in the bio-med fields. We have applied Wikipedia and UMLS as sources for *SciConceptMiner* system because of their data quality and comprehensive coverage on scientific terms and phrases. Other semi-structured sources can be integrated with the current system design seamlessly as long as they pass the quality and relevancy examination of their contents.

Wikipedia: Wikipedia[10] is the largest collaboratively edited online encyclopedic knowledge. It contains contents in more than 300 languages and has over 6 million English articles as of July 2020. It was the first external data source being integrated into MAG considering its comprehensive coverage on academic topics spanning from social sciences to natural sciences, as well as technology and applied sciences. Each topic in Wikipedia (as a separate article) is written in high quality and has rare duplication (Lewoniewski, 2018). The key challenge of mining quality academic concepts from Wikipedia is to identify the right type of entities, as most articles in Wikipedia are missing entity type information. We used graph link analysis (Milne and Witten, 2008) for type prediction and had expanded the concepts from an initial 3K to over 200K. The details are described in the *Concept Discovery* section in (Shen et al., 2018). For concepts from Wikipedia, we did not use the search engine API to further filter as the resulting concept list is already with high quality and rare duplication.

UMLS:

The Unified Medical Language System (UMLS) is a repository of biomedical vocabularies developed by the US National Library of Medicine (NLM) with sources from multiple datasets and standards. The latest 2020AA release contains approximately 4.28 million medical concepts and 15.5 million unique concept names from over 200 sources. A system with large, complex data sources typically has various inherent limitations on the data quality. For UMLS, these include structural inconsistencies such as cycles in graph hierarchy, semantic inconsistencies between different vocabularies, and missing hierarchical relationships (Bodenreider, 2004, 2007; Humphreys et al., 1998).

In the concept candidate discovery stage, we take the full list of the concept names from UMLS and first clean it with simple rules such as removing digit-only terms, two-char terms, too long terms (over 30 chars), etc. We further filter the remaining terms with a corpus consisting of titles and abstracts from 170 million English scientific articles in MAG and only keep terms that appeared at least N times in above academic corpus. The resulting list is ready to be sent to a search engine API for duplication detection and concept selection in the second stage.

2.1.2 Self-supervised Concept Extractor Learning

The volume of new research being published is rapidly increasing, with MAG adding over 1 million new papers every month. This creates a unique challenge to identify, describe, and categorize an ever-evolving set of emerging concepts in a timely fashion.

To tackle this challenge, we formulate the concept detection as a self-supervised sequence labeling problem that allows us to extract concept candidates directly from unstructured academic documents. This is motivated by the recent development of deep learning (DL) based *Named Entity Recognition (NER)* models, which become dominant and achieve state-of-the-art results (Lample et al., 2016; Chiu and Nichols, 2016; Yadav and Bethard, 2019). NER is the task of identifying named entities of a specific type, such as person or location, in text. A most recent survey (Li et al., 2020) proposed a new taxonomy of DL-based NER with three parts: *distributed representations for input*, *context encoder*, and *tag decoder*. We adopt this taxonomy to design our concept extractor learning model.

Instead of a typical NER model which would learn to identify several entity types at the same

[5] https://www.wikidata.org/
[6] https://wiki.dbpedia.org/
[7] https://yago-knowledge.org/
[8] https://www.nlm.nih.gov/mesh/meshhome.html
[9] https://www.nlm.nih.gov/research/umls/index.html
[10] https://www.wikipedia.org/

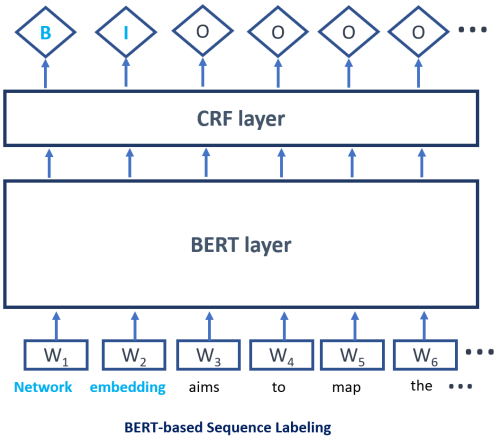

Figure 3: Concept extractor learning with a BERT-based sequence labeling model.

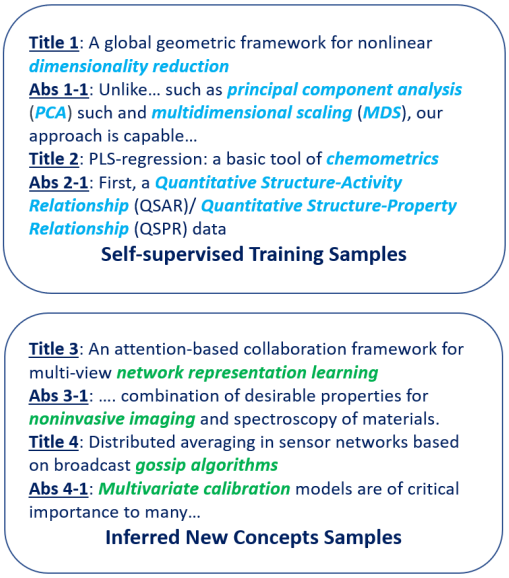

Figure 4: Self-supervised concept labeling samples.

time, we reduce our model design to identify a single entity type - scientific concept type. We propose to treat scientific concept extraction as a sequence labeling task. Tokens in the text are labeled with the BIO notation. 'B', 'I', and 'O' represent the beginning, inside, and outside of a scientific concept chunk respectively. On a sampled set of scientific articles in MAG, we do lexical matching using the synonyms of our existing concepts harvested from *Wikipedia* and *UMLS* as self-supervised labels. We fine-tune a transformer-based BERT model (Devlin et al., 2018) (e.g. BERT-Large) as a *context encoder* and use a Conditional Random Field (CRF) layer as a *tag decoder* to train a binary classifier on each word in a sentence to detect concept mentions.[11] Figure 3 illustrates the design of our concept extractor learning model. We infer scientific concept candidates using the trained model on a larger set of high-quality MAG documents, i.e. those published in prestigious journals/conferences. Figure 4 provides some self-supervised concept labeling samples as well as sample sentences with inferred new concepts. These new concept candidates are ready to be used in the next stage.

2.2 Synonym Detection and Concept Selection

In the second stage, we classify the scientific concept candidates detected in the first stage (either from UMLS or from automatic concept extractor models) into three broad categories: (1) synonyms of existing concepts, (2) new concepts, or (3) low-quality words/phrases we shall discard.

This is accomplished by searching for each concept candidate using the Bing Web Search API[12] and clustering candidates into scientific concept "identities" based on the URL relevance/reputation and the consistency of the mentions among top search results.

More specifically, if K out of top N URLs returned by two concept candidates is the same, we consider these two candidates are synonyms of a concept. We also curate the allowed-list and block-list of URL domains. The concept candidates whose top search results are from well-known domains of high-quality academic knowledge (in the allowed-list) would be accepted, and otherwise, they would be rejected. The block-list is used to reject terms that also have results from domains in the allowed list. That is usually the case for common words and phrases which returned with pages in online dictionary domains.

This simple yet effective approach can help trim around 92%-97% concept candidates as noisy terms and keep 3%-7% of high-quality concepts, synonyms, and well-written descriptions from domains containing credible academic knowledge and are in the allowed-list.

3 Evaluation and Analysis

3.1 Self-supervised concept extractor learning

We use the BERT-Large-Cased as the pre-trained language model and fine-tune the described con-

[11] We re-use the BERT vocabularies and their pre-trained embedding without regenerating and retraining on academic corpus.

[12] https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/

cept extractor learner model with 4 epochs. We generate the training corpus from MAG from CS and Medicine domain respectively and split them in 8:1:1 for train/dev/test. Table 1 shows the corpus size used for training and inference.

Training Corpus	CS	Med
# of articles	500K	414K
# of sentences	3.4M	3.6M
# of tokens	72.8M	82.7M
# of concept tokens	8.9M	9.7M
Inference Corpus	**CS**	**Med**
# of articles	2.56M	2.07M
# of sentences	17.6M	18.1M
# of tokens	373.8M	413.4M
# of concept tokens	26.2M	91.2M
Inferred Concept Terms	**CS**	**Med**
# of distinct terms	1.06 M	4.66M
# of cur. concept terms	73,167	88,350
# of new concept terms	48,531	34,744
# of new distinct concepts	46,182	31,302
# of new terms for cur. concepts	16,021	11,389
# of discarded terms	921k	4.53K

Table 1: Training and Inference Corpus Stats.

To ensure that this model works for documents across various scientific domains, we conduct experiments training our model using documents in different top domains (e.g. *computer science* and *medicine*). We observe that higher-quality candidates are generated using models trained from the same domain corpus. For example, when we apply the model trained with a CS corpus to predict concepts in the medicine corpus, the F1 score drops from 0.942 to 0.682. Therefore, we train different models on the corpus from an individual top-level domain, and the F1 scores of inference results on in-domain and out-of-domain corpus are shown in Table 2.

	CS-Model	Medicine-Model
CS-Test	**0.942**	0.649
Medicine-Test	0.682	**0.912**

Table 2: F1 scores of test sets on different models.

We have only conducted model training and inference on CS and medicine corpus. Continued training on other discipline corpora as well as exploring more effective concept extractor learning models are among our ongoing efforts.

3.2 Concept Analysis Based on Data Sources

In this section, we conduct an evaluation of the concept quality in terms of accuracy and coverage. We estimate the coverage by evaluating potential missed opportunities on discarded terms. We also leverage MAG data to conduct the analysis of top domain distribution and topic age distribution conditioned on different data sources.

The stats in this section are collected on four groups of concepts by their data sources: *Wikipedia*, *UMLS*, automatically extracted concepts on *Computer Science* (*AutoCS* or *A-CS*) and *Medicine* (*AutoMed* or *A-Med*) corpus respectively. Since the concepts discovered in *SciConceptMiner* are already integrated into MAG, we use the paper-concept relationship, concept hierarchy, and paper metadata such as publication year in MAG to facilitate this analysis. The details on how to obtain these relationships and meta-data are out of the scope of this work and please refer to (Wang et al., 2019; Shen et al., 2018) for more information.

3.2.1 Size, Impact, and Accuracy

In Table 3, we report the number of concepts, average number of papers associated with a concept, average citation received of a paper tagged with a concept, as well as the accuracy of concepts. The independent knowledge sources (*Wikipedia* and *UMLS*) provide similar topic sizes on a scale of hundreds of thousands, while the automatic extraction models identify about one-tenth of the size from external sources. On average, the concepts from *Wikipedia* are broader (with more papers associated) and have a higher impact (with more citations received), while concepts from *UMLS* are more fine-grained with slightly smaller influence. We evaluate the accuracy with the same approach described in (Shen et al., 2018) and it achieves a similar accuracy level between 94% and 95% across all data sources.

Data Source	Size	Paper	Cit.	Acc.
Wiki	226,466	3,386	15.6	94.8%
UMLS	433,468	59	9.1	94.5%
AutoCS	46,182	1,462	10.1	94.8%
AutoMed	31,302	1,498	10.7	94.2%

Table 3: Concept size, impact, and accuracy.

3.2.2 Potential Opportunities on Discarded Contents

It is generally challenging to evaluate the coverage of such a large-scale concept discovery system since it is nearly impossible to identify the "ground truth" of full coverage, even in a narrowed subdomain. In order to estimate the coverage, we identify the potential opportunities that we may have missed by sampling and inspecting the discarded inferred terms from learned concept extractor models. We sample 300 discarded terms in *AutoCS* and *AutoMed* respectively and report the size and accu-

racy[13] in Table 4. In all terms with a positive label, roughly one quarter to one third are new concepts not in the current system, and the remaining 66% to 75% are synonyms. Hence, we estimate that we might have missed about 100K concepts and 200K synonyms from the inference results of our concept extractor models.

Source	Discarded Size	Accuracy
Auto-CS terms	4.53 M	3.3%
Auto-Med terms	921 K	12.7%

Table 4: Discarded term size and accuracy.

3.2.3 Topic Domain Distribution

About 75% of 740K concepts in MAG are organized into a six-level DAG (directed acyclic graph) structure taxonomy, with top two levels manually curated (19 domains and 270 sub-domains). We use this taxonomy to aggregate all concepts to top-level 19 domains and report the percentage distribution on top 5 domains per data source and for all concepts. As shown in Table 5, Bio-Med-Chem 3 domains dominate all concepts (67%),*Wikipedia* (51%), *UMLS* (90%), and auto-extracted *AutoMed* (73%). Technology and applied sciences such as *Computer Science* and *Material Science* are the second biggest categories for all concepts. These two applied sciences together with *Mathematics* and *Engineering* dominate the *AutoCS* data source (58%).

	ALL	Wiki	UMLS	AutoCS	AutoMed
Bio	**28.4%**	**28.3%**	35.4%	-	**41.3%**
Med	24.2%	11.0%	**35.9%**	7.5%	16.2%
Chem	14.7%	11.6%	18.6%	-	15.3%
ComSci	7.0%	9.3%	-	**25.8%**	4.9%
MatSci	5.1%	-	2.6%	13.8%	7.8%
Math	-	6.0%	-	8.5%	-
Engr	-	-	-	9.3%	-
Other	20.7%	33.9%	7.5%	35.0%	14.5%

Table 5: Top domain distribution of concepts.

3.2.4 Topic Age Distribution

In Table 6, we report the average age of the papers associated with a concept. The average publication year (rounded off to the floor), as well as 5%, 50% (the median), and 95% publication year of a concept are also reported. It shows that concepts from *UMLS* are generally discovered and used in earlier years, lasting longer (25 years for the middle 90%), while *AutoCS* and *AutoMed* contain newer concepts with shorter life span (17-18 years for the middle 90%).

[13] We split the sampled data of each category to 3 groups with 100 each and they are evaluated by 3 judges. We report the average of positive label ratios.

Source	Age	Avg Y	5% Y	50% Y	95% Y
Wiki	18.2	2002	1983	2003	2013
UMLS	21.0	1999	1982	1997	2007
A-CS	14.1	2006	1990	2008	2017
A-Med	15.7	2004	1989	2006	2016

Table 6: Age distribution of concepts.

Figure 5 provides a yearly distribution from 2010 to 2019. It represents the percentage of papers (associated with concepts in respective sources) over the past 10 years.[14] This is consistent with our expectation as one of our primary goals of leveraging the automatic concept extraction is to discover emerging concepts in the latest scientific documents.

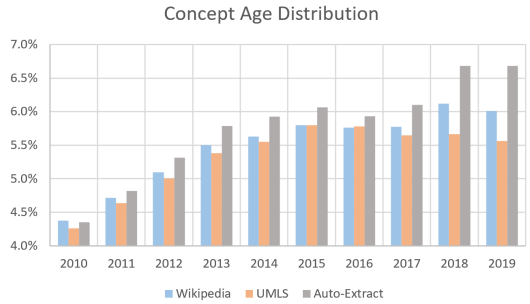

Figure 5: Concept Age Distribution 2010-2019.

4 Conclusion

In this work, we demonstrated a large-scale scientific concept discovery production system, *SciConceptMiner*, for automatically capturing academic concepts from both semi-structured data and unstructured documents. The system has two parts: the first is the concept candidate identification, and the second is synonym detection and concept selection. We used a BERT-based sequence labeling model to learn concept phrases with self-supervision and leverage a Web search API to cluster synonyms and identify valid concepts.

SciConceptMiner has discovered more than 740K scientific concepts across all research domains from *Wikipedia*, *UMLS*, and scholarly articles with high accuracy (94.7%). These concepts are integrated to build the *Microsoft Academic Graph*, which publishes one of the largest cross-domain scientific taxonomy. It enables easy exploration of scientific knowledge as well as facilitates many downstream applications like information retrieval, question answering, and recommendations.

[14] Please note that the percentage of papers of each year is calculated by dividing by all papers for a source. Since the earlier years' distributions are very close, we do not plot them. The sum of each source over the past 10 years is less than 1.

References

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Olivier Bodenreider. 2007. The unified medical language system what is it and how to use it? *Tutorial at Medinfo*.

Jianfei Chen, Jun Zhu, Jie Lu, and Shixia Liu. 2018. Scalable training of hierarchical topic models. *Proceedings of the VLDB Endowment*, 11(7):826–839.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Betsy L Humphreys, Donald AB Lindberg, Harold M Schoolman, and G Octo Barnett. 1998. The unified medical language system: an informatics research collaboration. *Journal of the American Medical Informatics Association*, 5(1):1–11.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Włodzimierz Lewoniewski. 2018. Measures for quality assessment of articles and infoboxes in multilingual wikipedia. In *International Conference on Business Information Systems*, pages 619–633. Springer.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.

Henry J Lowe and G Octo Barnett. 1994. Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *Jama*, 271(14):1103–1108.

David Milne and Ian H Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518.

Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A web-scale system for scientific knowledge exploration. *arXiv preprint arXiv:1805.12216*.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.

Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller, and Rudi Studer. 2006. Semantic wikipedia. In *Proceedings of the 15th international conference on World Wide Web*, pages 585–594.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Kuansan Wang, Zhihong Shen, Chi-Yuan Huang, Chieh-Han Wu, Darrin Eide, Yuxiao Dong, Junjie Qian, Anshul Kanakia, Alvin Chen, and Richard Rogahn. 2019. A review of microsoft academic services for science of science studies. *Frontiers in Big Data*, 2:45.

Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, 1(1):396–413.

Vikas Yadav and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.