# ADAPTING MEDLINKER TO MEDLINKER-SOCIAL

**Daniel Loureiro**
LIAAD-INESCTEC
dloureiro@fc.up.pt

## 1 INTRODUCTION

This document describes the changes introduced to MedLinker (Loureiro & Jorge, 2020) in order to optimize it for Medical Entity Linking on the Social Domain. Considering RECAP's focus on preterms, we also include some specialization for preterm-related concepts.

The most divergent feature of MedLinker-Social, with regards to MedLinker, is the withdrawal of the Contextual Neural Language Modeling (NLM) components in favour of improvements on the Approximate Dictionary Matching (ADM) components.

This departure was motivated by the following factors:

1. Unavailable pretrained Contextual NLMs for the social domain (during our development), similar to those employed in MedLinker for Mention Detection and Entity Linking.

2. In MedLinker, we found that ADM can be competitive with solutions based on Contextual NLMs when annotations are very limited, and there's no substantial corpus with medical concept annotations in the social domain.

3. Our early attempts employing Contextual NLMs trained on the general domain (i.e. BERT, Devlin et al. 2019), or medical literature (i.e. SciBERT, Beltagy et al. 2019), to the social domain demonstrated no gain over using the ADM components exclusively.

4. Domain adaptation solutions (i.e. VecMap, Artetxe et al. 2018) produced inconsistent results, seemingly due to reduced coverage of our annotation dataset (MedMentions, Mohan & Li 2019), coupled with the wide variety of concepts in our target ontology (UMLS).

5. Solutions involving Contextual NLMs expect use of high-end GPUs, which limit applications and seem hard to justify given our early experiments showing no obvious improvement for our use case.

In the following sections we describe the modifications we've found to be more adequate for Mention Detection in the social domain, as well as how we improved the ADM Entity Linking described in MedLinker (see Section 4 of Loureiro & Jorge 2020) to better accommodate the differences between writing in medical literature and in social networks (see Table 1).

Furthermore, we report on experiments performed on a sample of submissions from the Reddit social network. The analyses we've produced for this project are designed to assist in understanding the context in which various preterm-related concepts are discussed in social media, including in comparison to their discussion in medical literature.

As was the case with MedLinker, we're still interested in performing social linking based on the UMLS ontology, so that our solution can be easily integrated with a variety of biomedical Natural Language Processing projects (e.g. MedType, Vashishth et al. 2020).

For more information on this project, including code, tutorials and concept-specific reports from our analyses, please check the website: http://danlou.github.io/medlinker-social/

# 2 Method

## 2.1 Mention Detection as Keyword Extraction

With MedLinker-Social we rely on unsupervised Keyword Extraction methods in place of supervised Mention Detection. While this modifications stems mostly from the lack of annotations on the social domain, it presents the advantage of avoiding biases learned from supervised annotations in medical literature that won't be present in the social domain. For example, MedMentions annotations tend to correspond to noun phrases, while many UMLS concepts correspond to prepositional or adjective phrases.

The SimString (Okazaki & Tsujii, 2010) method we use to perform ADM provides a similarity score which we use to adjust the confidence of detected mentions. We use YAKE (Campos et al. 2020) to perform unsupervised keyword extraction, admitting subparts (not limited to longest sequences). In our code we also provide the option of adding YAKE's extraction score to SimString's similarity for increased matching relevance in some cases (favours longer sequences).

## 2.2 Improving String-based Linking on the Social Domain

We improved the fit of our ADM string-based linker for the social domain through the integration of aliases and variants from additional resources beyond UMLS (release 2017 AA).

### 2.2.1 Updated UMLS Metathesaurus

The latest UMLS release (2020 AA) contains an enriched set of aliases for many concepts in comparison to the version we used for MedLinker. Most notably, many concepts were expanded with more colloquial aliases, much more likely to match the terminology found on the social domain.

### 2.2.2 MedMentions Annotations

In our development of MedLinker, we were concerned with evaluating the performance of our solution, and focused on the MedMentions dataset for this purpose due to its size and difficulty. However, MedLinker-Social isn't designed to be evaluated in the same setting, so we use all splits of the MedMentions annotations to complement UMLS aliases. While these annotations from MedMentions are still from the medical domain (i.e. PubMed abstracts), we expect these annotations to still contribute some useful diversity.

### 2.2.3 Health Thesaurus

The Centers for Disease Control and Prevention's National Center for Health Marketing (CDC) has compiled a thesaurus of medical terms for the purpose of improving health communication. This thesaurus, called Plain Language Thesaurus for Health Communications, contains over 1,000 entries for terms related to all types of medical concepts.

The thesaurus is released in a Word (.docx) document, with no UMLS identifiers (or for any other formal ontology), so this integration required some considerable effort on converting the document to a more parsable format, followed by associating entries to UMLS identifiers. The UMLS association step was assisted by the string-matching method described up to this section, although the final set of processed thesaurus entries still required further manual corrections.

### 2.2.4 Lexical Variants

The work of Klein et al. (2019) recently collected and annotated tweets for the purpose of training NLP methods to detect reports from mothers about birth defects related to their children. In that effort, the authors also contributed a list of commonly observed variations (mostly mispellings) for expressions of medical terms on social media (Twitter). While our project's interest in *preterms* isn't perfectly aligned with the topic of *birth defects*, we still find enough of an overlap in these topics to warrant integrating this resource into our solution.

Unlike the previously described resource integrations, these variations aren't added to specific concept's alias sets. Instead, we use the variations to correct input before applying our Entity Linker.

| System | Resources | # Concepts | # Aliases |
|---|---|---|---|
| MedLinker | UMLS 2017 | 1,394,544 | 1,537,763 |
| MedLinker-Social | UMLS 2020 | 2,854,770 | 4,423,132 |
| | +MedMentions | 2,854,770 | 4,451,383 |
| | +Health Thesaurus | 2,854,770 | 4,452,284 |

Table 1: Impact of resource integration and comparison against MedLinker.

## 3 Experiments

Below we describe the experiments we've run based on extractions obtained using MedLinker-Social. Here we also provide more details about the corpora, packages and parameters used in this project, built on Python 3.6.5.

### 3.1 MedLinker-Social Parameters

The YAKE-based Mention Detection was performed with an ngram size of 5, and the remaining parameters were used with default values (v0.4.3). The SimString component of MedLinker-Social's Entity Linking was performed with the same parameters as used in Loureiro & Jorge (2020) (also regular n-grams of size 5 and char n-grams of size 3). All extractions reported here were obtained with a confidence threshold of 0.5 (see Sec. 2.1).

### 3.2 Corpora

#### 3.2.1 Reddit Corpus

We compiled a set of submissions from the Reddit social networks in order to run experiments employing MedLinker-Social. Considering our focus on preterms, we collected submissions from the most related communities within Reddit (called subreddits). Additionally, given that most of these communities target pregnancy more generally, we attempted to filter submissions by keywords related to preterms. This set of keywords was compiled semi-automatically from various sources, including frequent keywords on both our Reddit and EuroPMC corpus, the lexicon of Klein et al. (2019), and selected concepts from EFCNI's article on preterm health conditions.

To avoid irrelevant submissions in our corpus, we also require submissions to have at least 2 upvotes. The final complete Reddit Corpus is composed of 282,399 submissions and over 56M tokens.

#### 3.2.2 EuroPMC Corpus

One of our goals in this project involves comparing how preterm related concepts are discussed in medical literature and social networks. As such, we also compiled a corpus from medical literature related to preterms based on abstracts collected by EuroPMC and made accessible through their official API. To focus the abstracts on our topic, we queried the API for papers with the keywords 'preterm' or 'low birth weight'. The resulting corpus is composed of 319,075 abstracts with over 75M tokens (in the same order of magnitude as our Reddit corpus).

### 3.3 Concept Semantics

#### 3.3.1 UMLS Embeddings

We use static word embeddings in our comparison of how preterm related concepts are discussed in medical literature and social networks. To learn these static embeddings, we follow the approach and parameters Zhang et al. (2019) which learn subword-level static word embeddings based on fastText (200 dimensions, window size of 20, v0.9.2, defaults for remaining). After learning word/subword embeddings from our Medical and Reddit corpora independently, we compute concept-level embeddings from the average of the embeddings corresponding to every one of its aliases, provided the alias occurs at least once in our Reddit corpus, and the concept at least 10 times.

### 3.3.2 Semantic Space Visualization

For visualizing the 200-dimensional spaces for each domain, we follow the standard practice in NLP of using T-SNE (van der Maaten & Hinton, 2008) to perform dimensionality reduction to just 2-dimensions that can be represented in a scatter plot. In addition to this, we also color each the points corresponding to each concept in accordance to the concept's semantic type (STY) as defined in the UMLS ontology.

We used scikit's implementation of T-SNE from the sklearn package with defaults parameters (v0.21.3), and the interactive plots we show on our website (and screen capture here) were built using the bokeh package (v2.2.3).

### 3.3.3 Semantic Space Correlation

We also report on correlations for select concepts with regards to their representations in the medical semantic space and the social semantic space, learned when computing these concept-level embeddings. We compute this correlation based on the cosine similarities of each of the selected concepts to its 1,000 nearest neighbors, in each of the two spaces. Then, we expand these sets of 1,000 similarities to the larger set considering the similarities of every concept in the union of the sets from the neighbors in each space (of size 2,000 in the worst case). Finally, we derive our correlation p-value from the Pearson correlation metric, using scipy's implementation (v1.1.0).

### 3.4 Keyword Frequencies

Word embeddings may provide us some insight into how the topics are being discussed, but this representation is based on how concepts are related between themselves. With word clouds, we aim to provide insights of a different nature. We use frequency-based word clouds for two purposes: 1) show which keywords are most influential for representing specific concepts; 2) show which keywords most often co-occur with the concept, at the document-level (i.e. within each submission). To better highlight the relevant concepts that most often co-occur, we first find the average distribution of related concepts for all concepts, and then normalize each concept's distribution by subtracting that average distribution.

We represent word clouds using the wordcloud package (v1.8.1). Our word clouds also take advantage of the confidence metric described in Section 2.1, by using each linking's confidence as weights to bias frequencies towards highest confidence keywords.

## 4 Analysis

We've run our experiments on a set of 32 UMLS concepts matched to the contents of an EFCNI article on preterm-related health conditions (linked below). In this section we present part of our analysis for a single concept (C0004044 - Suffocating) and discuss some findings. Full analysis for the 32 concepts is available at this project's website.

- Article: `https://www.efcni.org/health-topics/in-hospital/health-conditions-of-preterm-infants/`

- Project: `http://danlou.github.io/medlinker-social/`

Due to time and resource constraints, these analyses aren't based extractions from our complete Reddit Corpus, just the most recent 34,605 submissions (resulting in 5.9M total concept matches).

### 4.1 Word Clouds

The word cloud in Figure 1, displaying words matched against the C0004044 (Suffocating) concept, shows the wide variety of ways that the concept can be referred to. This figure also showcases how the low confidence threshold we've set for these experiments can be beneficial by allowing for additional context (e.g. 'afraid baby would suffocate' vs. 'son died due to asphyxiation'). However, for other concepts, such as C0428977 (Bradycardia) which has the alias 'slowed pulse', the low confidence threshold produces too many false-positive matches (the word 'slowed' is sufficient to hit the threshold).

The keywords from related concepts displayed in Figure 2 show that C0004044 (Suffocating) is highly associated with sleeping conditions, which matches our expectations considering the topic. This particular example also shows that the proposed method for creating related word clouds is able to overcome the dominant influence of keywords very common across all submissions (e.g. 'baby').

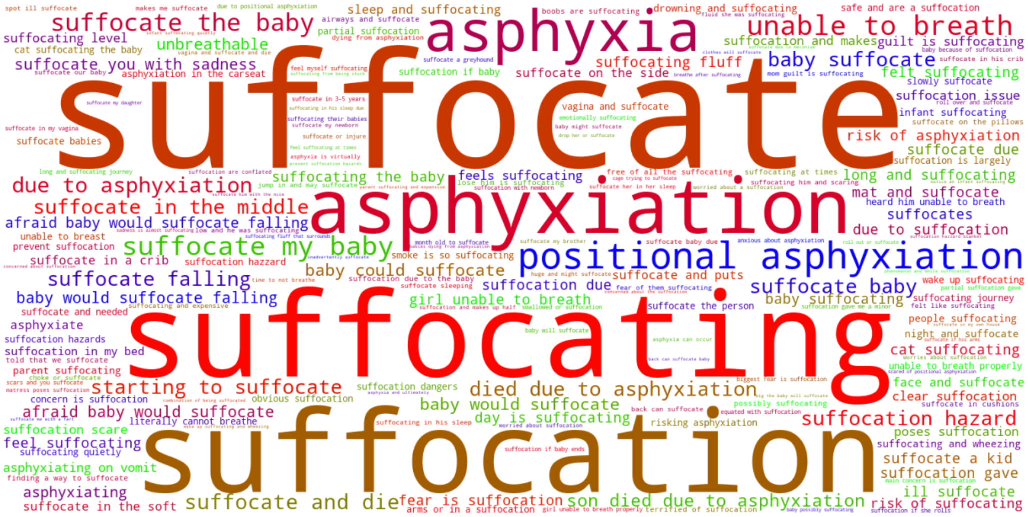

Figure 1: Keywords referring to C0004044 (Suffocating), sized according to weighted frequency.

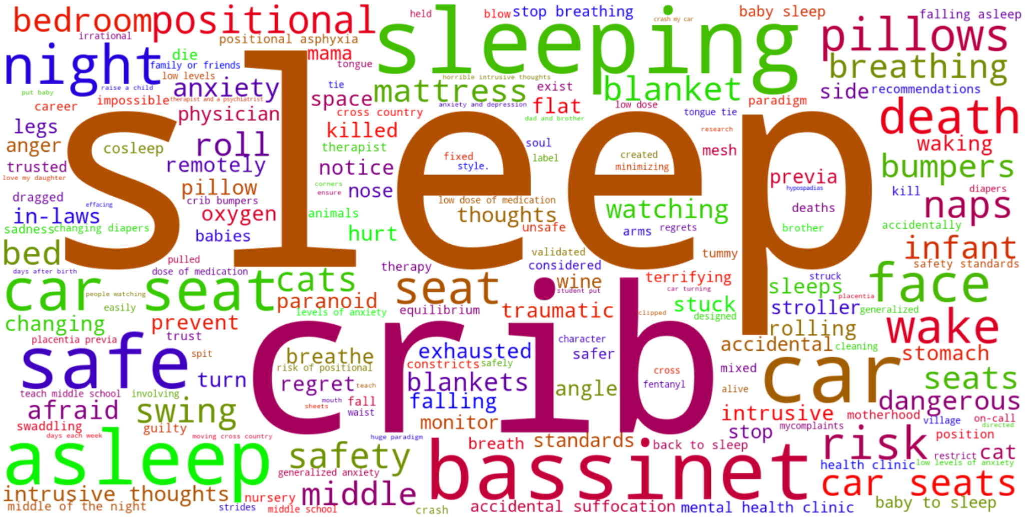

Figure 2: Keywords for concepts co-occurring with C0004044 (Suffocating), sized according to weighted frequency.

## 4.2 Embedding Visualization and Correlation

The embedding visualizations we developed for this project are designed to be experienced through the project website, where users can navigate the space and preview information about every neighbor. Still, we can report here that there are very significant differences in the sets of neighbors from each domain space. For example, on the medical space (Figure 3), the clusters closest to C0004044 (Suffocating) are associated with drowning, accidents, falls and poisoning, whereas on the social space (Figure 4) we find concepts related to distressed emotional states in much closer proximity, while the other causal associations appear less clustered and more distant.

This finding about the neighbors of C0004044 is also in tune with the negative correlation computed between the neighbors of this concept in both spaces (as described in Section 3.3.3). Conversely, concepts such as C0000832 (Placental Abruption) or C0022876 (Premature Labor) show much higher positive correlation.

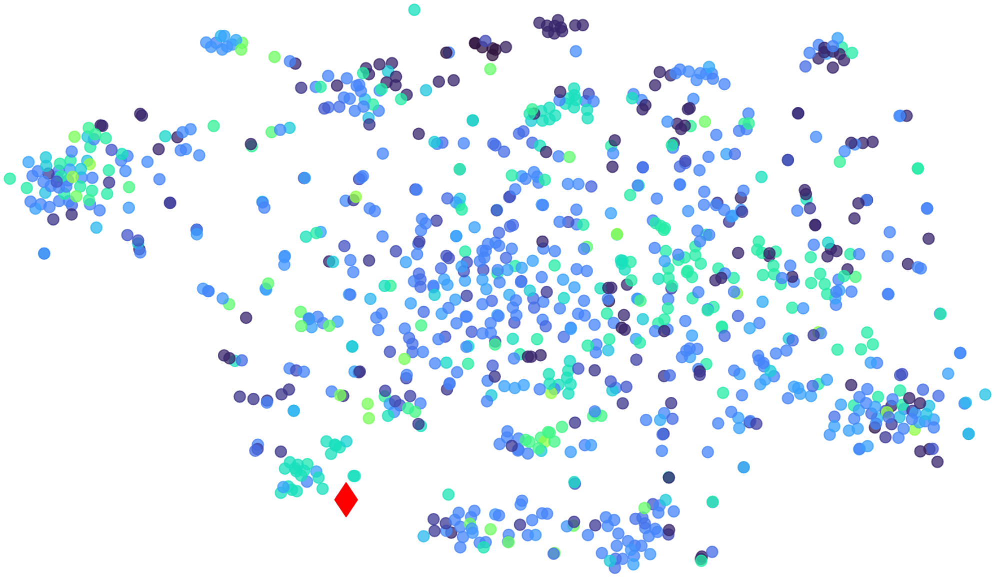

Figure 3: T-SNE visualization of 1,000 closest neighbors (concepts) of C0004044 in the medical domain (represented as red diamond). Colored according to semantic type.

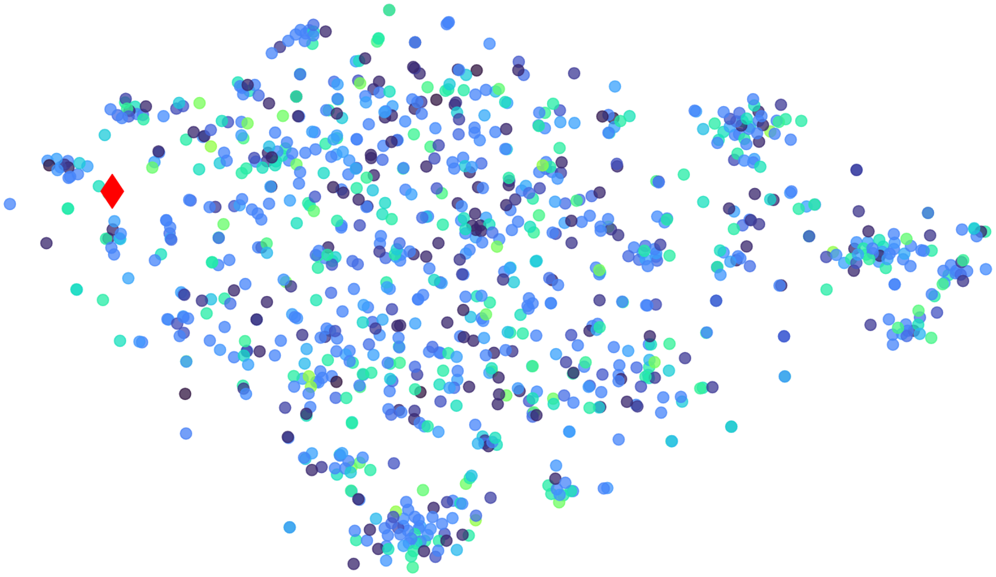

Figure 4: T-SNE visualization of 1,000 closest neighbors (concepts) of C0004044 in the social domain (represented as red diamond). Colored according to semantic type.

## 5 Conclusion

Our contributions in this project are threefold: 1) Compilation of an unsupervised corpus of social media posts targeting preterm-related concepts; 2) Development of MedLinker, and the MedLinker-Social adaptation, to more accurately identify highly-specific and versatile UMLS concepts from social media; 3) Showcase different analyses exploring information obtained by applying our tools to social media.

In this report's introduction, we already covered the reasons motivating our reinforcement of string-based matching methods, namely ADM, instead of pursuing more sophisticated solutions taking advantage of the latest developments in NLP, similarly to what we accomplished successfully with MedLinker for the medical domain. While we're certain that MedLinker-Social is better suited for social media than MedLinker, we should note that we also expect our efforts to improve ADM for the social domain through vocabulary expansion to be near the tipping point of the classical precision-recall trade-off: *expanding* the vocabulary (or relaxing confidence thresholds) will significantly increase false-positives, while *trimming* it (or increasing confidence thresholds) will significantly increase false-negatives (and tend towards direct dictionary lookups).

For this reason, we conclude that future work should start by focusing on strategies to build an annotated dataset of UMLS concepts, similar to MedMentions, but for the social domain. Furthermore, contextual NLMs trained on the social domain have very recently been released (e.g. Barbieri et al. 2020), and it's certain that the number and quality of these will improve in the short-term. With these two resources at disposal, we recommend replicating the MedLinker approach described in Loureiro & Jorge (2020).

### Acknowledgments

The research leading to these results has received funding from the European Union's Horizon 2020 - The EU Framework Programme for Research and Innovation 2014-2020, under grant agreement No. 733280.

## References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 789–798, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1073. URL https://www.aclweb.org/anthology/P18-1073.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL https://www.aclweb.org/anthology/2020.findings-emnlp.148.

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. 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)*, pp. 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://www.aclweb.org/anthology/D19-1371.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Ari Z Klein, Abeed Sarker, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. Towards scaling twitter for digital epidemiology of birth defects. *NPJ digital medicine*, 2(1):1–9, 2019.

Daniel Loureiro and Alípio Mário Jorge. Medlinker: Medical entity linking with neural representations and dictionary matching. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (eds.), *Advances in Information Retrieval*, pp. 230–237, Cham, 2020. Springer International Publishing. ISBN 978-3-030-45442-5.

S. Mohan and Donghui Li. Medmentions: A large biomedical corpus annotated with umls concepts. *ArXiv*, abs/1902.09476, 2019.

Naoaki Okazaki and Jun'ichi Tsujii. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 851–859, 2010.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL `http://www.jmlr.org/papers/v9/vandermaaten08a.html`.

Shikhar Vashishth, Rishabh Joshi, Ritam Dutt, Denis Newman-Griffis, and C. Rosé. Medtype: Improving medical entity linking with semantic type prediction. *ArXiv*, abs/2005.00460, 2020.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9, 2019.