Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2018”

Moscow, May 30—June 2, 2018

LEVERAGING DEEP NEURAL NETWORKS AND SEMANTIC SIMILARITY MEASURES FOR MEDICAL CONCEPT NORMALISATION IN USER REVIEWS

Miftahutdinov Z. (zulfatmi@gmail.com),
Tutubalina E. (elvtutubalina@kpfu.ru)

Kazan Federal University, Kazan, Russia

Nowadays a new yet powerful tool for drug repurposing and hypothesis generation emerged. Text mining of different domains like scientific libraries or social media has proven to be reliable in that application. One particular task in that area is medical concept normalization, i.e. mapping a disease mention to a concept in a controlled vocabulary, like Unified Medical Language System (UMLS). This task is challenging due to the differences in language of health care professionals and social media users. To bridge this gap, we developed end-to-end architectures based on bidirectional Long Short-Term Memory and Gated Recurrent Units. In addition, we combined an attention mechanism with our model. We have done an exploratory study on hyperparameters of proposed architectures and compared them with the effective baseline for classification based on convolutional neural networks. A qualitative examination of the mentions in user reviews dataset collected from popular online health information platforms as well as quantitative one both show improvements in the semantic representation of health-related expressions in user reviews about drugs.

Key words: medical concept mapping, medical concept normalization, deep learning, UMLS, recurrent neural networks, information extraction

1. Introduction

There were many novel applications of Natural Language Processing (NLP) to biomedical information in recent years. Most of researchers' attention attracts task of Named Entity Recognition (NER). Many applications of NER have been applied to scientific literature and electronic health records. And comparatively little work was carried out on social media texts of individuals undergoing medical treatment.

Social media in recent years had become a virtually inexhaustible source of people's opinions on the wide variety of topics. In this work, our focus is patients' opinions on drug effects, i.e. patients' reports. Progressive improvement of text mining approaches applied to patient reports in social media by the terms of accuracy and recall has multiplicative effect on several areas including pharmacovigilance (especially, for new drugs), drug repurposing, and understanding drug effects in the context of important and yet not well studied other factors such as concurrent use of other drugs, diet, and lifestyle.

We study the patients' comments on social media in an aspect of discovering disease-related medical concepts from. In the context of this problem, we map a text written in the informal language of social media (e.g. "I can't fall asleep all night" or "head spinning a little") to formal medical language (e.g. "insomnia" and "dizziness" respectively).

This goes beyond simple straightforward matching of natural language expressions with vocabulary elements: string matching approaches may not be able to link the social media language to the medical concepts due to few or an absence of overlapping words. We call the task of mapping everyday life language to medical terminology medical concept normalization (or medical concept mapping). The main benefit of solving this task is bridging the gap between the language of lay public and medical professionals.

The described task seems to be uneasy since patients post in social media texts on different illness concepts (a wide variety of one's from conditions like major depressive disorder to informal phrases describing specific symptoms such as "woke up too early" or "mucus building up in my lungs") and a wide diversity of drug reactions (e.g., "excessive sweating at night", "slept like a baby", or "clearing up an infection"). Also, we should mention that the data from social networks typically contain a lot of noise such as typos, misspellings, incorrect grammar, hashtags, abbreviations, and different variations of the same word.

Formally speaking, this task is related to several well-known NLP challenges including paraphrase detection, word sense disambiguation, and entity linking where an entity mention is mapped to a unique concept in an ontology after solving the disambiguation problem [1, 2]. In recent studies, there were proposed some approaches to this challenge treating the task of linking a one- or multi-word expression to a knowledge base as a supervised sequence labeling problem. Miftahutdinov and Tutubalina [3] proposed an encoder-decoder model based on bidirectional recurrent neural networks (RNNs) to translate a sequence of words from a death certificate into a sequence of medical codes. Two recent works present similar approaches [4, 5] that utilize RNNs for normalization of tweets' phrases at the AMIA 2017 Social Media Mining for Health Applications workshop, while Limsopatham and Collier [6] experimented with convolutional neural networks (CNNs) on social media data. These works demonstrated usage of deep learning techniques for medical concept normalization. In this paper, we experimented with more complex RNN architectures with

an attention mechanism and additional linguistic knowledge. Moreover, we study the impact of different word embeddings. We conduct extensive experiments on a real-life dataset from Askapatient.com and demonstrate the effectiveness of the proposed method for medical concept mapping.

2. Background

The most popular knowledge-based system for mapping texts from scientific literature and clinical records to medical identifiers are MetaMap [7] and DNorm [8]. MetaMap was developed by the National Library of Medicine (NLM) in 2001 and has become a de-facto baseline method for many recent studies. This system is based on UMLS and a linguistic approach using lexical lookup and variants by associating a score with phrases in a sentence. Leaman et al. introduced a DNorm system for assigning disease mentions from PubMed abstracts a unique identifier from a MEDIC vocabulary, which combines terminology from Medical Subject Headings (MeSH) and Online Mendelian Inheritance in Man (OMIM) [8]. DNorm consists of a text processing pipeline, including the named entity recognizer to locate diseases in the text, and a normalisation method. The normalisation method is based on a pairwise learning-to-rank technique using the tokens from all mentions as features. DNorm outperformed MetaMap as the baseline.

While there has been a lot of work on named entity recognition from social media posts that has been done over the past 7 years [5, 9, 10, 11, 12, 13, 14, 15, 16], relatively few researchers have looked at assigning social media phrases to medical identifiers. First Social Media Mining shared task workshop (organized as part of the Pacific Symp. on Biocomputing 2016) was designed to mining pharmacological and medical information from social media, with a competition based on a published dataset [13]. Task 3 is devoted to medical concept normalisation, where participants were required to identify the UMLS concept for a given ADR. The evaluation set consisted of 476 ADR instances. Sarker et al. [13] noted that there had been no prior work on normalisation of concepts expressed in social media texts, and task 3 did not attract much attention from the researchers.

Recently, two teams namely UKNLP [4] and gnTeam [5] participated in the Second Social Media Mining for Health (SMM4H) Shared Task and submitted their systems for automatic normalisation of ADR mentions to MedDRA concepts. For the task 3, Sarker et al. [17] created a new dataset of tweets' phrases. The training set for this task contains 6,650 phrases mapped to 472 concepts, while the testing set consisted of 2,500 phrases mapped to 254 classes. We also note that organizers of this task did not describe the corpus creation in details as well as not providing corpus statistics, e.g., the overlap percentage between training and testing sets. Teams' systems showed similar results. The gnTeam's approach contained three components for preprocessing and classification. The first two components corrected spelling mistakes and converted sentences into vector-space representation, respectively. For the third step, GnTeam adopted multinomial logistic regression model which achieved the accuracy of 0.877, while the bidirectional GRU achieved the accuracy of 0.855. As input, the network adopted the GoogleNews embeddings trained on a Google News corpus

due to higher results the highest performance over embeddings trained on tweets. The ensemble of both classifiers showed slightly better performance and achieved the accuracy of 0.885. The UKNLP's system adopted hierarchical LSTM in which a phrase is segmented into words and each word is segmented into characters. Word embeddings were trained on a Twitter corpus. Hierarchical Char-LSTM achieved the accuracy of 0.872, while hierarchical Char-CNN performed slightly better and achieved the accuracy of 0.877. We note this corpus of tweets for future work since the official test data is available for the shared task participants only by the time of publication.

Recently, Limsopatham and Collier [6] experimented with Convolutional Neural Networks (CNN) and pre-trained word embeddings for mapping social media texts to medical concepts. For evaluation, three different datasets were used. The authors created two datasets with 201 and 1,436 Twitter phrases which mapped to concepts from a SIDER database. The third dataset is the CSIRO Adverse Drug Event Corpus (CADEC) [2] which consists of user reviews from askapatient.com. The authors observed that training can be effectively achieved at 40–70 epochs. As input, the network concatenated embeddings of words. The GoogleNews embeddings improved results significantly over embeddings on medical articles. Experiments showed that CNN (accuracy 81%) outperformed DNorm (accuracy 73%), RNN (accuracy 80%) and a multi-class logistic regression (accuracy 77%) on the AskAPatient corpus (as well as corpora of tweets). This work is the closest to ours in the use of deep learning technology and semantic representation of words. However, we found that only approximately 40% of expressions in the test data are unique, while the rest of expressions occur in the training data. Therefore, the presented accuracy may be too optimistic. We believe that future research should focus on developing extrinsic test sets for medical concept normalisation.

3. Methods

In this section, we will discuss major challenges in this task and applied neural architectures.

3.1. Recognition of Different Word Variances

The task of medical concept normalisation is closely related to the problem of word sense disambiguation and terminological variance. There are major challenges which disease mention recognition methods as well as term extraction methods face:

(i) exical, morphological, and syntactic variants;
(ii) paraphrases, synonyms;
(iii) abbreviations;
(iv) ambiguity;
(v) misspellings.

The examples of three-form phrases from the CADEC corpus are presented in Table 1.

Table 1. Examples of three-form phrases with corresponding medical concepts

Free-form Phrases	Medical concept	SNOMED ID
lower pelvic pain	Pain in pelvis	30473006
uterus contractions	Uterine spasm	29542008
something wrong with my uterus	Uterus problem	289621007
stomach issues	Stomach problem	300306001
slightly heavier menstrual cycle	Menorrhagia	386692008
inflammation in my back muscles	Muscle cramp	55300003
inflammation in my neck	Cervical arthritis	387801000
heavy menstrual bleeding	Menorrhagia	386692008
acidic bile in my mouth	Acid reflux	698065002
could only walk less than 100 meters	Reduced mobility	8510008
very painful joints	Arthralgia	57676002
starting to upset my stomach	Stomach ache	271681002
can't sleep	Insomnia	193462001
high BP	Increased venous pressure	69791001
pulse is still extremely high	Pulse fast	86651002

3.2. Proposed Model for Concept Mapping

We propose a deep approach for mapping entity mentions to medical codes. We first convert each mention into a semantic representative vector using bidirectional LSTM or GRU [18–23] with attention mechanism on top of the embedding layer. We use the hyperbolic tangent as activation function. Then, a set of features are extracted using the cosine similarity between mentions and medical concepts from the UMLS Metathesaurus. For model training, we use the cross-entropy error between gold distribution and predicted distribution as the loss function. The model is depicted in Fig. 1.

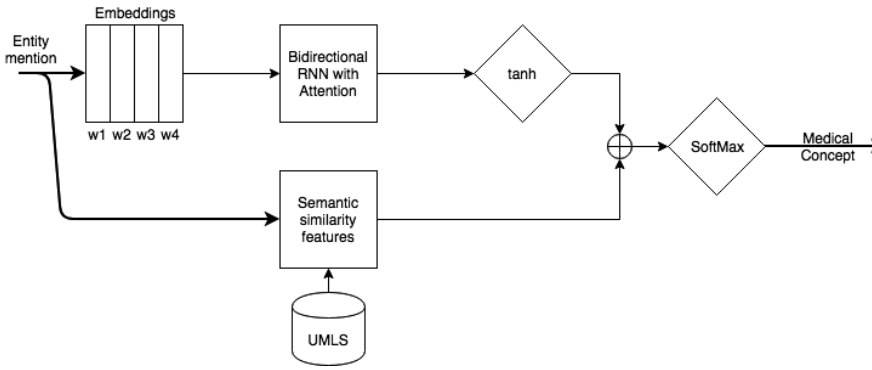

Figure 1: Proposed architecture for medical concept normalization

3.3. Semantic Similarity Features

We extract a set of features to enhance the representation of the phrases. These features consist of cosine similarity between the vectors of the input phrase and a concept in a medical terminology dictionary. This dictionary includes medical codes and synonyms from the UMLS Metathesaurus (version 2017 AA), where codes are presented in the CADEC corpus. We apply the following strategy to create representations of a concept and a mention and compute cosine similarity between the representations of each pair: present a medical code as a single document by concatenating synonymous terms. Then, we apply the TF-IDF transformation on the code and the entity mention and compute the cosine similarity.

Neural networks require word representations as inputs. We investigate the use of several different pre-trained word embeddings. Recent advances have made *distributed word representations* into a method of choice for modern NLP [24, 25, 26]. We utilize word embeddings named *HealthVec*, which are publicly available 200-dimensional embeddings that were trained on 2,607,505 unlabeled user comments (93,526 terms) from health information websites using the CBOW model in [14]. We also experimented with another published 200-dimensional embeddings named *PubMedVec* (2,351,706 terms) trained on biomedical literature indexed in PubMed [27].

4. Experimental Evaluation

The purpose of our evaluation is to determine how well recurrent neural networks can identify the corresponding medical concepts based on informal language from patients' texts.

4.1. Data Set

We conducted experiments on a collection of user reviews obtained from the CADEC corpus [2]. This corpus contains 1,250 reviews and consists of four predefined disease-related types: ADR (6,318 entities), Disease (283 entities), Symptom (275 entities), and Clinical Finding (435 entities). Authors reported that only 39.4% of the annotations (including drugs) were unique; people generally discussed similar reactions. Disease and Symptom specify the reason for taking the drug. Patients may mention the name of a disease or the symptoms that led to them taking a drug. Findings are any adverse side effects, diseases, or symptoms that were not directly experienced by the reporting patient. We did not distinguish between these types and join them into one class of annotations named *Disease*.

All entities in the CADEC corpus were mapped to SNOMED CT-AU (SCT-AU) by a clinical terminologist. SNOMED CT is a clinical terminology that provides codes, synonyms, and definitions of clinical terms, and can be accessed through the UMLS Metathesaurus. Additionally, concepts identified in the SNOMED CT were associated with MedDRA identifiers. In this work, we adopted only SNOMED CT identifiers and removed 'concept less' or ambiguous mentions for evaluation purpose. Table 2 shows final statistics for the CADEC corpus. The total number of unique codes was 1,029.

Table 2: Statistics of the dataset used in the experiments

Entity type	Total	Unique phrases	Unique SNOMED codes
ADR	5,838	3,241	788
Disease	266	165	108
Drug	1,657	290	124
Finding	399	270	180
Symptom	251	128	78

4.2. Preprocessing and Experiment Settings

Preprocessing includes spelling correction and lemmatization using the Natural Language Toolkit (NLTK). We performed a 5-fold cross-validation to evaluate the methods. We found that a standard cross-validation method creates a high overlap of expressions in an exact matching between training and testing parts. Therefore, the split procedure has a specific feature in our setup. First, we removed all duplicates in each dataset. Second, we grouped medical records into sets which are related to a specific medical code. Every such set was split independently into k folds, and all these folds were merged into final k folds. The created folds are publicly available[1].

4.3. Baseline System

For comparison, we applied state-of-the-art baselines based on convolutional neural networks. In [6], experiments showed that CNN outperformed existing strong baselines such as DNorm and Logistic Regression. In order to obtain local features from a text with CNNs, we used multiple filters of different lengths [28].

4.4. Model Configuration and Training

Since neural networks, especially deep neural networks, have a very large number of free parameters, problems with overfitting are inevitable, and some form of regularization is required. We used a dropout rate [29] of 0.5 after the embedding layer (before networks' layers).

Another standard technique in modern deep learning, batch normalisation [30], was designed to cope with a problem known as covariate shift. For all networks, we set the mini-batch size to 128 to minimize the negative log-likelihood of correct predictions.

The last important set of advances deal with actually training the model. We used a popular adaptive gradient descent variations, Adam [31]. Embedding layers are trainable for all networks. The number of outputs of the layer with the softmax activation equals to the number of unique concept codes. Additionally, we separated out 10% of the training set to form the validation set which was used to evaluate different model parameters. The number of epochs is determined by early stopping on the

[1] https://yadi.sk/d/oLBTUpXg3RtCzd

validation set. We employed early stopping after two epochs with no improvement on the validation set. The final number of epochs is 15.

For RNN, we utilized either a 100- or 200-dimensional hidden layer for each RNN chain. For CNN, we adopted effective parameters from [28, 6]. We used the filter w with the window size h of [3, 4, 5], each of which had 100 feature maps. Pooled features were fed to a fully connected feed-forward neural network (with dimension 100) to make an inference, using rectified linear units as output activation.

We found 91% and 88% of words from the CADEC corpus vocabulary in the word embeddings HealthVec and PubMedVec, respectively. For other words, their representations were uniformly sampled from the range of embedding weights [32].

4.5. Results

The standard technique for evaluating concept normalisation is to compare correctly normalised disorder mentions against the gold standard entities [7, 33]. Accuracy which is defined as follows:

$$Accuracy = \frac{N_{correct}}{T_g}, \quad (1)$$

where $N_{correct}$ is the number of correctly normalised disorder mentions and T_g is the total number of disorder mentions in the gold standard.We present the experimental results of neural networks in Table 3. The attention-based GRU with UMLS-based features achieved an accuracy of 69.92%.

Table 3: The accuracy performance of neural networks

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	HealthVec, 100 hidden units	64.83
GRU+Attention	HealthVec, 200 hidden units	67.08
with semantic similarity features		
LSTM+Attention	HealthVec, 100 units, similarity TF-IDF	67.63
LSTM+Attention	HealthVec, 200 units, similarity TF-IDF	66.83
GRU+Attention	HealthVec, 100 units, similarity TF-IDF (ALL)	69.92
GRU+Attention	HealthVec, 200 units, similarity TF-IDF (ALL)	69.42

The best results were obtained while using vectors trained on social media posts. GRU consistently outperformed CNNs and LSTM in terms of accuracy. Attention mechanism and prior knowledge from the UMLS Metathesaurus indeed led to quality improvements for both GRU and LSTM.

5. Conclusion

In this work, we have demonstrated that RNN-based architectures, LSTM- and GRU-based in particular, have promising performance on the task of medical concept normalization of free text mentions in social media. The experiments have shown qualitative and quantitative improvement over a strong baseline. We see three possible ways to next research to improve and expand the achieved results. The natural way to extend our models is to integrate a linguistic knowledge into them. We plan to concatenate RNN's output with a semantic similarity vector. We might focus on the development of extrinsic test sets for medical concept normalization. This future work looks promising also in consideration of paraphrase generation and other encoder-decoder applicable tasks.

Acknowledgements

This work was supported by the Russian Science Foundation grant no. 18-11-00284. The authors thank Valentin Malykh for useful discussions during writing this paper.

References

1. *Pradhan S., Elhadad N., Chapman W. W., Manandhar S., Savova G.* (2014), Semeval-2014 task 7: Analysis of clinical text, SemEval COLING, pp. 54–62.
2. *Karimi S., Metke-Jimenez A., Kemp M., Wang C.* (2015), Cadec: A corpus of adverse drug event annotations, Journal of biomedical informatics, Vol. 55, pp. 73–81.
3. *Miftakhutdinov Z., Tutubalina E.* (2017), Kfu at clef ehealth 2017 task 1: Icd-10 coding of english death certificates with recurrent neural networks. CLEF, 2017.
4. *Han S., Tran T., Rios A., Kavuluru R.* (2015), Team uknlp: Detecting adrs, classifying medication intake messages, and normalizing adr mentions on twitter, CEUR Workshop Proceedings, Vol. 1996, pp. 49–53.
5. *Belousov M., Dixon W., and Nenadic G.* (2017), Using an ensemble of generalised linear and deep learning models in the smm4h 2017 medical concept normalisation task, CEUR Workshop Proceedings, Vol. 1996, pp. 54–58.
6. *Limsopatham N., Collier N.* (2016), Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation, ACL.
7. *Aronson A.* (2001), Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, Proceedings of the AMIA Symposium, p. 17.
8. *Leaman R., Doğan R. I., Lu Z.* (2013), DNorm: disease name normalisation with pairwise learning to rank, Bioinformatics, 29(22), pp. 2909–2917.
9. *Leaman R., Wojtulewicz L., Sullivan R., Skariah A., Yang J., Gonzalez G.* (2010), Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts to health-related social networks, Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, BioNLP'10, pp. 117–125.
10. *Nikfarjam A., Sarker A., O'Connor K., Ginn R., Gonzalez G.* (2015), Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features, Journal of the American Medical Informatics Association, page 41.

11. *Oronoz M., Gojenola K., Pérez A., Díaz de Ilarraza A., Casillas A.* (2015), On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions, Journal of biomedical informatics, Vol. 56, pp.318–332.
12. *Korkontzelos I., Nikfarjam A., Shardlow M., Sarker A., Ananiadou S., Gonzalez G. H.* (2016), Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts, Journal of biomedical informatics, Vol. 62, pp. 148–158.
13. *Sarker A., Nikfarjam A., Gonzalez G.* (2016), Social media mining shared task workshop, Proc. Pacific Symposium on Biocomputing, pp. 581–592.
14. *Miftahutdinov Z., Tutubalina E., Tropsha A.* Identifying disease-related expressions in reviews using conditional random fields, Proceedings of International Conference on Computational Linguistics and Intellectual Technologies Dialog, Vol. 1, pp. 155–167.
15. *Tutubalina E., Nikolenko S.* (2017), Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews, Journal of Healthcare Engineering, 2017.
16. *VanDam C., Kanthawala S., Pratt W., Chai J., Huh J.* (2017), Detecting clinically related content in online patient posts, Journal of Biomedical Informatics.
17. *Sarker A., Gonzalez-Hernandez G.* (2017), Overview of the second social media mining for health (smm4h) shared tasks at amia 2017, CEUR Workshop Proceedings, pp. 43–48.
18. *Goodfellow I., Bengio Y., Courville A.* (2016), Deep Learning, MIT Press.
19. *Bengio Y., Courville A., and Vincent P.* (2013), Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell., 35(8), pp. 1798–1828.
20. *Bengio Y., Simard P., and Frasconi P.* (1994), Learning long-term dependencies with gradient descent is difficult, IEEE transactions on neural networks, 5(2), pp. 157–166.
21. *Greff K., Kumar Srivastava R., Koutník J. R., Steunebrink B., Schmidhuber J.* (2015), LSTM: A search space odyssey, CoRR.
22. *Cho K., Van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y.* (2014), Learning phrase representations using RNN encoder-decoder for statistical machine translation, arXiv preprint arXiv:1406.1078.
23. *Schuster M., Paliwal K. K.* (1997), Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, Vol. 45(11), pp. 2673–2681.
24. *Goldberg Y.* (2015), A primer on neural network models for natural language processing, CoRR, 2015.
25. *Rubenstein H., Goodenough J. B.* (1965), Contextual correlates of synonymy, Commun. ACM, 8(10), pp. 627–633.
26. *Mikolov T., Chen K., Corrado G., Dean J.* Efficient estimation of word representations in vector space, CoRR, abs/1301.3781.
27. *Moen S., Salakoski T., Ananiadou S.* (2013), Distributional semantics resources for biomedical text processing.
28. *Kim Y.* (2014), Convolutional neural networks for sentence classification, arXiv preprint arXiv:1408.5882.

29. *Srivastava N., Hinton G. E., Krizhevsky A., Sutskever I., Salakhutdinov R.* (2014), Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, Vol. 15(1), pp. 1929–1958.
30. *Ioffe S., Szegedy C.* (2015), Batch normalisation: Accelerating deep network training by reducing internal covariate shift, International Conference on Machine Learning, pp. 448–456.
31. *Kingma D. P., Ba J.* (2014), Adam: A method for stochastic optimization, CoRR, abs/1412.6980.
32. *He K., Zhang X., Ren S., Sun J.* (2015), Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, Proceedings of the IEEE international conference on computer vision, pp. 1026–1034.
33. *Suominen H., Salanterä S., Velupillai S., Chapman W. W., Savova G., Elhadad N., Pradhan S., South B. R., Mowery D. L., Jones G. J., et al.* (2013), Overview of the share/clef ehealth evaluation lab 2013, International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 212–231.