Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2019"

Moscow, May 29—June 1, 2019

LANGUAGE MODELS FOR UNSUPERVISED ACQUISITION OF MEDICAL KNOWLEDGE FROM NATURAL LANGUAGE TEXTS: APPLICATION FOR DIAGNOSIS PREDICTION

Tarasov D. (dtarasov3@gmail.com),
Matveeva T., **Galiullina N.**

Meanotek, Kazan, Russia

Following recent success of neural language models in various downstream language understanding tasks, including common sense reasoning, we investigate possible utility of such models in domain specific reasoning task—proposing of preliminary diagnosis based on patient complains, presented as natural language text. We demonstrate that language model, trained on the texts collected from online medical forums posses significant accuracy in this task (73% at top 10 suggestions), when evaluated on dataset, constructed from clinical case reports, published in specialized medical journals. While preliminary, these findings indicate a possible new method that can be used to augment online symptoms checkers and clinical decision support systems.

Keywords: symptom checkers, neural language model, medical diagnosis

МОДЕЛИ ЯЗЫКА ДЛЯ ИЗВЛЕЧЕНИЯ МЕДИЦИНСКИЙ ЗНАНИЙ ИЗ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ С ЦЕЛЬЮ ПРИМЕНЕНИЯ В ЗАДАЧАХ ПРЕДСКАЗАНИЯ ДИАГНОЗА

Тарасов Д. (dtarasov3@gmail.com),
Матвеева Т., **Галиуллина Н.**

ООО «Меанотек», Казань, Россия

В связи с недавним успехом нейронных моделей языка в решении различных задач понимания естественного языка, включая рассуждения на основе здравого смысла, мы исследуем возможную полезность таких моделей в задаче рассуждений, специфичных для предметной области—предварительной медицинской диагностики на основе жалоб пациента, представленных в виде текста на естественном языке. Мы демонстрируем, что языковая модель, обученная на текстах, собранных на медицинских форумах в Интернете, обладает значительной точностью в выполнении этой задачи (73 % из 10 лучших предположений) при оценке по набору данных, построенному на основе отчетов о клинических случаях, опубликованных в специализированных медицинских журналах. Эти результаты указывают на возможный новый метод, который можно использовать для расширения возможностей онлайн-проверки симптомов и систем поддержки принятия клинических решений.

Ключевые слова: нейронная модель языка, медицинская диагностика, системы поддержки принятия клинических решений

1. Introduction

In this paper we consider the task of proposing preliminary diagnosis based on patient complaints, presented as text in natural language (Russian). Such systems are useful for their users, because they can give them better understanding on possible causes of their symptoms, as well as provide advice on the most appropriate point of care. Currently, English-based symptom checkers have insufficient accuracy. According to recent studies [Semigran et al, 2015], most systems provide correct diagnosis first in 34% of cases, while average accuracy in top-20 (correct diagnosis within the top 20 diagnoses given) is 58%.

While new tests were published [Razzaki et al, 2018] recently for Babylon Triage and Diagnostic System, claiming near-human expert accuracy, these results are still below that of best human doctors and concerns were raised about validity of published results, due to methodological flaws [Fraser et al, 2018].

In recent years, approaches, based on deep learning gained popularity in the field of medical diagnostics. While majority of applications of deep learning are in the area of medical image processing tasks, end-to-end language processing diagnostic approaches were also proposed. For example Deep Patient [Miotto et al, 2016] system uses auto-encoders to train representation of patient medical history and then predicts probability for a given patient to develop a new disease.

However, deep learning based models require huge amounts of training data to reach good performance, and for the task, considered in this paper, such datasets are very difficult to collect. Ideally, one want a dataset of 100,000 or more patient complaints matched with clinically confirmed diagnosis. Such dataset is hard to create because of privacy issues and even if these can be solved by anonimization, hospitals rarely have full descriptions of patient conditions in their own words (as opposed to description made by physician).

It was suggested that big language models can learn useful knowledge from text in completely unsupervised manner. In particular, [Trinh et al, 2018] demonstrated that language models can obtain higher accuracy on Winograd Schema Challenge [Levesque et al, 2012] then competing methods. It is tempting to assume that such models can also learn domain specific knowledge.

In this paper we propose a method that learns diagnosis classification task using data found on online medical forums, where people discuss their health problems. Such data is available in abundance in Russian language and in many other languages. Based on this data, this paper makes following contributions:

1. We develop term extraction model, that extracts mentions of diagnoses and symptoms from online forum postings in Russian.
2. Using this model we construct Knowledge base (KB) based on co-occurrence of diagnosis and symptoms in online forums
3. We train large language model on medical forums text collection
4. We compare co-occurrence based and new proposed language-model based approaches to diagnosis prediction. Our findings indicate that language model-based diagnosis prediction is superior to co-occurence baseline.
5. We demonstrate that diagnosis prediction from patient complaints, presented as text in natural language can have accuracy comparable to current symptom checkers based on series of multiple-choice questions

2. Related work

Symptom checkers usually use large hand-crafted knowledge bases (KB) of medical facts, than need constant revision [Blum, et al, 1991]. Semi-automatic methods for construction of such KBs were proposed [Ramnarayan et al, 2016], [Middleton et al, 2016], that are based on co-occurrence statistics in PubMed, Wikipedia, and Electronical Medical Records data. Drawbacks of these methods are requirements for complex NLP pipelines to interpret data, and limited ability of co-occurrence statistics to capture disease-symptom relationship, such as timing of symptom appearance. During their operation, most symptom checkers use multiple choice lists to collect symptoms, or allow to enter symptoms in natural language, but only one symptom in time [Kafle et al, 2018]. And even with that restriction, to achieve free-form symptoms interpretation capability, complex paraphrase generation model are employed to generate large number of paraphrases and then do look-up.

Such restrictions in data representation limit the extent to which symptom checker can leverage information that user provides about specific circumstances of the user and about disease development process. As a result, existing approaches only achieve accuracy around 59% at top 10 diagnosis suggestions [Kafle et al, 2018] and can not work with free-form descriptions of users conditions, that are typically found in online forums.

Few authors attempted to use deep learning for solving natural language based disease diagnosis problem. Current research is mostly focused on clinical question answering [Hasan et al, 2016] and medical data mining [Barnickel et al, 2009]; [Mallory et al, 2015].

3. Methods and algorithms

3.1. Datasets and data annotation

3.1.1. Online posting forums data

We used internal dataset of forum posts, collected from Russian online medical forums during September 2018. Dataset contains descriptions of medical complaints, discussions and advices from doctors on variety of medical topics. Total number of forum posts in the dataset is 30,756 (68 MB of UTF-8 encoded text).

We annotated mentions of diagnoses, symptoms and body parts in first 200 posts (60,000 words) from this dataset, and made separate training and test set (45,000 words training and 15,000 words test).

3.1.2. Diagnosis prediction test set

We collected 50 case reports published in Russian medical journals, for 50 different medical conditions randomly selected from list of diagnoses found in online forums. Each case report has final diagnosis verified by careful medical evaluation. For dataset construction, for each case report, patient complaints, described in that case report were rewritten using informal language by the person without special medical knowledge (as if written from patient perspective). In this way, we are trying to avoid systematic bias of having test set prepared by doctors (rather then lay persons for whom system is intended), which was one of important methodological issues with previous evaluation methodologies [Fraser et al, 2018]. In the same way, we avoid possible bias that can be introduced by having fictional cases that generally fit diagnosis criteria, but not based on real data, because we use data from real cases with clinically confirmed diagnosis. While larger test set is desirable, it is very labor-intensive to construct, and similar studies has used dataset of comparable size before [Semigran, 2015]; [Kafle et al, 2018], thus we consider it to be acceptable for preliminary studies.

3.2. Language model

Language modeling (LM) is one of the important tasks of natural language processing. The task involves predicting the (n+1)th token in a sequence given the n preceding tokens, where tokens can be words, subwords or characters. More formally, the goal of a language model is to estimate a distribution $P(x_{0:T})$ over sequences of tokens (x_0, x_1, ..., x_T).

The joint distribution over long text spans can then be represented as a product of the predictive distribution over tokens conditioned on the preceding tokens:

$$P(x_{0:T}) = \prod_{t=0}^{T} P(x_t \vee x_{0:t-1}) \tag{1}$$

Neural language models [Sutskever et al, 2011] usually use recurrent neural networks (RNN) for sequence modeling. Given a sequence of vectors {$x(t)$}, where t=$1..T$, an RNN computes memory and output sequences:

$$h(t) = f(Wx(t) + Vh(t{-}1) + b) \tag{2}$$

$$y(t) = g(Uh(t) + c) \tag{3}$$

where f is a nonlinear function, such as the sigmoid or hyperbolic tangent function and g is the output function. W and V are weight matrices between the input and hidden layer, and between the hidden units. U is the output weight matrix, b and c are bias vectors connected to hidden and output units. $h(0)$ in equation (1) can be set to constant value that is chosen arbitrary or trained by backpropagation.

Recently, it was shown that by learning to predict the next character given previous characters, neural network based language models can learn internal representations that capture syntactic and semantic properties [Radford et al, 2017].

We use Long Short Term Memory (LSTM) [Hochreiter et al, 1997] based neural network. The structure of the LSTM [9] allows it to train on problems with long term dependencies. In LSTM simple activation function f from above is replaced with composite LSTM activation function. Each LSTM hidden unit is augmented with a state variable $s(t)$ The hidden layer activations correspond to the 'memory cells' scaled by the activations of the 'output gates' o and computed in following way:

$$h(t) = o(t) \times f(c(t)) \tag{4}$$

$$c(t) = d(t) \times (c(t{-}1) + i(t)) \times f(Wx(t) + Vh(t{-}1) + b) \tag{5}$$

where × denotes element-wise multiplication, $d(t)$ is dynamic activation function that scales state by "forget gate" and $i(t)$ is activation of input gate.

We train LSTM-based character level language model with 3 hidden layers, with 3,192 LSTM cells per each layer. Given that it is hard to capture all basic language structure with relatively small dataset, we pre-trained our model on subset of Russian Wikipedia, containing 2 billion characters. Model was trained by using truncated backpropogation thought time with learning rate controlled by Adam [Kingma et al, 2014] algorithm. We halted training by tracking the performance on the validation set, stopping when negligible gains were observed. Then, we use trained weights to initialize new model, that was trained on medical forums texts.

3.3. Diagnoses and symptoms extraction

We trained the sequence tagging model using mini-batch gradient descent with one sentence per mini-batch. We used simple learning rate annealing method in which we multiple the learning rate by 0.85 if test loss does not fall for 2 consecutive epochs. By performing model selection on separate development set, we found optimal number of hidden units per layer of the LSTM to be 128, and the number of LSTM layers to be 2.

We used two different sets of input features—word embeddings trained over forum texts using word2vec algorithm [Mikolov et al, 2013] and activations of last layer of LSTM language model.

3.4. Co-occurrence based diagnosis prediction

We first extracted all diagnoses from forum data using trained extraction model. We then manually assigned ICD-10 (International Statistical Classification of Diseases and Related Health Problems, revision 10) codes to each unique term extracted. We then took top 200 most frequent diagnoses and calculated co-occurrence table with symptoms, where symptoms were listed as mentioned in the text, without normalization.

To predict diagnosis for a new text, we first extract all mentioned symptoms and then find top 40 most similar entries in symptoms/diagnoses co-occurrence table, using cosine similarity between symptoms embeddings, obtained by summing embeddings of each word for a given symptom. Each diagnosis in the list was scored according to the number of matched symptoms and extend to which individual symptoms were similar. We use this method as baseline to compare against language model based method.

3.5. Language model based diagnosis prediction

Following previously proposed approaches for common-sense reasoning [Trinh et al, 2018], we concatenate full description of person's condition with diagnosis and compute joint probability of resulting text using trained language model.

3.6. Evaluation metrics

For measuring quality of term extraction models we use F-measure, computed using *Proportional Overlap*—a metric that imparts a partial correctness, proportional to the overlapping amount, for each match [Irsoy and Cardie, 2014].

For measure of diagnostic accuracy, our main outcomes were whether the system listed the correct diagnosis first or within the first 10 of potential diagnoses.This metric sometimes defined as *diagnosis recall at top N* [Middleton et al, 2016], while other authors use the term "*diagnostic accuracy at top N*" [Semigran et al, 2015], [Kafle et al, 2018]. We will use the term diagnostic accuracy here. The choice of metric is dictated by the need to compare our results to others and the fact that text descriptions alone do not provide enough information to exclude all possible causes but one, so we are interested to measure the ability of system to successfully narrow list of possible causes to a few possibilities.

4. Results and discussion

4.1. Extraction of diagnoses and symptoms mentions

After training terms extraction model, we obtained results, presented in **Table 1**.

Table 1. Term extraction accuracy

Input features type	F1, diagnosis	F1, symptom
Word2vec trained on forum data	0.55	0.65
Word2vec trained on Wikipedia	0.51	0.58
Activations of language model top layer (Wikipedia)	0.50	0.59
Activations of language model top layer (pre-trained on Wikipedia, fine-tuned on forum posts)	**0.57**	**0.68**

We observe that using language model contextual embeddings improves term extraction accuracy, although these improvements are relatively minor and only present when model is fine-tuned on in-domain texts. We found that term extraction models that use language model features generally have high precision (0.75 for diagnosis) compared to models that use skip-gram embeddings (0.62), which makes language model features based models more suitable for construction of co-occurence tables.

4.2. Accuracy of diagnosis prediction

Results of diagnosis prediction on test set are presented in Table 2. We found that accuracy of language-model based method is superior to that of simple co-occurrence baseline, suggesting that language model is capable of leverage additional information contained in descriptions of patients conditions, that is not present in co-occurrence statistics and skip-gram based word embeddings.

Another surprising finding is that accuracy that we obtained from such a noisy dataset is generally high and comparable to that of much more complex systems [Kafle et al, 2018], even through our system operates directly on natural language descriptions and is not allowed to ask additional questions to the user.

We also found that model trained on Wikipedia alone does not have good performance, and using forum posts alone also leads to low accuracy, while fine-tuning Wikipedia model on forum posts leads to superior performance. This could be due to the fact that random subsample of Wikipedia contains very few medical facts (it mostly contains history, sports, and media topics) but is helpful for acquiring representations of basic natural language structure.

It is worth noting, that are goal here is not to achieve better accuracy *per se*, but to establish if unsupervised learning based on language model can obtain knowledge useful for the task of diagnosis prediction. While supervised methods may well be capable of obtaining better accuracy given proper training set, our focus here is primary on unsupervised learning.

Table 2. Accuracy of diagnosis prediction

Method	Diagnostic accuracy @ top1	Diagnostic accuracy @ top10
Co-occurrence + similarity of symptoms	22%	58%
Language model (trained on wikipedia)	1%	5%
Language model (trained on forum posts)	10%	45%
Language model (pre-trained on Wikipedia, fine-tuned on forum posts)	**33%**	**73%**

4.3. Analysis of individual cases and failure modes

In this sections we examine sample results from the system on 3 test cases and analyze possible causes of failures.

Case 1

In the first case the following description was presented to the system:

> *«У меня такая ситуация. Дикие боли в эпигастральной области, больше слева, от поджелудочной вниз к кишечнику. Отрыжка, метеоризм, запоры; во время приступов—расстройства желудка. Боли до еды и после(с тяжестью)»*
> (approximate English translation: *"I have this situation. Wild pain in the Epgaastrin area, more left, from the pancreas down to the intestines. belching, flatulence, constipation; During the attacks-indigestion. Pain before and after eating (with weight)"*)

Correct diagnosis in this case was the diagnosis of "gastritis", which coincides with the first diagnosis proposed by the system. However, it should be noted that the diagnosis of gastritis in this case is not particularly difficult. Among all the suggestions in top 10 plausible options were "colitis", "pancreatitis", "reflux disease". However, system also suggested improbable diagnoses, such as "thyroiditis".

Case 2

> *«першение в горле, усиление кашля с небольшим количеством мокроты сероватого цвета, повышение температуры тела до 37,8 °C, потливость. В течение примерно 25 лет беспокоит кашель, преимущественно по утрам, с небольшим количеством мокроты»*
> (*"Sore throat, increased cough with a small amount of sputum grayish color, increased body temperature up to 37,8°C, sweating. For about 25 years, worries cough, mostly in the mornings, with a little sputum"*)

The description for this case was compiled on the basis of an article in the Medical Journal (Internal Medicine Archive No. 2 (22) 2015-"Clinical case of tuberculosis development under the mask of exacerbation of chronic bronchitis"). As the title of the paper suggests, the reference diagnosis in this case was tuberculosis. We consider the answer of the system in this case to be correct, because the diagnosis of tuberculosis is present in the top 10 suggestions, despite the fact that it is in the ninth place, after bronchitis, common cold, and other respiratory diseases. According to the paper, the diagnosis of tuberculosis in this case poses difficulties, and it was not established by a physician, initially despite the fact that he had the opportunity to accurately examine the patient and conduct laboratory tests, while the system relies solely on a short description.

Case 3

«Потеря сознания, продолжавшаяся 4 мин, сопровождавшаяся судорогами, которые длились около 4 с, затем утихали и потом снова появлялись еще 1–2 раза, а затем исчезали, заведением глазных яблок вверх, слюноотделением, прикусыванием языка, постприступной сонливостью, повышением температуры тела до 37,1°С. История развития настоящего заболевания. Первый приступ произошел 26 августа 2013 г. без видимой причины. Больной в это время отдыхал на море, загорал на пляже. Со слов его матери, приступ длился 1 мин: отмечалось заведение глазных яблок вверх, «потрясывание» всего тела, сначала ног, затем рук, через несколько секунд появилась пена изо рта, немного обмочился. Сам пациент не помнит приступ»

(*"The loss of consciousness lasted 4 minutes, accompanied by convulsions, which lasted about 4 s, then calmed down and then reappeared again 1–2 times, and then disappeared, the establishment of eyeballs up, salivation, the bite of the tongue, drowsiness, increase of body temperature up to 37,1°C. History of development of the present disease. The first attack occurred on August 26, 2013 for no apparent reason. The patient at this time rested on the sea, sunbathing on the beach. From the words of his mother, the attack lasted 1 minute: It was noted the establishment of eyeballs up, "shaking" the whole body, first legs, then hands, after a few seconds appeared foam from the mouth, a little wet. The patient himself does not remember the attack"*)

This description is also an adaptation of the description of the clinical case of epilepsy. Despite the fact that in this case the assumption of the diagnosis of "epilepsy" is not particularly difficult for a human doctor, the correct diagnosis was not in to 10 suggestions. Among the suggestions in this case were: "hypothermia", "arrhythmia", "mitral valve prolapse", "anxiety disorder" and "depression". It is noteworthy that this description is long and poorly adapted, as it contains the text of the description of the doctor rather than the patient. As an experiment, we introduced only a part of this description beginning with the fragment "from the words of his mother" In this case correct diagnosis was obtained on 2nd place, following allergic reaction (anaphylaxis). We hypothesize, that language model may have difficulties in processing too long texts where it is hard to select relevant symptoms, which can be mitigated in the future by using attention-based language model.

5. Conclusions

1. We found that texts, posted in online medical forums can be valuable source of data for training symptoms checkers (diagnosis prediction models), despite their noisy content.
2. Language-model based systems, trained on online medical forums posts, have considerable (73% at top-10) accuracy in detecting correct diagnosis based on user's natural language description of medical condition. This accuracy exceeds that of simple co-occurence based baseline model and possibly approaches accuracy of more complex symptom checkers, for chronic conditions, while still being inferior in diagnosing acute medical emergencies.
3. In line with previous findings, language-model features in form of activations of LSTM top layer hidden units improve medical term extraction accuracy, albeit to a small extent.
4. In summary, our findings, while being preliminary, seems to indicate that large language models can acquire significant domain-specific knowledge, possibly pointing to a completely new way for improving existing diagnostic systems.

References

1. *Barnickel, Thorsten, et al.* (2009). "Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts." PLoS One 4.7: e6393.
2. *Blum, B. I., & Semmel, R. D.* (1991, May). Medical informatics, knowledge, and expert systems. In [1991] Computer-Based Medical Systems@ m_Proceedings of the Fourth Annual IEEE Symposium (pp. 212–218). IEEE.
3. *Fraser, H., Coiera, E., & Wong, D.* (2018). Safety of patient-facing digital symptom checkers. The Lancet, 392(10161), 2263–2264.
4. *Hasan, S. A., Zhao, S., Datla, V. V., Liu, J., Lee, K., Qadir, A., & Farri, O.* (2016). Clinical Question Answering using Key-Value Memory Networks and Knowledge Graph. In TREC.
5. *Hochreiter, S., & Schmidhuber, J.* (1997). Long short-term memory. Neural computation, 9(8), 1735–1780
6. *Irsoy, O., & Cardie, C.* (2014). Opinion mining with deep recurrent neural networks. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 720–728).
7. *Kafle, S., Pan, P., Torkamani, A., Halley, S., Powers, J., & Kardes, H.* (2018). Personalized symptom checker using medical claims. In Proceedings of the Third International Workshop on Health Recommender Systemsco located with Twelfth ACM Conference on Recommender Systems (HealthRec-Sys'18), Vancouver, BC, Canada, October 6, 2018, 5 page.
8. *Kingma, Diederik P., and Jimmy Ba* (2014). "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980.

9. *Levesque, H., Davis, E., & Morgenstern, L.* (2012). The winograd schema challenge. In Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning.
10. *Mallory, Emily K., et al.* (2015). “Large-scale extraction of gene interactions from full text literature using DeepDive.” Bioinformatics: btv476.
11. *Middleton, K., Butt, M., Hammerla, N., Hamblin, S., Mehta, K., & Parsa, A.* (2016). Sorting out symptoms: design and evaluation of the babylon check automated triage system. arXiv preprint arXiv:1606.02041.
12. *Mikolov, T., Chen, K., Corrado, G., & Dean, J.* (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
13. *Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T.* (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports, 6, 26094.
14. *Radford, A., Jozefowicz, R., & Sutskever, I.* (2017). Learning to generate reviews and discovering sentiment. arXiv preprint arXiv:1704.01444.
15. *Ramnarayan, P., Kulkarni, G., Tomlinson, A., & Britto, J.* (2004). ISABEL: a novel Internet-delivered clinical decision support system. Current perspectives in healthcare computing, 245–256.
16. *Razzaki, S., Baker, A., Perov, Y., Middleton, K., Baxter, J., Mullarkey, D. & DoRosario, A.* (2018). A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. arXiv preprint arXiv:1806.10698
17. *Semigran, Hannah L., et al.* (2015) “Evaluation of symptom checkers for self diagnosis and triage: audit study.” bmj 351: h3480.
18. *Sutskever, I., Martens, J., & Hinton, G. E.* (2011). Generating text with recurrent neural networks. In Proceedings of the 28th International Conference on Machine Learning (ICML-11) (pp. 1017–1024).
19. *Trinh, T. H., & Le, Q. V.* (2018). A simple method for commonsense reasoning. arXiv preprint arXiv:1806.02847.