KAZAN FEDERAL UNIVERSITY

as a manuscript

Miftahutdinov Zulfat

ENTITY LINKING MODELS IN BIOMEDICAL DOMAIN

PhD Dissertation Summary
for the purpose of obtaining academic degree
Doctor of Philosophy in Computer Science

Kazan — 2022

The PhD Dissertation was prepared at Kazan Federal University
Academic Supervisor: Tutubalina Elena Viktorovna, PhD, Kazan Federal University.

1 Introduction

Topic of the thesis

This work focuses on the development of named entity linking methods in a biomedical domain. Named entity linking in a biomedical domain is also called medical concept normalization. The task of named entity linking is to match the natural language phrase with the corresponding concept from the knowledge base if there is one. In this case, a concept is an element of the knowledge base that reflects some notion in a specific area of knowledge. For example, in the UMLS knowledge base, the concept with ID C0004057 corresponds to the medical notion of the drug Aspirin. In addition to the identifier and name, a concept can also have various synonyms and relationships with other concepts. Thus, the task of medical concept normalization is to associate a fragment of the text with a specific concept from the knowledge base.

Even though the entity linking task is widely studied in the general domain, within the medical domain there are some characteristic features: (i) a large variety of knowledge bases, which are often not static and are updated at different intervals; (ii) the complexity of creating appropriate corpora with a sufficient level of coverage due to the high qualification requirement for annotators; (iii) a high variability of forms of use within one concept - the same drug can have different names and many trade names adopted by chemists. The dissertation work analyzes, modifies, and synthesizes existing approaches to solve this problem. In particular, in works [1–4] the evaluation of the classification approach was carried out and a vector of semantic similarities was proposed, which characterizes the degree of similarity of an entity with each concept of terminology. Significant shortcomings were detected in quality evaluation procedures while studying the classification approach. Most of the dataset contains a high intersection (up to 60%) between the pre-defined training and test subsets. For a more realistic assessment of the models, we proposed to split the sample into non-overlapping training and test parts. The works listed above show the effectiveness of the proposed vectors of semantic similarity and the method of incorporating them into the classification approach. Nevertheless, a significant drawback of the classification approaches is the inability to recognize the concepts that are absent in the training set. In this regard, in the works [5; 6] we presented an approach based on metric learning for solving the problem of linking named entities. This approach is based on the construction of a single vector space for entities and concepts. The common space allows for similarity-based normalization and treats the task of named entities linking as a ranking task. In the work [7], we proposed a method for combining the classification and metric approaches based on a threshold value. The evaluation of proposed approach was carried out on Social Media Mining for Health Applications (SMM4H) 2019 Shared task (Task 3), 2020 (Task 3), and 2021 (Task 1c) [8–10]. It showed the best results among all teams of participants.

Relevance

The massive amount of textual data in various sources provides plentiful opportunities for their use as a health care resource. As data sources, we can consider social networks, databases of scientific articles, patents, or clinical trials.

Through Internet resources, users get the opportunity to exchange opinions and get nearly unlimited access to information about segments of the pharmaceutical market and medical information. In addition, clinical trials do not always provide a complete list of side effects. This is due to the fact that side effects often appear after prolonged use of the drug or have an effect only on a certain group of patients who have not participated in clinical trials. This fact leads to the large volumes of comments containing unexplored side effects for specific drugs. Using comments on medical products available on the Internet and going beyond simple keyword searching is both an opportunity and a challenging area of natural language processing. The application of such techniques to the Internet resources will make it possible to identify new side effects, find cases of drug misusage, and generate candidates for drug repurposing.

The second crucial resource for public health is scientific article databases. One such database is PubMed [11] which indexes biomedical articles. According to its content, it is of great interest to scientists engaged in medical research or the development of new drugs. The key point when using such databases of scientific articles is the ability to quickly access the needed information. You can try to solve this problem with general search engines. However, since the aim is to obtain more precise information, there are difficulties with the query formulation in general search engines. For instance, a scientist might examine all works that contain joint research on a particular gene and disease, or works in which genes interacted in a specific way. The nature of the queries largely determines the methods and tools used to build search engines. In particular, solving the problem of extracting and linking named entities is crucial in such search scenarios. The same is true for databases of medical patents and clinical trials.

One of the essential and often necessary stages in extracting structured information from a large amount of textual data is named entity linking. It remains crucial for processing all types of resources: Internet resources, databases of scientific articles, patents, and clinical trials. Of course, the named entity recognition should precede this step. However, it is not considered in this dissertation.

Traditional approaches to the medical concepts normalization are based on the use of dictionaries and knowledge bases. The most common system based on knowledge for mapping entities to concepts from the UMLS is MetaMap. This linguistic system uses a lexical search based on the generation of various variants of the input phrase. In this case, each generated variant is assigned a score that characterizes its proximity to the original phrase. Then all phrases that do not have an exact match in the UMLS knowledge base are filtered out. Among the remaining variants, we select the one with the highest similarity score. A major drawback of this approach is the low value of the recall metric. The next approach used in the medical concept normalization task is learning to rank. This approach was first applied to the normalization problem in the paper [12]. The DNorm system developed by the authors uses pairwise learning to rank, which utilizes vector representations of mentions and candidate terms from the UMLS. The vector representations are created based on TF-IDF metrics. The TaggerOne system described in the work [13] is an extension of the work of [12]. TaggerOne differs from DNorm in that TaggerOne uses Markov and semi-Markov models to jointly learn the task of named entity recognition and medical concept normalization. In recent years there has been a tendency to treat the problem of medical concept normalization from the classification approach point of view. For instance, convolutional neural networks used in work [14]. In the article, the authors have shown that the use of deep learning models leads to a significant increase in the F-measure compared to classical approaches. The works considered in this thesis also study the classification approaches. Namely, we examined convolutional and recurrent architectures of neural networks based on vector representations and pre-trained language models ELMo and BERT. We proposed semantic similarity vectors and an integration method for these vectors into the classification approach. The proposed methods showed best results among other teams participating in the medical concept normalization shared tasks CLEF 2017 Task 1, SMM4H 2019 Task 3, SMM4H 2020

Task 3. During the studies, however, we noted the shortcomings of the classification approach and standard evaluation methods. One of the major problems of the classification approach is the lack of training examples covering all possible medical concepts. Due to the stated limits of the corpora, the classification approach is unable to recognize concepts that are not present in the training sample. Rule-based methods are free from this kind of limitation. However, rule-based approaches have low recall metrics. Therefore, new methods need to be developed to solve the problem of medical concept normalization based on modern natural language processing approaches that do not require all medical concepts in the training sample. Metric learning is one such approach since it does not require all concepts in the training set and, as shown in the work [5], is resistant to vocabulary changes. In the work [5] considered in the thesis, to solve the problem of named entity linking task, we proposed an approach based on metric learning. This approach constructs a common vector space for entities and concepts. This common space allows normalization based on similarity and treats the named entity linking task as a ranking task. In the work [7], we proposed a method for combining the classification and metric learning approaches based on a threshold. We evaluated the proposed approach in the SMM4H 2019 (Task 3), SMM4H 2020 (Task 3), and SMM4H 2021 (Task 1c) shared tasks. Based on the test results, the proposed approach showed the best results among all teams. The approaches proposed in the work [5; 6] have been integrated into Insilico Medicine's data processing pipelines. The thesis describes some of the models used in this platform and provides quality metrics on standard datasets.

This work aims to develop a set of efficient methods based on metric learning and negative sampling to solve the named entity linking task.

2 Key results and conclusions

Contributions. The main contribution of the work is the named entity linking models:

1. Models for the named entity linking based on the classification approach. We have proposed vectors of semantic similarity that have proven their effectiveness in the CLEF eHealth 2017 Task 1, SMM4H 2019 Task 3, SMM4H 2020 Task 3, SMM4H 2021 Task 1c shared tasks. We highlighted the drawbacks of the classification approach and standard evaluation

methods. In particular, the lack of training examples in datasets covering all possible medical concepts, and a large number of test examples duplicating elements of the training example. We proposed a method for evaluating models that eliminates the high level of the intersection of training and test samples.

2. A named entity linking model based on a metric learning approach. We showed the robustness of this model to vocabulary switches and the ability to recognize concepts that were not present in the training sample.
3. A named entity linking model based on a combination of classification and metric learning approaches. We demonstrated the effectiveness of this method in the SMM4H 2020 Task 3, SMM4H 2021 Task 1c shared tasks.

Theoretical and practical significance

The practical significance of the results stems from the fact that the developed models aimed to analyze texts from open sources, including the Internet, which contains an extensive set of medical information that can be used in research projects, and to improve healthcare. The theoretical significance lies in the new models for named entity linking task proposed in the thesis. Primarily, we have improved the models based on the classification approach and pointed out the drawbacks of such methods and the evaluation methodology. We proposed more reasonable evaluation strategies for medical concept normalization tasks. We proposed a method based on the metric learning approach to solve the problem of limited training data. Finally, we proposed an approach that allows combining the strengths of both solutions – classification and metric learning.

Key aspects/ideas to be defended.

1. A named entity linking model based on the classification approach using the features of semantic similarity.
2. A named entity linking model based on the metric learning approach.
3. A named entity linking model based on a combined approach.

Personal contribution

In the first article, the author proposed vectors of semantic similarity and models that integrate the proposed vectors into a classification approach. All experiments were carried out by the author. In the second and third articles, the author proposed the models trained using metric learning, triplet loss and negative sampling. All experiments in these articles were carried out by the author.

Publications and probation of the work

The author of the thesis is the primary author of 2 main articles on the topic of the thesis.

First-tier publications

1. **Miftahutdinov Z.** et al. Drug and Disease Interpretation Learning with Biomedical Entity Representation Transformer //European Conference on Information Retrieval. – Springer, Cham, 2021. [Scopus, ECIR - Core A conf.]
2. **Miftahutdinov Z.**, Kadurin A., Kudrin R., Tutubalina E. Medical concept normalization in clinical trials with drug and disease representation learning //Bioinformatics. – 2021. – T. 37. – №. 21. – C. 3856-3864 DOI: 10.1093/bioinformatics/btab474 (Q1, Impact Factor 2021 6.64) [Scopus]
3. Tutubalina, E., **Miftahutdinov, Z.**, Nikolenko, S., & Malykh, V. (2018). Medical concept normalization in social media posts with recurrent neural networks. Journal of biomedical informatics. — Vol. 84. — Pp. 93–102 DOI:10.1016/j.jbi.2018.06.006 (Q1, Impact Factor 2019 3.5) [Scopus, WOS]

Reports at conferences

1. The 10th International Conference on Analysis of Images, Social Networks and Texts, December 16, 2021, keynote. "Drug and Disease Interpretation Learning with Biomedical Entity Representation Transformer".
2. European Conference on Information retrieval, March 28, 2021. "Drug and Disease Interpretation Learning with Biomedical Entity Representation Transformer".
3. European Conference on Information retrieval, April 14, 2020. "On biomedical named entity recognition: experiments in interlingual transfer for clinical and social media texts".

4. The 28th International Conference on Computational Linguistic, December 8, 2020. "Fair Evaluation in Concept Normalization: a Large-scale Comparative Analysis for BERT-based Models".
5. The 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, July 28, 2019. "Deep Neural Models for Medical Concept Normalization in User-Generated Texts".

3 Content of the work

3.1 Classification approach and Semantic similarity features

One of the recently established approaches to medical concept normalization is the classification approach. Recent classification algorithms are based on deep neural networks with a softmax activation function on top of the last layer. Softmax generalizes the logistic function for the multidimensional case. In this case, the model is trained to approximate the probability distribution over the set of all concepts for the entity m. The advantage of this approach is the high precision of the concepts presented in the training set. However, one of the major problems of the classification approach is the lack of corpora covering all possible medical concepts. As shown in the work [1], the CADEC corpus [15] contains less than 5% of the SNOMED-CT concepts. Similar statements hold for most other corpora. For example, the PsyTAR corpus [16] contains less than 1% of the controlled vocabulary, the SMM4H corpus – less than 0.5% [17], the TwiMed corpus – less than 0.5% [18], the NCBI corpus [19] - less than 0.01%, TwADR-L corpus - less than 0.01% [14]. To circumvent such limitations in the work [7], considered in this thesis, we proposed a method for combination of classification and metric learning approaches.

We first considered the classification methods in the medical concept normalization task in the work [1] and enriched the first work in the article [2]. In this works, the proposed approach is based on two different vector representations of the entity span. The first is deep neural networks representation derived from the pre-trained language model BioBERT.

The second representation is based on the semantic similarities between the entity span and the medical concepts from the UMLS ontology. In this case, the semantic similarity is defined as the cosine distance between the TF-IDF or word2vec representation of the medical concepts and the entity span. Thus, each element of the semantic similarity vector indicates the degree of closeness of the entity span to the concepts from the UMLS ontology. Based on the obtained vector representations, we classify the entity span to the corresponding medical concept. The described approach is illustrated in the figure 1.

The architectures proposed in the articles [1;2] share the processing pipeline. At the first stage, the source text is converted into a vector representation using

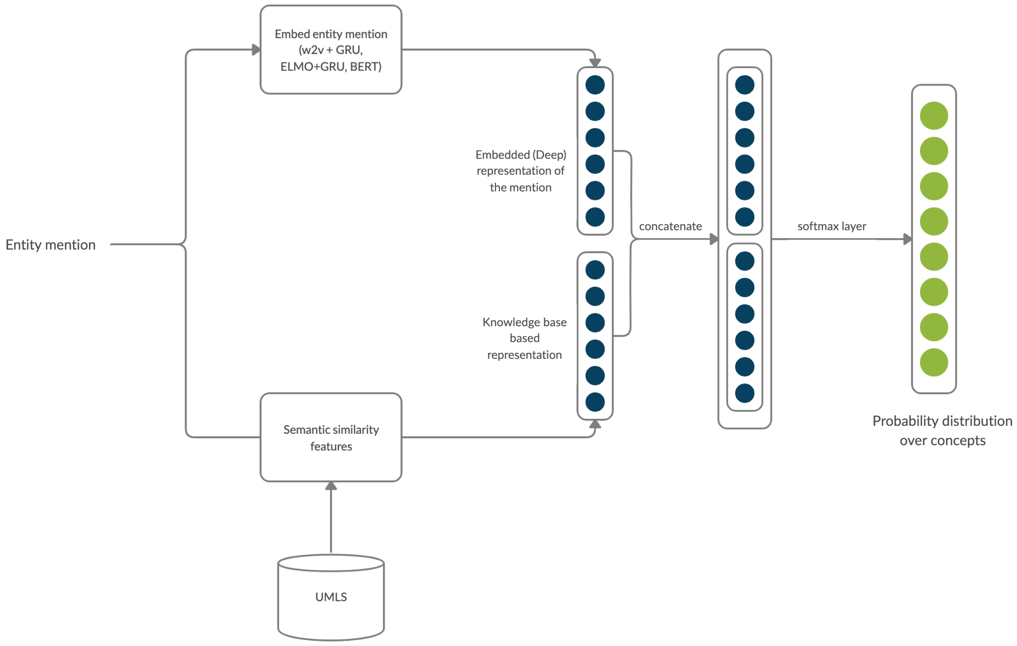
Embed entity mention (w2v + GRU, ELMO+GRU, BERT)
Entity mention
Embedded (Deep) representation of the mention
Knowledge base based representation
concatenate
softmax layer
Probability distribution over concepts
Semantic similarity features
UMLS

Figure 1: General scheme of the classification approach

deep neural networks according to the formula below:

$$y_m = red(Encoder(m)), \tag{1}$$

where Encoder is some trainable deep neural network, $red(\cdot)$ is a function that reduces a sequence of vectors into one vector. As an encoder network we considered several architectures:

- convolutional neural network based on word2vec representations
- recurrent neural network based on GRU with word2vec representations
- recurrent neural network based on GRU with word representations obtained from pre-trained language model ELMo [20].
- BERT pre-trained language model.

For convolutional and recurrent models, words vector representations were initialized with pre-trained word2vec vectors from the work [21].

There are several different variations of the $red(\cdot)$ function. In particular, in the works [1; 2] the following options have been utilized: select the output corresponding to the CLS token for the BERT pre-trained language model; apply the attention mechanism presented in the article [22] for a recurrent neural network.

Further, the vector of semantic similarity between the original expression and medical concepts from the UMLS ontology concatenated to the obtained vector representation based on deep neural networks. At the last stage, based on the obtained vector representations, entity span is classified to the corresponding medical concept using the $softmax$ activation function, as shown in the formula 2:

$$c_p = softmax([y_m : y_s]), \tag{2}$$

where y_m corresponds to the representation obtained from the deep neural network, y_s is the semantic similarity vector, $[y_m : y_s]$ is the concatenation operation.

Moreover, a new evaluation approach is proposed in the paper [1]. A detailed examination of the CADEC dataset revealed a high degree of overlap between the training and test examples - namely, that 60% of the test part consisted of samples from the training section. With such a ratio, it is difficult to predict the quality of the model on new data since the model that remembers the training set already reaches 60% accuracy. For a more realistic estimation of the models, we split the dataset into non-overlapping training and test parts. For this purpose, we removed duplicates from the original sample and grouped all entities within one concept. Thus, we got n groups, where n is the number of concepts presented in the set. Each group we divide into test and training splits. Further, we combine all training and test splits. This division guarantees that the test part will not overlap with the training part. Moreover, all concepts will be presented both in the test part and in the training part.

The tables 2 and 1 contain the results presented in the works [1] and [2] respectively. At the same time, in addition to the results on the disjoint split, table 1 shows the results on a random split into training and test samples. As can be seen, convolutional neural networks achieve 46% accuracy on the CADEC corpus, the lowest quality among all models. Recurrent neural networks achieve 64.5% accuracy. Recurrent neural networks with the attention mechanism and word2vec representations achieve 70.05% accuracy in the case of using a vector of semantic features. The model, which does not use semantic features, achieves 66.56% accuracy. The recurrent network based on the contextual vector representations of tokens from the ELMo network on the CADEC corpus achieves 71.68% and 74.70% accuracy with and without semantic similarity features respectively. The BERT model reaches 79.83 % with semantic features and 79.25% - without the features. Similar results were obtained on the PsyTAR and SMM4H 2019 Task 3 corpora. As you can see from the results, ELMo contextual vector representations

Table 1: Evaluation metrics for classification approach on CADEC, PsyTAR, SMM4H 2019 Task 3 datasets. Results for random and non-overlapping (Custom) dataset split into test and training parts presented.

Model	**CADEC**		**PsyTAR**		**SMM4H 2019 Task 3**
	Random	Custom	Random	Custom	Official
Baseline: match with training set annotation	66.09	0.0	56.04	2.63	67.12
DNorm [14]	73.39	-	-	-	-
CNN [14]	81.41	-	-	-	-
RNN [14]	79.98	-	-	-	-
Attentional Char-CNN [23]	84.65	-	-	-	-
Hierarchical Char-CNN [24]	-	-	-	-	87.7
Ensemble [25]	-	-	-	-	88.7
GRU+Attention	82.19	66.56	73.12	65.98	83.16
GRU+Attention w/ TF-IDF (MAX)	84.23	70.05	75.53	68.59	86.28
ELMo+GRU+Attention	85.06	71.68	77.58	68.34	86.60
ELMo+GRU+Attention w/ TF-IDF (MAX)	85.71	74.70	79.52	70.05	87.52
BERT	88.69	79.83	83.07	77.52	89.28
BERT w/ TF-IDF (MAX)	88.84	79.25	82.37	77.33	89.64

give better quality metrics than word2vec non-contextual representations. However, the best quality metrics are obtained with the BERT model. It should also be noted that the vector of semantic features increases quality metrics only in the case of using word2vec or ELMo representations. The above observations are the same for all corpora used in the article [2].

The evaluation of the proposed model based on BERT contextual representations was also carried out at the SMM4H 2019 shared task subtask 3. Subtask 3 consisted of two parts: named entity recognition and medical concept normalization. To extract named entities, we utilized the BERT language model with a classification layer for each token according to the BIO scheme. Table 3 shows the results for both the named entity recognition component and the evaluation of the end-to-end solution. As can be seen from the table, the classifier built on the pre-trained language model BERT that utilizes semantic similarity features showed the best results among the rest of the participants.

We also utilized the semantic similarity features in the variation of the medical concept normalization task. In some cases, the medical concept normalization task requires matching a phrase with several concepts from a medical knowledge base. This variation is often encountered in the processing of clinical texts and requires linking entities with the International Classification of Diseases (ICD). The International Classification of Diseases is a diagnostic tool used to monitor and classify health problems and death causes, and to provide information for clinical purposes.

Table 2: Evaluation metrics for classification approach on CADEC corpus. The results are given for a non-overlapping split of the dataset into test and training parts

Model	**Parameters**	**Accuracy**
CNN	HealthVec, 100 feature maps	46.19
CNN	PubMedVec, 100 feature maps	45.79
LSTM	HealthVec, 200 hidden units	64.51
LSTM	PubMedVec, 200 hidden units	64.24
GRU	HealthVec, 200 hidden units	63.05
GRU	PubMedVec, 200 hidden units	62.73
LSTM+Attention	HealthVec, 200 hidden units	65.73
LSTM+Attention	PubMedVec, 200 hidden units	64.92
LSTM+Attention	HealthVec, 100 hidden units	64.83
GRU+Attention	HealthVec, 200 hidden units	**67.08**
GRU+Attention	PubMedVec, 200 hidden units	66.55
GRU+Attention	HealthVec, 100 hidden units	66.56
with prior knowledge		
LSTM+Attention	HealthVec, 100, similarity: TF-IDF (ALL)	67.63
LSTM+Attention	HealthVec, 200, similarity: TF-IDF (ALL)	66.83
GRU+Attention	HealthVec, 100, similarity: TF-IDF (ALL)	69.92
GRU+Attention	HealthVec, 200, similarity: TF-IDF (ALL)	69.42
GRU+Attention	HealthVec, 100, similarity: W2V (ALL)	69.14
GRU+Attention	HealthVec, 100, similarity: TF-IDF (MAX)	**70.05**

This problem is used to be solved by rule-based methods. In particular, in the work [26], authors applied the Solr full-text search system to solve the medical concept normalization problem. The approach described in the paper achieved 84.8% accuracy on the [27] dataset. Cabot et al. used a combination of a dictionary approach and fuzzy matching algorithms, in their work [28]. Their system has achieved 80.38% accuracy. Along with the rule-based approach, plenty of works used a classification approach. In particular, [29] used support vector machine (SVM) multilabel classification along with a text preprocessing stage (removing stop words, removing diacritics, correcting some spelling errors). The method proposed by the authors reached 84.7 % F-measure.

In the articles [3; 4] considered in this thesis we proposed a new method for linking the texts from medical documents to formal medical concepts, based on

Table 3: Evaluation metrics of the classification approach on the SMM4H 2019 Task 3. The results of other methods are taken from the work [8]. EF_1 is the partial F-measure for the NER task, NF_1, NP, NR is the F-measure, the precision and recall of the end-to-end evaluation respectively.

Model	**E $\mathbf{F_1}$**	**N $\mathbf{F_1}$**	**N P**	**N R**
KFU@NLP Team	**66.0**	**43.0**	**36.0**	**54.0**
ensemble RNN & Few-Shot Learning	-	35.0	34.0	36.0
BERT + Flair + RNN	63.0	31.0	37.0	27.0
encoder-decoder (W biLSTM + attention)	60.0	21.0	22.0	20.0

deep neural networks of encoder-decoder architecture. We tackled the problem of medical concept normalization as a sequence to sequence problem. In this case, the source sequence is a sequence of tokens, and the target sequence is a sequence of ICD codes. The objective is to learn the correct dependencies between the source and the target sequences.

We utilized the architecture presented in the work [30] to implement the encoder-decoder model. A bidirectional LSTM network is used as an encoding network. A unidirectional LSTM network is used as a decoding network.

The semantic similarity features were integrated into the proposed method. The semantic similarity features were based on ICD 10 medical terms. The implemented medical concept normalization system achieved the best result in the first subtask of the CLEF eHealth 2017 shared task.

3.2 Metric learning and negative sampling

The medical concept normalization task can be viewed from the point of information retrieval. In this instance, an entity that needs to be linked to a concept from a medical knowledge base can be treated as a query, a set of concepts as a collection of documents that need to be sorted by relevance. The first element in the ranked list must be a concept corresponding to the entity. The document consists of fields representing the concept: a unique concept identifier and a set of concept names. In addition to these two fields, various medical knowledge bases may contain additional information. For example, the hierarchical structure of concepts or the semantic concept type [31]. However, in this work, the main fields for the medical concept normalization are a unique concept identifier and a variety of concept names (synonyms).

This approach was first applied to the normalization problem in the work [12]. The DNorm system is based on pairwise learning to rank approach, as features, this system utilizes TF-IDF representations of the entity mention and synonyms from the UMLS. The TaggerOne system described in the work [13] is a continuation of the previous work [12]. TaggerOne differs from DNorm in that TaggerOne uses Markov and semi-Markov models to jointly learn the task of named entity recognition and medical concept normalization. The system reaches 82.9% F-measure on the NCBI corpus. However, these works are based on n-gram text representations, which do not provide plentiful opportunities for modern text processing approaches.

The article [5], considered in this dissertation, proposes an approach to fine-tune the language model based on the Transformer architecture using metric learning, triplet loss and negative sampling, in particular. The model obtained during training with the metric learning approach on medical data hereinafter is referred to as DILBERT. The proposed approach allows building a common semantic vector space for entities and concepts from the knowledge base, where texts with similar meanings are located close to each other. This property allows to rank concepts based on distance function s and to solve the problem of medical concept normalization.

Following the notation suggested in [32], both entities and concepts are mapped to vector representations as follows:

$$y_m = red(T(m)); y_c = red(T(c)), \tag{3}$$

where T is a deep neural network of the transformer architecture, the weights of which can be updated during fine-tuning, $red(\cdot)$ is a function that reduces a sequence of vectors into one vector, m is an entity that needs to be linked to the corresponding concept, c - concept name. There are several implementations of the $red(\cdot)$ function, such as selecting the output corresponding to the CLS token or element-wise average pooling over all vectors to get a fixed-size vector. It has been empirically established that average pooling is the optimal option of the $red(\cdot)$ function. As a pre-trained model of the Transformer architecture, the BioBERT v1.1 model is used, pre-trained on the PubMed corpus of annotations of biomedical articles.

The relevance score of the candidate c_i for the entity m is given by the distance function applied to the corresponding vector representations. In the Ph.D. thesis, Euclidean and cosine distances are considered. However, the experiments did not show any significant difference between these two options in terms of the quality metrics. Consequently, the quality metrics are given only for the Euclidean distance.

$$s(m, c_i) = ||y_m - y_{c_i}||, \tag{4}$$

As shown in the [33], the BERT language model could be tuned in such a way that vector representations obtained from this model will more accurately express semantic similarity. It is possible to fine-tune the BERT language model so that the vector representations obtained from this model more accurately express semantic similarity. We utilize a triplet objective function to train the network that reflects the semantic similarities and differences between concepts and entities. Suppose there is a mention of the entity m, the name of the corresponding (positive) concept c_g, the name of not corresponding (negative) concept c_n, the triplet objective function adjusts the neural network so that the distance between m and c_g is less than the distance between m and c_n for a given threshold. The following loss function is expressed as:

$$max(s(m, c_g) - s(m, c_n) + \epsilon, 0), \tag{5}$$

where ϵ is an offset that ensures that c_g is at least ϵ closer to m than to c_n. In our experiments, $\epsilon = 1$. The model is schematically illustrated in 2.

The selection of positive and negative examples is an essential component of the triplet loss function. Let us consider the generation procedure for each of them. Suppose we have a pair: an entity with its corresponding concept identifier and the dictionary. We limit the dictionary to concepts that correspond to the entity. The positive examples are generated from the limited part of the dictionary. The rest of the dictionary is used to generate the negative [34] examples. We investigated several strategies to select positive and negative examples

- **Random sampling**: positive and negative examples are randomly selected from the corresponding parts of the dictionary;
- **Hierarchy random sampling** (random sampling + n parents): Parent concept names are added to randomly generated positive samples. Negative examples are randomly generated.
- **Resampling** (resampling): in this case, the model trained on randomly sampled triplets is used to generate hard case examples. The generation process consists of several steps: (i) all entity mentions and concept names are encoded using the current model, (ii) as positive examples, we select

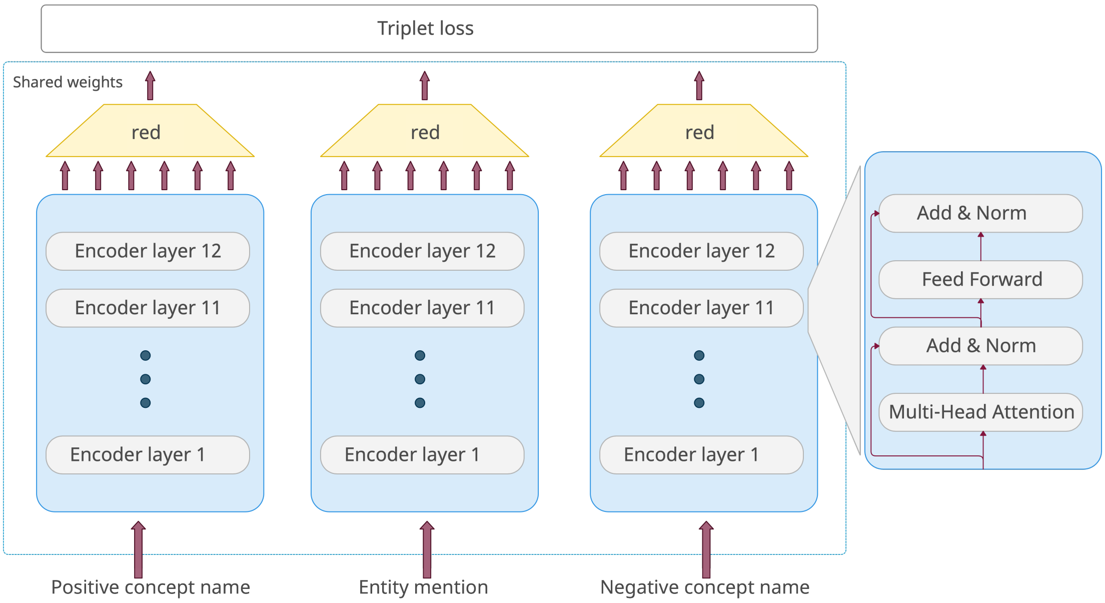

Figure 2: Architecture of the DILBERT. The model consists of three encoding networks with shared weights. The network on the left encodes the positive example, the network in the middle encodes the entity, and the network on the right encodes the negative example. All of the encoders are transformer based networks.

the closest to the entity concepts with the same concept identifier. (iii) negative examples are selected from the closest to the entity concepts that have a different concept identifier.

- **Resampling concerning hierarchy** (resampling + n siblings): positive and negative examples are generated similarly to the Resampling strategy. However, additionally, the names of concepts that have a common parent with the corresponding concept to the entity are selected as negative examples

The evaluation of the proposed method was carried out on the CDR corpus. The most commonly used quality measurement in the medical concept normalization task is accuracy. As shown in [35], the CDR corpus contains a large number of duplicates and a substantial overlap between train and test subsamples within the same corpus. To eliminate the influence of the model's ability to remember the training set and obtain more realistic quality metrics, we evaluated all models on the "refined" test set presented in [35]. The BioBERT v1.1 language model was used as the base model for comparison. The quality of the model was also compared with the state-of-the-art BioSyn [36] model. The results are shown in the table 4.

Table 4: Accuracy for the DILBERT model on CDR Disease and CDR Chemicals corporas.

Model	CDR Disease	CDR Chemical
BioBERT ranking	66.4	80.7
BioSyn	74.1	**83.8**
DILBERT, random sampling	75.5	81.4
DILBERT, random + 2 parents	75.0	81.2
DILBERT, random + 5 parents	73.5	81.4
DILBERT, resampling	**75.8**	83.3
DILBERT, resampling + 5 siblings	75.3	82.1

The key feature of the metric learning approach is the ability to detect the entities with no suitable concept in the vocabulary. The detection methodology naturally follows from the assumption that similar elements are close to each other in a latent space. Consequently, if all dictionary objects are far enough from the entity, then this is the out-of-vocabulary case. The notion of far enough is formalized via threshold value. Thus, if all concepts are at a distance greater than the threshold value t, we can conclude that none of them corresponds to the entity. To determine the threshold, the maximum distance of true-positive cases d_{tp} and the minimum distance of false-positive cases d_{fp} are used. The threshold value is set equal to the weighted sum:

$$t = a_1 * d_{tp} + a_2 * d_{fp}, \tag{6}$$

where a_1 is the proportion of true positive examples among entities whose closest concept is at a distance of $s \in [d_{fp}; d_{tp}]$; a_2 – proportion of false positives in the same entity set. If the given set of entities is empty, then the coefficients are set equal to $\frac{1}{2}$. The evaluation of the proposed approach of out-of-vocabulary cases detection was carried out on the corpus of clinical trials [5]. The results are shown in the 5 table.

Another key feature of the proposed approach is the independence from the terminology presented in the training set. A model obtained on one terminology can be applied to another without additional re-training. In particular, the 5 table shows the results for models trained on the CDR corpus, tied to the MEDIC and

Table 5: The metrics of the DILBERT model on the corpora of clinical trials. The results are presented for the disease (CT Condition) and drug (CT Intervention) entity types. Quality metrics are shown both for a subset consisting only of entities with a single concept (single concept) and for the entire corpus (full set)

Model	**CT Condition**		**CT Intervention**	
	single concept	full set	single concept	full set
BioBERT ranking	72.60	71.74	77.83	56.97
BioSyn	86.36	-	79.58	-
DILBERT with different sampling strategies				
random sampling	85.73	84.85	**82.54**	**81.16**
random + 2 parents	86.74	86.36	81.84	79.14
random + 5 parents	**87.12**	**86.74**	81.67	79.14
resampling	85.22	84.63	81.67	80.21
resampling + 5 siblings	84.84	84.26	80.62	76.16

CTD terminologies, and evaluated on the corpus with InSilico's in-house terminology and MeSH terminology. The model achieved 87.12% and 90.53% accuracy (accuracy) on the in-house and MeSH terminologies respectively.

As it can be seen from the 4 and 5 tables, the DILBERT model shows results on par with the BioSyn model on the CDR Chemical corpora and outperforms on the CDR Disease, CT Condition, CT Intervention corporas.

The studied models calculate the similarity between entities and concept names at the word and sub-word level. This feature allows to link entity that has a surface form similar to one of the concept names from the terminology, such as “visual defects” and “visual impairment”. However, these models may incorrectly associate entities with knowledge base concepts in two major cases: (i) the surface form of the entity and the concept name are similar but have different meanings (e.g., “chlorfenac” (C041190) and “chlorferon” (C305311)), (ii) both expressions have the same meaning but differ in the surface form (“methindol” and “indomethacin” are the same anti-inflammatory drug (D007213)). Knowledge bases as well may be outdated, and coverage of synonyms may be incomplete.

We carried out a series of experiments with incomplete vocabulary. We used models trained on the CDR Disease & Chemical corpora. These experiments show the model tendency to remember the concept names derived from the training vocabulary. To create incomplete dictionaries, we grouped the original

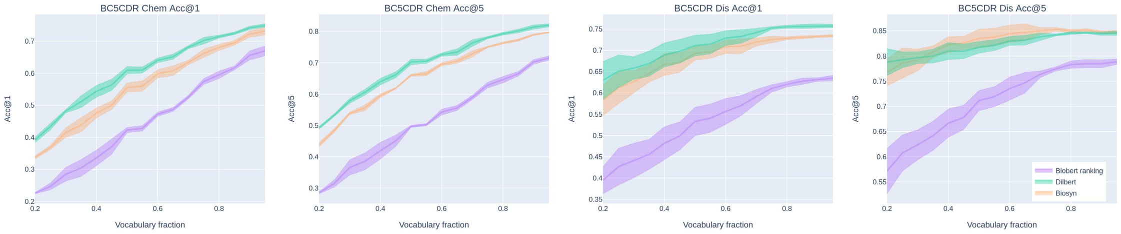

Figure 3: Evaluation of the impact of the dictionary completeness on the quality of the model.

versions of the dictionaries by concept identifier, in each group we randomly selected Vocabulary fraction × group size synonyms, where group size is the size of the group of the corresponding concept, and Vocabulary fraction is the proportion of synonyms remaining in the incomplete dictionary. Note that if the number of concept names after the selection turned out to be fractional, then the number of concept names was rounded to the smallest integer. For example, 95% of 10 concept names are 9. The fraction of synonyms varies from 0.95 to 0.20, with a step of -0.05. For each Vocabulary fraction value, the procedure of random generation of an incomplete dictionary was performed four times, then the results were averaged. According to the results presented in the figure 3, we can draw following conclusions.

First, the results for the Acc@1 metric show that the performance degradation from a complete dictionary to a dictionary containing 30% of synonyms is significant. This phenomenon is explained by the fact that, when assigning a concept, models rely more on the name that is close by surface form. However, because of the incompleteness of the dictionary, the required synonyms may be missed during prediction. We can also see that the effectiveness of drug normalization models has decreased more than disease normalization models, because drug names are very heterogeneous (there are names of active compounds, trademarks, proprietary identifiers, etc.). Second, from the point of view of Acc@5, the models show not as much significant drop in the quality metric. Finally, as we can see from the graphs, the DILBERT model shows the smallest drop in quality when using an incomplete dictionary, compared to other models.

It should be noted that the proposed approach allows caching and indexing of vector representations of concept names. Which in this case significantly speeds up the time spent on processing one entity. Thus, when using the FAISS library with GPU support for fast search in multidimensional space, it takes 3 hours to process 10 million entities, or about 1000 entities per second. Processing is carried out on a single Nvidia TITAN X video card.

3.3 Combined approach

As noted, classification approaches have a high accuracy in recognition of the concepts present in the training set, and cannot recognize the rest of the concepts. Approaches based on metric learning are an addition to the classification approaches. Since they allow one to recognize concepts that are outside the training sample. As a result, in the work of 2020 [7], the author proposed a method for combining the classification and metric learning approaches. In this case, normalization was carried out using the predictions of both predictors based on the threshold value t. Suppose that, according to the metric learning approach, c_m be the concept closest to the entity m, at a distance $S(m, c_k)$ - the distance is determined according to the formula 4. The concept c_c is the most probable within the classification model. Then the output of the combined model is determined by the following formula:

$$c_{cm} = \begin{cases} c_m, \text{if } s(m, c_k) < t \\ c_c, \text{otherwise} \end{cases} \tag{7}$$

According to the formula 7, the output of the combined approach will be the concept chosen by the metric approach if its synonym is at a distance not more than t from the entity. Otherwise, the output of the combined model will be the most likely concept according to the classification model. Thus, the preference is given to concepts obtained by the metric learning approach and having a closer synonyms, and then the most probable by the classification approach. The maximum distance in this case is a hyper-parameter and obtained on the validation data set.

The proposed method was evaluated at Social Media Mining for Health Applications (#SMM4H) Shared Tasks 2020 Task 3 and Social Media Mining for Health Applications (#SMM4H) Shared Tasks 2021 Task 1c. The evaluation of SMM4H 2020 and SMM4H 2019 addressed a task that included two subtasks: named entity recognition and medical concept normalization. To extract named entities an approach based on the EnDR-BERT [7] language model with a classification layer for each token according to the BIO scheme was used. The combined approach described above is used to solve normalization problems. As a language

Table 6: Quality metrics of the combined model on the SMM4H 2020 Task 3. The results of other approaches are taken from the article [9]

Model	**E F$_1$**	**N F$_1$**	**N P**	**N R**
KFU@NLP Team	**76.0**	**46.0**	**48.0**	**45.0**
BERT, CADEC, SMM4H'17 corpus	73.0	38.0	34.0	44.0
RoBERTa, multi-task learning	69.0	35.0	33.0	38.0
BERT ensemble, fastText-based similarity metrics, CADEC	58.0	22.0	24.0	20.0
BiLSTM, CRF, GloVe and EXT word embeddings, QuickUMLS	46.0	20.0	35.0	14.0
-	56.0	15.0	15.0	14.0
dictionary	16.0	0.0	0.0	0.0

Table 7: Quality metrics of the combined model on the SMM4H 2021 Task 1c. The results of other approaches are taken from the article [10]

Model	**E F$_1$**	**N F$_1$**	**N P**	**N R**
KFU@NLP Team	40.0	**29.0**	30.1	27.5
BERT with joint NER and Normalization	37.0	24.0	37.1	17.8
RoBERTa, multi-task learning	69.0	35.0	33.0	38.0
BERTweet and similarity measures	42.0	20.0	13.9	34.2
Multi-task learning with selective oversampling	51.0	16.0	16.0	17.0

model BERT, BioBERT, Scibert were studied. The best results for the normalization problem were shown by BERTs, outperforming the rest by 1-2% in terms of accuracy. On the SMM4H 2020 validation dataset, the proposed approach showed the following results: the F-measure of exact match reached 57.81% when extracting named entities, the accuracy when solving the normalization problem was 45.17%. The evaluation on the test data set was carried out by the authors of the problem in two ways: separately for entity recognition task and in for the entire problem. The described approach showed the best results in the SMM4H 2020 Task 3 and SMM4H Task 1c competitions. On the SMM4H 2020 test set for extracting named entities, the proposed approach showed the best results, reaching 75.5% F-measure of partial match. The average score among all teams was 56.4% F-measure. For complex evaluation, the described method showed a 46.3% F-measure on the test dataset, with an average of 29.2%. Detailed results on the test data set are given in the 7 table. On the SMM4H 2021 test set proposed model achieved a 40% F-score for an exact match in the named entity recognition task, a 29% F-score, a 30% accuracy, and a 27.5% recall for combined evaluation. It is worth noting that despite the low performance on the named entity recognition part in the SMM4H 2021, the proposed approach achieved the highest performance among all participants in the full system assessment. The obtained results testify the high quality of the named entity normalization component in comparison with the solutions of other participants.

4 Conclusion

The final section presents the main results of the work.

1. In the works [1; 2], the effectiveness of classification approach in the medical concept normalization task was studied. As well, the semantic similarity features were proposed, which have shown their performance boost in CLEF eHealth 2017 Task 1, SMM4H 2019 Task 3, SMM4H 2020 Task 3, SMM4H 2021 Task 1c. The papers further show the limitations of the classification approach and standard model quality assessment methods. In particular, the lack of training samples covering all possible medical concepts and significant overlaps between the test and training subsets of existing datasets lead to false quality measurements. A method for evaluating models is proposed that eliminates the disadvantage of high intersection.
2. A named entity linking model DILBERT, which is based on a metric learning approach, was proposed. The work [5] shows the sustainability of the model to vocabulary substitution and the ability to recognize concepts that were not present in the training set. The effectiveness of the model on the CDR Disease, CDR Chemical, CT Intervention, CT Condition corporas were shown. It is also shown that the DILBERT model is less prone to remember vocabulary.
3. Based on the classification and metric learning approaches, a combined model of named entity linking was proposed. The effectiveness of this method was shown in the SMM4H 2020 Task 3 and SMM4H 2021 Task 1c.

Bibliography

1. Medical concept normalization in social media posts with recurrent neural networks / Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, Valentin Malykh // *Journal of biomedical informatics.* — 2018. — Vol. 84. — Pp. 93–102.

2. *Miftahutdinov Zulfat, Tutubalina Elena.* Deep Neural Models for Medical Concept Normalization in User-Generated Texts // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. — 2019. — Pp. 393–399.

3. *Miftahutdinov Zulfat, Tutubalina Elena.* Deep learning for ICD coding: Looking for medical concepts in clinical documents in English and in French // International Conference of the Cross-Language Evaluation Forum for European Languages / Springer. — 2018. — Pp. 203–215.

4. *Miftahutdinov Z, Tutubalina E.* KFU at CLEF eHealth 2017 Task 1: ICD-10 coding of English death certificates with recurrent neural networks // CEUR Workshop Proceedings. — 2017.

5. Drug and Disease Interpretation Learning with Biomedical Entity Representation Transformer / Zulfat Miftahutdinov, Artur Kadurin, Roman Kudrin, Elena Tutubalina // *Proceedings of the 43rd European Conference on Information Retrieval.* — 2021.

6. Medical concept normalization in clinical trials with drug and disease representation learning / Zulfat Miftahutdinov, Artur Kadurin, Roman Kudrin, Elena Tutubalina // *Bioinformatics.* — 2021. — Vol. 37, no. 21. — Pp. 3856–3864.

7. *Miftahutdinov Zulfat, Sakhovskiy Andrey, Tutubalina Elena.* KFU NLP Team at SMM4H 2020 Tasks: Cross-lingual Transfer Learning with Pretrained Language Models for Drug Reactions // Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task. — 2020. — Pp. 51–56.

8. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019 / Davy Weissenbacher, Abeed Sarker, Arjun Magge et al. // Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task. — 2019. — Pp. 21–30.

9. Overview of the fifth social media mining for health applications (# smm4h) shared tasks at coling 2020 / Ari Klein, Ilseyar Alimova, Ivan Flores et al. // Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task. — 2020. — Pp. 27–36.

10. Overview of the sixth social media mining for health applications (# smm4h) shared tasks at naacl 2021 / Arjun Magge, Ari Klein, Antonio Miranda-Escalada et al. // Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task. — 2021. — Pp. 21–32.

11. *Canese Kathi, Weis Sarah.* PubMed: the bibliographic database // The NCBI Handbook [Internet]. 2nd edition. — National Center for Biotechnology Information (US), 2013.

12. *Leaman Robert, Islamaj Doğan Rezarta, Lu Zhiyong.* DNorm: disease name normalization with pairwise learning to rank // *Bioinformatics.* — 2013. — Vol. 29, no. 22. — Pp. 2909–2917.

13. *Leaman Robert, Lu Zhiyong.* TaggerOne: joint named entity recognition and normalization with semi-Markov Models // *Bioinformatics.* — 2016. — Vol. 32, no. 18. — Pp. 2839–2846.

14. *Limsopatham Nut, Collier Nigel.* Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — 2016. — Pp. 1014–1023.

15. Cadec: A corpus of adverse drug event annotations / Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, Chen Wang // *Journal of biomedical informatics.* — 2015. — Vol. 55. — Pp. 73–81.

16. The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications / Maryam Zolnoori, Kin Wah Fung, Timothy B Patrick et al. // *Data in brief.* — 2019. — Vol. 24. — P. 103838.

17. *Sarker Abeed, Gonzalez-Hernandez Graciela.* Overview of the second social media mining for health (SMM4H) shared tasks at AMIA 2017 // *Training.* — 2017. — Vol. 1, no. 10,822. — P. 1239.

18. *Alvaro Nestor, Miyao Yusuke, Collier Nigel.* TwiMed: Twitter and PubMed comparable corpus of drugs, diseases, symptoms, and their relations // *JMIR public health and surveillance.* — 2017. — Vol. 3, no. 2. — P. e24.

19. *Doğan Rezarta Islamaj, Leaman Robert, Lu Zhiyong.* NCBI disease corpus: a resource for disease name recognition and concept normalization // *Journal of biomedical informatics.* — 2014. — Vol. 47. — Pp. 1–10.

20. Deep Contextualized Word Representations / Matthew Peters, Mark Neumann, Mohit Iyyer et al. // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). — 2018. — Pp. 2227–2237.

21. *Miftahutdinov ZS, Tutubalina EV, Tropsha AE.* Identifying disease-related expressions in reviews using conditional random fields // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. — 2017. — Pp. 155–166.

22. Hierarchical attention networks for document classification / Zichao Yang, Diyi Yang, Chris Dyer et al. // Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. — 2016. — Pp. 1480–1489.

23. Multi-task Character-Level Attentional Networks for Medical Concept Normalization / Jinghao Niu, Yehui Yang, Siheng Zhang et al. // *Neural Processing Letters.* — 2018. — Pp. 1–18.

24. Team UKNLP: Detecting ADRs, classifying medication intake messages, and normalizing ADR mentions on twitter / S. Han, T. Tran, A. Rios, R. Kavuluru // *CEUR Workshop Proceedings.* — 2017. — Vol. 1996. — Pp. 49–53.

25. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task / Abeed Sarker, Maksim Belousov, Jasper Friedrichs et al. // *Journal of the American Medical Informatics Association.* — 2018. — Vol. 25, no. 10. — Pp. 1274–1283.

26. LITL at CLEF eHealth2017: automatic classication of death reports / Lydia-Mai Ho-Dac, Cécile Fabre, Anouk Birski et al. // CLEF eHealth 2017. — 2017.

27. CLEF eHealth 2017 Multilingual Information Extraction task Overview: ICD10 Coding of Death Certificates in English and French. / Aurélie Névéol, Aude Robert, Robert Anderson et al. // CLEF (Working Notes). — 2017.

28. *Cabot Chloé, Soualmia Lina Fatima, Darmoni Stéfan Jacques.* SIBM at CLEF eHealth Evaluation Lab 2017: Multilingual Information Extraction with CIM-IND. // CLEF (Working Notes). — 2017.

29. *Zweigenbaum Pierre, Lavergne Thomas.* Multiple Methods for Multi-class, Multi-label ICD-10 Coding of Multi-granularity, Multilingual Death Certificates. // CLEF (Working Notes). — 2017.

30. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation / Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre et al. // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). — 2014. — Pp. 1724–1734.

31. *Bodenreider Olivier.* The unified medical language system (UMLS): integrating biomedical terminology. — Vol. 32. — Oxford University Press, 2004. — Pp. D267–D270.

32. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring / Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, Jason Weston // *CoRR abs/1905.01969. External Links: Link Cited by.* — 2019. — Vol. 2. — Pp. 2–2.

33. *Reimers Nils, Gurevych Iryna.* Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). — 2019. — Pp. 3973–3983.

34. Distributed representations of words and phrases and their compositionality / Tomas Mikolov, Ilya Sutskever, Kai Chen et al. // Advances in neural information processing systems. — 2013. — Pp. 3111–3119.

35. *Tutubalina Elena, Kadurin Artur, Miftahutdinov Zulfat.* Fair Evaluation in Concept Normalization: a Large-scale Comparative Analysis for BERT-based Models // Proceedings of the 28th International Conference on Computational Linguistics. — 2020. — Pp. 6710–6716.

36. Biomedical Entity Representations with Synonym Marginalization / Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, Jaewoo Kang // ACL. — 2020.