Biomedical Concept Normalization over Nested Entities with Partial UMLS Terminology in Russian

Natalia Loukachevitch[1], Andrey Sakhovskiy[2,3], Elena Tutubalina[2,4,5]
[1] Moscow State University, Moscow, Russia
[2] Sber AI, Moscow, Russia
[3] Skolkovo Institute of Science and Technology, Moscow, Russia
[4] Kazan Federal University, Kazan, Russia
[5] Artificial Intelligence Research Institute, Moscow, Russia
{andrey.sakhovskiy,tutubalinaev}@gmail.com

Abstract

We present a new manually annotated dataset of PubMed abstracts for concept normalization in Russian. It contains over 23,641 entity mentions in 756 documents linked to 4,544 unique concepts from the UMLS ontology. Compared to existing corpora, we explore two novel annotation characteristics: the nestedness of named entities and the incompleteness of the Russian medical terminology in UMLS. 4,424 entity mentions are linked to 1,535 unique English concepts absent in the Russian part of the UMLS ontology. We present several baselines for normalization over nested named entities obtained with state-of-the-art models such as SapBERT. Our experimental results show that models pre-trained on graph structural data from UMLS achieve superior performance in a zero-shot setting on bilingual terminology.

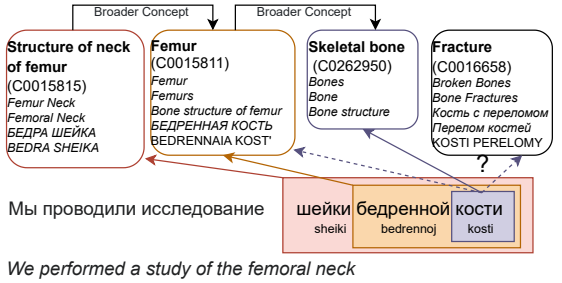
Keywords: medical concept normalization, entity linking, dataset, Russian language, PubMed abstracts

1. Introduction

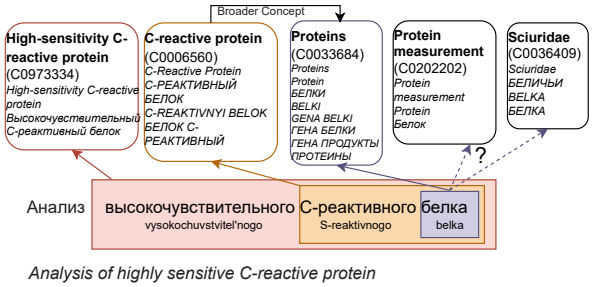
Biomedical entities are used in a variety of biomedical applications, including relational knowledge discovery (Chen et al., 2016; Bonner et al., 2022), clinical decision making (Sutton et al., 2020; Peiffer-Smadja et al., 2020), and information retrieval (Lee et al., 2016; Fiorini et al., 2018; Soni and Roberts, 2021). However, mentions of diseases, drugs, and other concepts in free-form texts are highly variable. This challenge can be addressed by medical concept normalization (MCN; also called medical concept linking), which is the task where entity mentions are mapped against a large set of medical concept names and their concept unique identifiers (CUIs) from a knowledge base (KB).

The biomedical domain is characterized by extensive KBs such as the Unified Medical Language System (UMLS) (Bodenreider, 2004). UMLS represents over 166 lexicons/thesauri with over 15M concept names from 27 languages. However, about 71% of the concept names are labeled in English. Other languages occur much less: the Russian part of UMLS includes translations of three sources and only amounts to 1.96% of the English UMLS in vocabulary and 1.62% in source counts (NIH).

MCN faces several significant challenges. Among them, the incompleteness of medical terminology stands out as a major challenge. Related studies follow annotation guidelines, which restrict the medical terminology to either (i) the source (e.g., English SNOMED-CT (Spackman et al., 1997) in SemEval-2014 Task 7 (Pradhan et al., 2014), SemEval-2015 Task 14 (Elhadad et al., 2015),



(a)



(b)

Figure 1: Examples of nested entities, incorrect candidates (dotted arrows) and their linking (solid arrows). UMLS concepts are represented by CUIs and a set of synonymous concept names.

MCN (Luo et al., 2019), COMETA (Basaldella et al., 2020)) or (ii) the target language (Nesterov et al., 2022; Miranda-Escalada et al., 2020a,b). In particular, a recent corpus for clinical MCN RuCCoN (Nesterov et al., 2022) focuses on concepts within the Russian part of UMLS, highlighting the chal-

LREC-COLING 2024, pages 2383–2389

lenge that many terms have not been translated from English into Russian. Despite the failure to normalize these mentions to a concept, the mention itself may still contain valuable and relevant information that would be ignored in NLP applications if CUI-less mentions are discarded.

The complex structure of entities presents another challenge for MCN. This structure would allow for multiple embedded entities but is much harder for manual labeling and has not been done in the abovementioned MCN corpora. As shown in Fig. 1, all nested entities help disambiguate each other. In Fig. 1a, an entity "kosti" (pale blue) is linked to C0262950 (Skeletal Bone), which lacks Russian translations, two alternative concepts: C0015811 (Femur) and C0016658 (Fracture), have partial lexical matches with the entity. In Fig. 1b, an entity "S-reaktivnogo belka" (pale orange) is linked to C0015811 (High-sensitivity C-reactive protein). A nested entity "belka" (pale blue) exactly matches C0036409 (Sciuridae, squirrels), while the correct concept C0033684 (Proteins) is the broader concept for C0015811 in UMLS. Therefore, nested entity annotations allow linking internal entities to equivalent UMLS concepts and can provide additional context for linking in a nested group of entities, especially in the absence of translation to the target language.

In this work, we present a new manually annotated dataset of PubMed abstracts for concept normalization in Russian. We design our annotation guidelines to account for a complex structure of nested disease, anatomy, and chemical mentions and the partial nature of medical terminology (Sec. 2.2). We present baselines for nested MCN and evaluate state-of-the-art BERT-based (Devlin et al., 2019) models (Sec. 3). Our dataset and code are available at: https://github.com/nerel-ds/NEREL-BIO.

2. Dataset

2.1. Basic Dataset with NER Labeling

We supplement with entity linking labeling the NEREL-BIO dataset, the only large-scale available dataset of Pubmed abstracts in Russian with nested named entity annotation (NER) labeling (Loukachevitch et al., 2023). NEREL-BIO dataset is annotated with 37 entity types, which can be nested or intersect each other. The nestedness of annotation (up to six layers of depth) allows more comprehensive coverage of biomedical concepts. Linking nested entities to biomedical concepts is important for biomedical retrieval systems based on UMLS. If we rely solely on flat entities, the concept ID would typically be assigned to longer entities, such as "pain in the head" (UMLS CUI C0018681). However, if a user searches for the disease concept ID for "pain" (CUI C0030193), this nested entity would be missed. In such cases, more complex annotations, including nested entities, become essential. By utilizing nested entities in the annotation process, we can overcome the limitations of flat entities and ensure that relevant information is captured accurately.

2.2. Annotation Process and Principles

The UMLS linking was made for three entity types in NEREL-BIO: ANATOMY, CHEM, and DISO entities among all. The selection of semantic types for annotation was based on several criteria, including frequency in the collection, popularity in biomedical research, and the "reliability" of annotation. UMLS semantic types and frequencies of these entity types are given in Tab. 1. Tab. 2 presents the statistics on the number of mentions and unique CUIs at different levels of nestedness. Annotators mapped each mention to a CUI using *Brat* (Stenetorp et al., 2012) with UMLS 2020 AB. To speed up labeling, automatic preprocessing steps were carried out, including the removal of markup for entities of other types and the elimination of duplicate entities to ensure a single mention for linking. Automatic linking to UMLS concepts in Russian was performed using a baseline that relied on ranking SapBERT representations (Liu et al., 2021) (Sec. 3). Following the markup completion, entity mentions were restored, and links for mentions of the same entities remained consistent, with an option to make corrections if needed.

Four terminologists and a moderator experienced in terminological studies, including the biomedical domain, were involved in the annotation. Before the annotation began, the annotators read annotation guidelines and received training from the moderator to ensure consistency. Each instance was carefully reviewed by an annotator and then checked by the moderator to maintain the quality of the annotations. Following the methodology described by Luo et al. (2019), we calculated the Inter-Annotator Agreement (IAA) as the accuracy of the annotations. For the purpose of calculation, we selected a sample of 11 documents with app. 850 entities, which were labeled by two independent annotators. The achieved IAA is 78.37%.

2.3. Analysis of Partial Coverage of the Russian language in UMLS

The problem with both manual and automatic linking was partial coverage of UMLS concepts with Russian translations of medical terms. In contrast to other studies restricted to a specific language, our study required annotators to identify an appropriate UMLS concept even when a Russian translation was not available, which often required significant effort. The UMLS concepts without Russian translations can be subdivided into different

Entity type	UMLS semantic type	#Mentions	#CUIs	#CUIs without Russian
ANATOMY	A1.2 Anatomical structure	8456	1260	619
CHEM	A1.4.1 Chemical	4748	1201	377
DISO	B2.2.1.2 Pathologic Function	10437	2153	539
Total		23641	4544	1535

Table 1: Summary on entity mentions of three types linked with the UMLS concepts.

Mention depth	# mentions	# uniq. CUIs
Non-nested	10695	2464
Nested, depth 0	4842	2211
Nested, depth 1	6407	1456
Nested, depth 2	1414	361
Nested, depth 3	169	63
Nested, depth 4	5	3

Table 2: Statistics on the number of mentions and unique CUIS at different levels of nestedness.

categories, and annotators had to take specific actions for accurate linking within each category:

1) **Well-known single-word terms** such as Complication (C0009566), Cirrhosis (C0023890), Bone (C0023890). Interestingly, it was found that some of these terms were absent from the Russian language translations. However, it was discovered that some of these terms could be found in the Russian variants of more complex terms. For instance, the Russian term 'oslozhnenie' (complication) was identified in the Russian variant 'nevrologicheskoe oslozhnenie anestezii' for concept C0854693 (Anaesthetic complication neurological), which is part of the translated MedDRA (Brown et al., 1999). We note that the concept Complication (C0009566) is included in several resources, such as the NCI thesaurus (Fragoso et al., 2004), but has not been translated into Russian.

During the automatic entity linking preprocessing, single-word terms that are absent in the UMLS are often mistakenly linked to concepts that have existing Russian translations, as seen in the example mentioned above. While the annotators can usually identify the correct concept, this process still requires careful attention to ensure accuracy.

2) **Multiword terms with transparent compositional semantics**, such as Vascular Endothelial Cells (C1257792) and Mild depressive disorder (C0588006), can be accurately translated using online translation systems. During manual annotation, the annotators are required to first translate the given term into English, query the resulting translation, and then select the correct concept from the UMLS search results.

3) **Drug names with clear symbol transliteration**, such as cytoflavin (C1701400) and mexidol (C0128329), can be accurately identified using translation. This enables the relevant UMLS concept to be detected with greater ease and accuracy.

4) **Latin abbreviations**: miR-155 (C2003121), let-7a (C1708690). These terms are usually correctly linked by an automatic linker.

5) In the case of **trademarks**, it is often possible to match them with the active substance of the drug they represent (e.g., 'furasol' can be matched with furazidin (C0878309)). Hence, annotators must search for the drug trademark in a drug registry.

6) **Difficult cases**. For some terms, simple translation does not give a correct English term. In this case, the annotators may employ the following tools and resources: (1) Searching for the term in Wikipedia and attempting to locate the corresponding English page; (2) Using the Latin term for anatomic structures; (3) Searching for a more general UMLS concept and checking its related narrower concepts using UMLS relations to determine the appropriate concept; (4) Searching for Russian scientific papers that discuss a target term and attempting to find translations or keywords in English. For example, the term 'vnutrenneye slukhovoye overstayed is translated into English by Google as 'internal auditory opening'. The correct translation is 'internal acoustic opening', which is a term variant for the UMLS concept 'Structure of porus acusticus internus' (C0229513). This link can be found via Latin terminology. The term 'vyvorot nizhnego veka' was translated as 'inversion of the lower eyelid', but the correct translation (e.g., as found in Wikipedia) is 'lower eyelid ectropion' (C0521736).

The proportions of the categories (estimated on a term sample) are as follows: • Category 2. Multiword terms – 43.4%; • Category 1. Single-word terms – 26.7%; • Category 6. Difficult cases – 15%; • All other categories – 15%.

3. Baselines & Evaluation

In this work, we approach the MCN task as the ranking task. Given a mention m, our goal is to find k closest concept names from a vocabulary. We evaluate models using top-k accuracy as the metric, following Liu et al. (2021); Nesterov et al. (2022); Sakhovskiy et al. (2023). We conduct two types of experiments: (1) zero-shot evaluation on the **whole NEREL-BIO set** (Sec 3.1); (2) evaluation of supervised baselines on two data splits (Sec. 3.2). These two data splits are as follows: (i) the **random split**, which is an 80/20 train-test (18976/4745 mentions) split with an additional constraint: each mention in the test set is part of a nested entity; (ii) the **hard split**, where each (CUI, mention) from the test set is unseen during training.

Model	Ru & Eng concepts		Non-Russian concepts	
	@1	@5	@1	@5
SapBERT	**.5379**	**.6652**	.1799	.3347
GEBERT	.5186	.6206	**.2251**	**.3388**

Table 3: Zero-shot evaluation on the whole NEREL-BIO and its subset of concepts absent in Russian UMLS using combined English-Russian dictionary of concept names from the UMLS.

Model	Data split			
	Random		Hard	
	@1	@5	@1	@5
Zero-shot SapBERT	.7520	.8471	.7038	.8031
BioSyn	.8703	.9185	.7618	.8504
BioSyn + S_{nested}	.8714	.9177	.7603	.8427
BioSyn + reranking	**.8725**	**.9208**	**.7622**	**.8508**

Table 4: Evaluation of simple supervised nested MCN baselines against non-nested BioSyn model with the SapBERT encoder on two data splits of NEREL-BIO with mono-lingual Russian dictionary. Each test set contains nested entities only.

For each concept, we sample a single random mention from a nested entity for the test set and put the remaining mentions into the train set (15031/3920 train/test mentions). All non-nested entities are included in the train set. For both splits, 10% of train mentions are used as a dev set for hyperparameter optimization. We drop mentions that are mapped to the non-Russian part of the UMLS.

3.1. Cross-lingual Zero-shot Evaluation

We compare ranking models based on two language models (LMs) pre-trained on synonymous multilingual concept names from the UMLS:

- *SapBERT*: a BERT-based metric learning framework that pretrains on the UMLS positive/negative triplets (Liu et al., 2021).
- *GEBERT*: a BERT-based model which trains on UMLS using graph neural networks and contrastive learning (Sakhovskiy et al., 2023).

Tab. 3 shows the Acc@1 and Acc@5 metrics on the whole NEREL-BIO and its subset of mentions with no Russian concept name in UMLS. We can see that the existing multilingual models used in the unsupervised regime are not truly cross-lingual: they fail to map mentions in one language to concept names in another language (see "Non-Russian concepts" column. Here, we have a mention in Russian and a proper dictionary entry in English only). For example, multilingual SapBERT degrades from 54% Accuracy@1 to 18% when no vocabulary in the target language is available. Cross-linguality is a crucial property since most existing languages either have low resources or no resources at all (i.e., UMLS provides concept names in about 20 languages only) and thus cannot offer a fine-grained normalization dictionary.

3.2. Nested MCN with Supervised Models

In the nested case, a mention m can be a part of a nested entity $M = (m, m_1, \ldots, m_n)$. For the MCN task, the goal is to find k candidates $(c_1, c_2, \ldots, c_k)$ from a vocabulary given m. As a non-nested MCN baseline, we adopt BioSyn (Sung et al., 2020) which iteratively re-ranks candidates based on a sum of two similarity scores: (i) the dot-product of TF-IDF representations of m and c; (ii) the dot-product $S_d(c, m)$ of BERT embeddings e^d_m and e^d_c of m and c, respectively. The model is trained with negative marginal log-likelihood loss to produce higher scores for positive (c, m) pairs. Let m' denote the longest entity from M. For the nested case, we define a nested mention m_n as $m\,[SEP]\,m'$, where $[SEP]$ is a BERT model's special separator token. Similarly, $m\,[SEP]\,m'$ is a nested candidate c_n. We propose two nested MCN baselines that build upon BioSyn:

- **Nested score** is obtained as the dot-product of embeddings e^n_m and e^n_c of n_c and n_m: $S_{nested} = e^n_m \cdot e^n_c$. We modify the BioSyn's scoring function as a sum of three terms: $S_d(c, m), S_{sp}(c, m), S_{nested}(c, m)$ and leave other model components unchanged.
- **Reranking** baseline adopts a frozen fine-tuned BioSyn to perform a nestedness-aware re-ranking, adding a nested score $S_{nested}(c, m) = MLP([e^n_m; e^n_c])$ to BioSyn scores. MLP is a 2-layer perceptron with GeLU (Hendrycks and Gimpel, 2016) activations and $[\cdot\,;\cdot]$ is the concatenation.

Each supervised baseline was trained for 20 epochs with a learning rate of $1 \cdot 10^{-5}$ and batch size of 16 using the Adam optimizer (Kingma and Ba, 2015). For prediction, we loaded the model parameters from the epoch with the highest Acc@1 on the dev set.

The evaluation results for nested MCN baselines are presented in Tab. 4. The re-ranking baseline gives an insignificant improvement over BioSyn, but the nested problem statement requires further exploration. The removal of overlapping mentions from train and test sets leads to a significant Acc@1 drop of more than 10%. As we discard mentions that are mapped to the non-Russian UMLS concepts, the zero-shot quality is higher than presented in Tab. 3 which further highlights the complexity of cross-lingual MCN.

4. Conclusion

In this paper, we presented a unique corpus for its annotation scheme and language which is curated to maintain high-quality annotations of biomedical nested entity mentions in PubMed abstracts with concepts from multilingual UMLS knowledge graph (KG). Different evaluation scenarios were designed to compare the performance of LMs augmented with KG. A substantial decrease in zero-shot performance (over -30% accuracy) of multilingual models between the whole corpus and its subset of concepts absent in the Russian UMLS highlights the need for future research in this area. This corpus can serve as a challenging yet reliable evaluation benchmark for the development of multilingual models specific to the biomedical domain.

Acknowledgments

The work has been supported by the Russian Science Foundation grant # 23-11-00358. We would also like to thank the anonymous reviewers for their comments on this paper.

Limitations

Other biomedical corpora in Russian The closest corpora are RuCCoN (Nesterov et al., 2022), where entities from clinical records are linked to the Russian part of UMLS, and the XL-BEL benchmark (Liu et al., 2021), which links Wikipedia mentions to the Russian part of UMLS. We note that these datasets differ from ours in three key aspects: the source of texts, the structure of nestedness, and the terminology used. As such, it was out of the scope of our work to evaluate our trained baselines for nested MCN on these existing datasets.

No state-of-the-art for nested MCN. In this paper, we have introduced several variations of models that consider the nested structure of entities. However, we have observed that these models do not differ significantly from the state-of-the-art models used for the classical MCN, where the nearest concept is predicted independently for each entity. Therefore, more complex neural architectures remain open research questions.

A cross-lingual benchmark for nested MCN remains to be built. In our dataset, we focus on linking entity mentions in Russian to English and Russian UMLS concepts. As we were unable to find similar biomedical corpora containing nested entities linked to UMLS, cross-lingual knowledge transfer from other languages was beyond the scope of our paper. It should be noted that while the initial NEREL-BIO dataset contains annotations for over 700 Russian and 100 English abstracts (Loukachevitch et al., 2023), only the Russian abstracts with mentions are currently publicly available on GitHub[1].

Transfer from general-domain data. The NEREL-BIO scheme expands on the annotation capabilities of the general-domain NEREL (Loukachevitch et al., 2021) by providing annotation guidelines for nested named entities in the biomedical domain. However, since NEREL links nested entity mentions to Wikidata, we did not evaluate our models in a cross-terminology setting, where MCN models trained with Wikidata terminology are evaluated with UMLS terminology.

Ethics Statement

The dataset introduced in this paper involved only new annotations on top of the existing, publicly available NEREL-BIO dataset of PubMed abstracts. Dataset annotation was conducted by annotators, and there are no associated concerns (e.g. regarding compensation). Each annotator was paid an hourly wage of $25, which corresponds $1000 monthly wage. The minimum monthly wage in Russia for full-time employment is under $200. As discussed in limitations, we believe these new annotated datasets serve as a starting point for the evaluation of LMs on biomedical texts with complex entity structure in a zero-shot setup with incomplete health terminology in a target language.
Our annotations, code, and annotation guidelines will be released upon acceptance of this paper.

5. Bibliographical References

Stephen Bonner, Ian P Barrett, Cheng Ye, Rowan Swiers, Ola Engkvist, Andreas Bender, Charles Tapley Hoyt, and William L Hamilton. 2022. A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *Briefings in Bioinformatics*, 23(6).

Huaidong Chen, Wei Chen, Chenglin Liu, Le Zhang, Jing Su, and Xiaobo Zhou. 2016. Relational network for knowledge discovery through heterogeneous biomedical and clinical features. *Scientific Reports*, 6(1):29915.

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*, pages 4171–4186.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. Semeval-2015 task 14:

[1] https://github.com/nerel-ds/NEREL-BIO

Analysis of clinical text. In *proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310.

Nicolas Fiorini, Kathi Canese, Grisha Starchenko, Evgeny Kireev, Won Kim, Vadim Miller, Maxim Osipov, Michael Kholodov, Rafis Ismagilov, Sunil Mohan, et al. 2018. Best match: new relevance search for pubmed. *PLoS biology*, 16(8):e2005343.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sunwon Lee, Donghyeon Kim, Kyubum Lee, Jaehoon Choi, Seongsoon Kim, Minji Jeon, Sangrak Lim, Donghee Choi, Sunkyu Kim, Aik-Choon Tan, et al. 2016. Best: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PloS one*, 11(10):e0164680.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238.

NIH. 2023. Nih umls statistics.

Nathan Peiffer-Smadja, Timothy Miles Rawson, Raheelah Ahmad, Albert Buchard, P Georgiou, F-X Lescure, Gabriel Birgand, and Alison Helen Holmes. 2020. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clinical Microbiology and Infection*, 26(5):584–595.

Sameer Pradhan, Noémie Elhadad, Wendy W Chapman, Suresh Manandhar, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. In *SemEval@ COLING*, pages 54–62.

Andrey Sakhovskiy, Natalia Semenova, Artur Kadurin, and Elena Tutubalina. 2023. Graph-enriched biomedical entity representation transformer. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings*, volume 14163 of *Lecture Notes in Computer Science*, pages 109–120. Springer.

Sarvesh Soni and Kirk Roberts. 2021. An evaluation of two commercial deep learning-based information retrieval systems for covid-19 literature. *Journal of the American Medical Informatics Association*, 28(1):132–137.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650.

Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17.

6. Language Resource References

Basaldella, Marco and Liu, Fangyu and Shareghi, Ehsan and Collier, Nigel. 2020. *COMETA: A Corpus for Medical Entity Linking in the Social Media*.

Bodenreider, Olivier. 2004. *The unified medical language system (UMLS): integrating biomedical terminology*. Oxford University Press.

Brown, Elliot G and Wood, Louise and Wood, Sue. 1999. *The medical dictionary for regulatory activities (MedDRA)*. Springer.

Fragoso, Gilberto and de Coronado, Sherri and Haber, Margaret and Hartel, Frank and Wright, Larry. 2004. *Overview and utilization of the NCI thesaurus*. Wiley Online Library.

Liu, Fangyu and Vulić, Ivan and Korhonen, Anna and Collier, Nigel. 2021. *Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking*.

Loukachevitch, Natalia and Artemova, Ekaterina and Batura, Tatiana and Braslavski, Pavel and Denisov, Ilia and Ivanov, Vladimir and Manandhar, Suresh and Pugachev, Alexander and Tutubalina, Elena. 2021. *NEREL: A Russian Dataset with Nested Named Entities, Relations and Events*.

Loukachevitch, Natalia and Manandhar, Suresh and Baral, Elina and Rozhkov, Igor and Braslavski, Pavel and Ivanov, Vladimir and Batura, Tatiana and Tutubalina, Elena. 2023. *NEREL-BIO: a dataset of biomedical abstracts annotated with nested named entities*. [link].

Yen-Fu Luo and Weiyi Sun and Anna Rumshisky. 2019. *MCN: A comprehensive corpus for medical concept normalization*.

Miranda-Escalada, Antonio and Farré, Eulàlia and Krallinger, Martin. 2020a. *Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results.*

Miranda-Escalada, Antonio and Gonzalez-Agirre, Aitor and Armengol-Estapé, Jordi and Krallinger, Martin. 2020b. *Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of CLEF eHealth 2020*.

Nesterov, Alexandr and Zubkova, Galina and Miftahutdinov, Zulfat and Kokh, Vladimir and Tutubalina, Elena and Shelmanov, Artem and Alekseev, Anton and Avetisian, Manvel and Chertok, Andrey and Nikolenko, Sergey. 2022. *RuCCoN: Clinical Concept Normalization in Russian*. Association for Computational Linguistics. [link].

Spackman, Kent A and Campbell, Keith E and Côté, Roger A. 1997. *SNOMED RT: a reference terminology for health care.* American Medical Informatics Association.