RuCCoN: Clinical Concept Normalization in Russian

Aleksandr Nesterov[1], **Galina Zubkova**[1], **Zulfat Miftahutdinov**[2], **Vladimir Kokh**[1], **Elena Tutubalina**[2,3,4], **Artem Shelmanov**[5,7], **Anton M. Alekseev**[6], **Manvel Avetisian**[1], **Andrey Chertok**[4,5], **Sergey Nikolenko**[6,7,8]

[1] Sber AI Lab, Moscow, Russia
[2] Kazan Federal University, Kazan, Russia
[3] HSE University, Moscow, Russia
[4] Sber AI, Moscow, Russia
[5] AIRI, Moscow, Russia
[6] St. Petersburg Department of Steklov Mathematical Institute, St. Petersburg, Russia
[7] ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia
[8] Neuromation OU, Tallinn, Estonia
{AAlNesterov,GVZubkova}@sberbank.ru

Abstract

We present RuCCoN, a new dataset for clinical concept normalization in Russian manually annotated by medical professionals. It contains over 16,028 entity mentions manually linked to over 2,409 unique concepts from the Russian language part of the UMLS ontology. We provide train/test splits for different settings (stratified, zero-shot, and CUI-less) and present strong baselines obtained with state-of-the-art models such as SapBERT. At present, Russian medical NLP is lacking in both datasets and trained models, and we view this work as an important step towards filling this gap. Our dataset and annotation guidelines are available at https://github.com/sberbank-ai-lab/RuCCoN.

1 Introduction

Electronic health records and other clinical texts contain patient histories through the progression of diseases and represent a treasure trove of information for medical specialists. This information is often unstructured, concealed in free-form text, which leads to the need for natural language processing on medical texts. Mentions of diseases, symptoms, drugs, and other concepts are highly variable, and since the medical vocabulary is very large, entity linking and concept normalization become hard and important problems. State-of-the-art models are increasingly successful in high-resource languages such as English or Spanish, where labeled datasets include ShARe/CLEF eHealth 2013 Task 1 (Suominen et al., 2013), SemEval-2014 Task 7 (Pradhan et al., 2014), SemEval-2015 Task 14 (Elhadad et al., 2015), MCN (Luo et al., 2019), CANTEMIST (Miranda-Escalada et al., 2020a), CodiEsp (Miranda-Escalada et al., 2020b), and others. However, little has been done for medical entity linking in many languages that are high-resource in other regards. One example is Russian: it is among top 10 languages in the world and has many NLP datasets and resources, but the medical part of Russian NLP is underdeveloped. The Russian UMLS includes translations of Medical Dictionary for Regulatory Activities (MedDRA) (Brown et al., 1999), Logical Observation Identifiers Names and Codes (LOINC) (Forrey et al., 1996), and Medical Subject Headings (MeSH) (Coletti and Bleich, 2001), but it still only amounts to 1.8% of the English UMLS in vocabulary and 1.36% in source counts (NIH 2021, a).

In this work, we present RuCCoN (**Ru**ssian **C**linical **Co**ncept **N**ormalization), a new labeled dataset for clinical concept normalization in Russian. We have employed medical professionals to label the dataset based on concepts from the Russian UMLS (Section 2). Moreover, we present several types of test sets for various settings, including stratified, zero-shot, and CUI-less settings (Section 2.4). We evaluate several state-of-the-art models on RuCCoN, including various fine-tuning variations, and check whether labeled data in Russian is necessary (spoiler alert: it is) by testing cross-lingual concept normalization from English (Section 3). Our results can serve as baselines for RuCCoN and cross-lingual concept normalization.

Findings of the Association for Computational Linguistics: ACL 2022, pages 239 - 245
May 22-27, 2022

2 RuCCoN Dataset

2.1 Basic Dataset with NER Labeling

We supplement with entity linking labeling the only large-scale available dataset of clinical free-text notes in Russian with named entity recognition (NER) labeling (Shelmanov et al., 2015), created by researchers and practitioners from the Scientific Center of Children Health (SCCH). The corpus is based on medical histories of over 60 SCCH patients with allergic and pulmonary disorders and diseases. It contains discharge summaries, radiology, echocardiography, and ultrasound diagnostic reports, recommendations, and other records from a number of different physicians. Documents in the corpus are deidentified: all names are removed, dates are altered. The corpus, freely available for research purposes[1], contains 160 fully annotated texts with nearly 250,000 tokens. It has over 18,200 annotated entities, over 7,400 attributes and 3,500 relations with 7 types of entities: "Disease", "Symptom", "Drug", "Treatment", "Body location", "Severity", "Course". The nearest counterparts for English are the corpus of the Shared Annotated Resources (ShARe) initiative (Pradhan et al., 2015) and the corpus of Strategic Health IT Advanced Research Project: Area 4 (SHARPn) (Rea et al., 2012).

2.2 Annotation Process and Principles

Annotators mapped each mention to a concept unique identifier (CUI) from the Unified Medical Language System (UMLS) ontology (Bodenreider, 2004). The goal of entity normalization is to assign the same identifier to different synonyms of a given medical concept; e.g., "anemic heart infarction" and "myocardial infarction" refer to the same concept with CUI C0027051. Annotation was carried out in *Brat* (Stenetorp et al., 2012) with UMLS 2020 AB release. To speed up labeling, each text fragment was linked to CUI from UMLS automatically with the *tf-idf* baseline method. Annotators were allowed to use web search and the *UMLS Metathesaurus Browser* (NIH 2021, b) for meta-information. Each entity was independently annotated by three annotators. Following (Luo et al., 2019), we calculate Inter-Annotator Agreement (IAA) as the accuracy of the markups matched by at least two annotators over all annotated mentions. At least two annotators linked an entity to the same concept from the ontology in 13,125 cases and annotated 1032 entities as CUI-less; IAA was 78.37%. In 3900 cases when all annotators disagreed, the expert annotator with Ph.D. in medicine (the first author of the paper) was asked to decide whether the CUI selected by one of the annotators was in fact correct. After this procedure we obtained a corpora with 16,028 entities linked to 2,409 concepts and 1,293 entities linked with no concept (CUI-less). Table 1 shows the basic statistics of the dataset; percentages are obtained by greedily choosing the first relevant semantic type for a given CUI. Semantic types best represented in our annotation are *Disease or Syndrome* (≈ 22%), *Body Part, Organ, or Organ Component* (17%), *Organic Chemical* (14.5%), *Finding* (7%), *Sign or Symptom* (6.5%), and *Pathologic Function* (4%). Annotation guidelines were created by an expert with Ph.D. in medicine. The dataset was labeled by three annotators with higher education in different fields of medicine, two of them with Ph.D. in medicine. Each annotator was paid an hourly wage of $55 for about 80 hours of labeling, so each annotator was paid $4400; the minimal monthly wage in Russia for full-time employment is under $200.

Semantic Type	Train		Test	
	#	%	#	%
Disease or Syndrome	2032	18.11	848	17.62
Body Part, Organ, etc.	1670	14.88	699	14.52
Organic Chemical	1502	13.38	694	14.42
Finding	896	7.98	373	7.75
Sign or Symptom	677	6.03	254	5.27
Therapeutic or Preventive Proc.	542	4.83	202	4.19
Pathologic Function	449	4	188	3.9
Am. Acid, Peptide, or Protein	358	3.19	160	3.31
Organ or Tissue Function	339	3.02	136	2.81
Body System	150	1.33	73	1.5

Table 1: Top 10 semantic types counts in RuCCoN.

2.3 Annotation Design and Challenges

During the annotation process, we have encountered a number of challenges that are specific to Russian and other low-resource languages.

Lack of Russian translation for UMLS concepts. Many terms have not been translated from English into Russian. This often holds for terms that characterize the severity of symptoms, morphological characteristics of anatomical formations, and body locations; examples include "regular shape" ("форма правильная") or "patent lumen"("просвет свободен"). In these cases, we obtain NER labeling for general entity types such as "Disease" or "Body

[1] http://nlp.isa.ru/datasets/clinical

location", but there are no CUIs that annotators could link in Russian.

Combining many related concepts into one NER fragment. Many phrases annotated in NER labeling as a single entity could be split into several separate and/or nested entities. These cases most often found in morphological descriptions of anatomical formations (e.g., "average size (left lobe = 44mm, 1-st segment 11, right lobe = 93mm), smooth contour, homogeneous parenchymatous tissue, average chogenicity") and cases where adjectives characterizing a concept are combined into a single fragment; e.g., "mild repolarization disorders" ("легкие нарушения процесса реполяризации") could be labeled as a single entity but here the adjective "mild" ("легкие") might also be separated from the main concept "repolarization disorders" ("нарушения процесса реполяризации"). This happens since NER labeling is usually done for "flat" NER rather than nested; nested NER would allow for multiple embedded entities but is much harder for manual labeling, and has not been done in this case. In our dataset, annotators were instructed to link several CUIs to a single text fragment in these cases.

Redundancy of the UMLS vocabulary. Some UMLS concepts in Russian have different CUIs even though they are phrased in exactly the same way, and the CUIs have different semantic types; for example, "amyloidosis" ("амилоидоз") appears for both C0002726 (type: Disease or Syndrome) and C0268381 (type: Neoplastic Process). Also, some concepts have different CUIs while they are synonymous in their meaning; for example, "acholic stool" ("ахоличный стул") has a code C2675627 and "pale stool" ("светлый стул") has a code C0232720. In such cases, annotators were advised to choose a more appropriate CUI based on its meta-information provided in the UMLS.

Complex rephrasing. In entity linking, annotators have to change the wording to establish correspondences between mentions and concepts, relying on their domain knowledge and comprehensive search for synonyms. In Russian, this is complicated by minor inconsistencies in the UMLS translation itself: several different CUIs may either have minor semantic differences that cannot be distinguished or overlap significantly in their meaning. E.g., "adenoid hypertrophy" ("гипертрофия аденоидов") may be annotated as "nasopharyngeal tonsil hypertrophy (adenoids)" ("гипертрофия глоточных миндалин (аденоиды)") or "hypertrophy of adenoids exclusively" ("гипертрофия исключительно аденоидов"), two different CUIs. This effect often leads to inconsistencies between annotators.

Subset	# entities	# unique entities	# concepts
Full train	12189	5435	2031
In-KB train	11220	4934	2030
Full test	5132	2689	1232
In-KB test	4808	2464	1231
Zero-shot test	434	417	379
Stratified test	1266	1199	576
RWN med.	2319	1666	635
XL-BEL	681	610	510
MCN	13609	5979	3792

Table 2: Dataset statistics.

2.4 Train/test Splits

We release the full annotated corpus along with three test sets, setting aside 30% of the corpus and then applying different filtering strategies. Table 2 shows the statistics for each split.

Stratified. In this case, we filter the test set so that each UMLS concept appears in the test set appears at least once in the training set, but not this specific mention from the test set (Miftahutdinov and Tutubalina, 2019). Thus, 100% of concepts in this test set are covered in the training set, but no mentions in the training set are literally the same as mentions in the test set.

Zero-shot. In this case, we filter the test set to contain only novel concepts that do not appear in the training set at all. In other words, the *Stratified* split is designed to ensure that the model encounters the same concepts in the training, development, and test sets, but with different surface forms, while the *Zero-Shot* split instead exposes models to unseen terms and concepts in the development and testing sets, making it the harder setting of the two.

CUI-less. In this case, we supplement the random train/test split with 30% of the cases where there is no CUI associated with an entity. This set is intended to test whether a linking system is able to refrain from linking to a concept when there is no suitable concept in the vocabulary ("CUI-less" category in CLEF/SemEval challenges). We call the "full test set" and "full train set" the subsets with this addition of CUI-less cases, and use "in-KB" for subsets without CUI-less mentions. Stratified and zero-shot settings are commonly used in general domain entity linking, but the CUI-less setting is specific for medical data.

XL-BEL, RuWordNet medical, and MCN train-

Model	In-KB test		Full test		Stratified test		Zero-shot test	
	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
Tf-Idf	37.58%	46.98%	-	-	25.83%	34.20%	26.27%	41.01%
Multilingual BERT	29.01%	33.74%	29.15%	33.16%	12.32%	16.35%	15.90%	19.35%
RuBERT	25.17%	28.22%	24.05%	25.66%	11.53%	14.53%	13.82%	17.51%
SapBERT	45.84%	56.41%	37.18%	37.47%	30.02%	40.44%	29.49%	40.78%
SapBERT+MCN	46.51%	56.45%	43.67%	53.23%	30.41%	40.60%	27.88%	41.47%
SapBERT+RWN	45.47%	55.12%	43.30%	50.19%	29.94%	39.42%	29.03%	38.48%
SapBERT+XL-BEL	47.77%	58.74%	40.80%	42.30%	32.54%	42.97%	29.95%	45.16%
SapBERT+RuCCoN	59.26%	68.99%	53.39%	60.02%	47.31%	61.45%	32.95%	47.47%
SapBERT+RuCCoN+RWN	57.84%	68.55%	52.67%	58.79%	47.79%	63.67%	32.49%	46.31%
SapBERT+RuCCoN+XL-BEL	58.78%	68.05%	53.20%	59.80%	46.52%	59.08%	33.41%	48.85%
SapBERT+RuCCoN+RWN+XL-BEL	58.55%	67.82%	52.65%	59.20%	50.32%	62.48%	33.41%	45.85%

Table 3: Evaluation results with test set filtering.

ing sets. In our basic evaluation setting, we use only our own labeled dataset for training and testing. However, we could also supplement the training set with other resources. First, the XL-BEL cross-lingual biomedical entity linking benchmark maps entity mentions from *Wikipedia* to UMLS in a number of languages, including Russian (Liu et al., 2021b). Second, we have applied the following linking procedure to the medical part of *RuWordNet* (RWN) (Loukachevitch et al., 2016): found all lemmas of RWN synsets from the medical domain, intersected these synsets with lemmatized Russian UMLS terms, composed the vocabulary of synsets that have at least one lemma in UMLS, and filtered out exact matches with UMLS, resulting in a set of senses not contained in UMLS but from synsets with another sense contained in both UMLS and RuWordNet. We note that both RWN and XL-BEL have small intersection of 13 and 9 CUIs with zero-shot RuCCoN test set, respectively.

Third, we also test cross-lingual entity linking with models trained on the MCN (Medical Concept Normalization) dataset (Luo et al., 2019), a large-scale manually annotated corpus in English for clinical concept normalization produced from a corpus released for the 4th i2b2/VA shared task (Uzuner et al., 2011). Statistics for all supplementary datasets are also shown in Table 2.

3 Evaluation

For entity linking, we use ranking based on embeddings of a mention and a possible concept. Each entity mention and concept name are first passed through a model that produces their embeddings and then through a pooling layer that yields a fixed-sized vector. The inference task is then reduced to finding the closest concept name representation to the entity mention representation in a common embedding space, where the Euclidean distance can be used as the metric. Nearest concept names are chosen as top-k concepts for entities. For ranking, we use the publicly available code[2] from (Tutubalina et al., 2020).

We compare ranking models based on several different embeddings:

1. *Tf-idf*: standard sparse *tf-idf* representations constructed on character-level unigrams and bigrams;
2. *BERT*: multilingual BERT embeddings with no fine-tuning (Devlin et al., 2019); this is a cross-lingual baseline that has not been trained on biomedical texts;
3. *RuBERT*: Russian BERT embeddings (Kuratov and Arkhipov, 2019) trained on the Russian part of Wikipedia and news data;
4. *SapBERT*: a BERT-based metric learning framework that generates hard triplets based on the UMLS for large-scale pre-training (Liu et al., 2021a) and also allows for a cross-lingual variant trained on XL-BEL (Liu et al., 2021b).

Additionally, we compare several variations of fine-tuning on datasets with training sets via synonym marginalization as suggested by the authors of *BioSyn* (Sung et al., 2020):

1. *SapBERT+RuCCoN*, with fine-tuning on our target train set of EHRs;
2. *SapBERT+MCN*, with tuning on the MCN set;
3. *SapBERT+WRN*, on the dataset extracted from the medical part of the RuWordNet thesaurus;

[2] https://github.com/insilicomedicine/Fair-Evaluation-BERT

4. *SapBERT+XL-BEL*, on the the Russian part of XL-BEL;

5. *SapBERT+RuCCoN+RWMXL-BEL*, on the combination of all three sets.

For training, we have used the publicly available code provided by the authors at https://github.com/dmis-lab/BioSyn with the following parameters: the number of top candidates k is 20, the mini-batch size is 16, the learning rate is 1e-5, the dense ratio for candidate retrieval is 0.5, the number of epochs is 5. To deal with *nil* prediction, we apply the strategy from (Miftahutdinov et al., 2021); a mention is out of KB if the nearest candidate is further than a threshold in terms of weighted average of two distances: minimum distance of false positives and maximum distance of true positives, as computed on the train set.

Following previous works on entity linking (Suominen et al., 2013; Pradhan et al., 2014; Wright et al., 2019; Phan et al., 2019; Sung et al., 2020; Miftahutdinov et al., 2021; Tutubalina et al., 2020), we use top-k accuracy as the evaluation metric: $\text{Acc@k} = 1$ if the correct CUI is retrieved at rank $\leq k$, otherwise $\text{Acc@k} = 0$. Table 3 shows the Acc@1 and Acc@5 metrics for our test sets. We see that SapBERT significantly outperforms other models, and steadily improves the results as more datasets are added for fine-tuning. Note how SapBERT trained on RuCCoN is much better on the full test set that SapBERT trained on other data, but the difference almost disappears on the zero-shot test, suggesting that it was almost entirely due to specific entities labeled in the training set. This confirms the need to label additional data to further improve the results of even the best state-of-the-art entity linking models, which is what RuCCoN itself provides for the Russian language. Another result is that fine-tuning on additional medical data is generally beneficial; e.g., we have found that SapBERT fine-tuned on English clinical notes outperforms basic SapBERT consistently across all datasets in our study.

4 Error Analysis

To better understand the quality of our best model, we analyzed its erroneous predictions. For analysis, we randomly selected 100 erroneous predictions, which were then analyzed by an expert annotator with Ph.D. in medicine. As can be seen from Table 4, most of the errors are related to the lexical similarity of incorrectly predicted entities (for example, the text "concor" was incorrectly associated with the entity C0009738 "Congo" due to the similarity of spelling; for the same reason, the text "decrease in EF" was incorrectly associated with the entity C0520837 "decrease in FEV"). An interesting fact is that in second place in the frequency of errors are predictions close in meaning to the source text. For example, the source text "bilateral acute maxilloethmoidal sinusitis" was associated with the entity C0155806 "acute ethmoiditis". This entity is not a complete synonym of the source text, but it is very close to its meaning. It should be noted that the errors described in the last two rows of the table are not inherently errors. Some of the errors are related to the complete synonymy of ground truth and model prediction. For example, the text "biliary tract dysfunction" was annotated as C0005395 "pathology of the biliary tract", while the model predicted the entity C0005424 "biliary tract disease", which in its meaning is a complete synonym, but does not coincide with the golden truth according to CUI.

Cause of error	Number of mentions
No obvious reason	18
Lexical similarity	38
Nested entity	11
Semantic similarity	19
Complete synonymy	9
Annotation error	5

Table 4: Manual evaluation of incorrect predictions of the SapBERT+RuCCoN model on 100 randomly selected mentions from the in-KB test set.

5 Conclusion

In this work, we have presented RuCCoN, a new clinical concept normalization dataset in Russian, labeled by medical professionals and accompanied with several train/test splits for fair evaluation in various settings. We make RuCCoN publicly available for research purposes, and we hope that future works will make use of RuCCoN as a training and evaluation resource.

Acknowledgements

The work on experiments with MCN, WRN, and XL-BEL corpora was done by Z.M. and supported by the Russian Science Foundation [grant number 18-11-00284]. We are grateful to annotators.

References

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

Elliot G Brown, Louise Wood, and Sue Wood. 1999. The medical dictionary for regulatory activities (meddra). *Drug safety*, 20(2):109–117.

Margaret H Coletti and Howard L Bleich. 2001. Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association*, 8(4):317–323.

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.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. Semeval-2015 task 14: Analysis of clinical text. In *proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310.

Arden W Forrey, Clement J Mcdonald, Georges DeMoor, Stanley M Huff, Dennis Leavelle, Diane Leland, Tom Fiers, Linda Charles, Brian Griffin, Frank Stalling, et al. 1996. Logical observation identifier names and codes (loinc) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clinical chemistry*, 42(1):81–90.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. 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.

Fangyu Liu, Ivan Vulic, Anna Korhonen, and Nigel Collier. 2021b. Learning domain-specialised representations for cross-lingual biomedical entity linking. In *ACL/IJCNLP (2)*, pages 565–574.

Natalia V Loukachevitch, German Lashevich, Anastasia A Gerasimova, Vladimir V Ivanov, and Boris V Dobrov. 2016. Creating russian wordnet by conversion. In *Computational Linguistics and Intellectual Technologies: papers from the Annual conference Dialogue*, pages 405–415.

Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. 2019. Mcn: A comprehensive corpus for medical concept normalization. *Journal of biomedical informatics*, 92:103132.

Zulfat Miftahutdinov, Artur Kadurin, Roman Kudrin, and Elena Tutubalina. 2021. Medical concept normalization in clinical trials with drug and disease representation learning. *Bioinformatics*, 37(21):3856–3864.

Zulfat Miftahutdinov and Elena Tutubalina. 2019. Deep neural models for medical concept normalization in user-generated texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 393–399, Florence, Italy. Association for Computational Linguistics.

Antonio Miranda-Escalada, Eulàlia Farré, and Martin Krallinger. 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. In *IberLEF SEPLN*, pages 303–323.

Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. 2020b. Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.

NIH 2021. a. Nih umls statistics.

NIH 2021. b. Umls metathesaurus browser.

Minh C Phan, Aixin Sun, and Yi Tay. 2019. Robust representation learning of biomedical names. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3275–3285.

Sameer Pradhan, Wendy Chapman, Suresh Man, and Guergana Savova. 2014. Semeval-2014 task 7: Analysis of clinical text. In *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval 2014*. Citeseer.

Sameer Pradhan, Noémie Elhadad, Brett R South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana Savova. 2015. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 22(1):143–154.

Susan Rea, Jyotishman Pathak, Guergana Savova, Thomas A Oniki, Les Westberg, Calvin E Beebe, Cui Tao, Craig G Parker, Peter J Haug, Stanley M Huff, et al. 2012. Building a robust, scalable and standards-driven infrastructure for secondary use of ehr data: the sharpn project. *Journal of biomedical informatics*, 45(4):763–771.

AO Shelmanov, IV Smirnov, and EA Vishneva. 2015. Information extraction from clinical texts in russian. In *Computational Linguistics and Intellectual Technologies*, pages 560–572.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Junichi 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.

Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 212–231. Springer.

Elena Tutubalina, Artur Kadurin, and Zulfat Miftahutdinov. 2020. Fair evaluation in concept normalization: a large-scale comparative analysis for bert-based models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6710–6716.

Ozlem Uzuner, Brett South, Shuying Shen, and Scott DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18:552–6.

Dustin Wright, Yannis Katsis, Raghav Mehta, and Chun-Nan Hsu. 2019. Normco: Deep disease normalization for biomedical knowledge base construction. In *Automated Knowledge Base Construction*.