Attentive Multiview Text Representation for Differential Diagnosis

Hadi Amiri[a,c], **Mitra Mohatarami**[b], **Isaac S. Kohane**[c]
[a]Department of Computer Science, University of Massachusetts, Lowell
[b]MIT Computer Science and Artificial Intelligence Laboratory
[c]Department of Biomedical Informatics, Harvard University
Massachusetts, USA
hadi_amiri@uml.edu, mitram@mit.edu, isaac_kohane@harvard.edu

Abstract

We present a text representation approach that can combine different views (representations) of the same input through effective data fusion and attention strategies for ranking purposes. We apply our model to the problem of *differential diagnosis*, which aims to find the most probable diseases that match with clinical descriptions of patients, using data from the Undiagnosed Diseases Network. Our model outperforms several ranking approaches (including a commercially-supported system) by effectively prioritizing and combining representations obtained from traditional and recent text representation techniques. We elaborate on several aspects of our model and shed light on its improved performance.

1 Introduction

Electronic Health Records (EHRs) (Dick et al., 1997) contain a wealth of documented information and insights about patients health and well-being. However, it is difficult to effectively process such data due to complex terminology, missing information, and imprecise clinical descriptions (Friedman et al., 2013; Rajkomar et al., 2019). In addition, an especially challenging class of diseases are orphan or rare diseases (Kodra et al., 2012; Walley et al., 2018), which are diverse in symptoms and affect a smaller percentage of the population.

In this paper, we investigate how well Natural Language Processing (NLP) algorithms could reproduce the performance of clinical experts in the task of *differential diagnosis*–the process of distinguishing a particular disease from others that present similar clinical features, given medical histories (descriptions) of individual patients. We formulate this task as a ranking problem where the aim is to find the most probable diseases given medical histories of patients (Dragusin et al., 2013).

We develop a novel *pairwise* ranking algorithm that combines different views of patient and disease descriptions, and prioritizes effective views through an **A**ttentive **M**ultiview **N**eural **M**odel (AMNM). We research this problem using data from the Undiagnosed Diseases Network (UDN) (Gahl et al., 2015; Ramoni et al., 2017)[1], which includes concise medical history of patients and their corresponding diseases in the Online Mendelian Inheritance in Man (OMIM) dataset (Amberger et al., 2015).[2] All diagnoses–mappings between each patient and corresponding diseases–are provided by a team of expert clinicians from the UDN.

The contributions of this paper are as follows:

- illustrating the impact of NLP in detecting the nature of illness (diagnosis) in patients with rare diseases in a real-world setting, and
- a novel neural approach that effectively combines and prioritizes different views (representations) of inputs for ranking purposes.

Our Attentive Multiview Neural Model employs traditional and recent representation learning techniques and outperforms current pairwise neural ranking approaches through effective data fusion and attention strategies. We conduct several experiments to illustrate the utility of different fusion techniques for combining patient (query) and disease (document) representations.[3]

2 Method

In many domains, entities can be represented from multiple views. For example, a patient can be represented by demographic data, medical history, diagnosis codes, radiology images, etc. We propose a neural model to effectively prioritize important views and combine them for ranking purposes.

[1] https://undiagnosed.hms.harvard.edu/
[2] https://www.omim.org/
[3] code: https://clu.cs.uml.edu/tools.html

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 1012–1019

Figure 1 shows our model, which comprises of three major components: (a): an attention network that estimates and weights the contribution of each view in the ranking process, (b): a fusion network the utilizes intra-view feature interactions to effectively combine query-document representations, and (c): a `softmax` layer at the end that estimates the query-document relevance scores given their combined representations. We first formulate the problem and then explain these components.

2.1 Problem Statement

Let $(\mathbf{q}', \mathbf{d}')$ and $(\mathbf{q}'', \mathbf{d}'')$ denote two different views of the same query and document (throughout the paper, we think of queries and documents as clinical descriptions of patients and diseases respectively).[4] These views can be obtained using traditional (Robertson and Walker, 1994) or recent (Devlin et al., 2019) representation learning techniques applied to textual descriptions or codified data of queries and documents. For example, $\mathbf{q}'$ and $\mathbf{d}'$ can indicate representations of the *texts* of a query and a document, and $\mathbf{q}''$ and $\mathbf{d}''$ can indicate representations of the *medical concepts and codes* associated with the same query and document. Our task is to determine a relevance score between each given query and document. Toward this goal, we effectively prioritize and combine these representations through Attention and Fusion neural networks, which are described blow.

2.2 Attention Model

We develop an attention sub-network to explicitly capture the varying importance of views by assigning attentive weights to them. Specifically, given the embedding vectors of a query $\mathbf{q}^i \in \mathbb{R}^l$ and a document $\mathbf{d}^i \in \mathbb{R}^m$ in the ith view, we use a Feed-forward network, i.e. function $f(.)$ in Figure 1, to estimate the vector $\mathbf{a}$ that captures attention weights across views as follows:

$$f(\mathbf{q}^i, \mathbf{d}^i) = \varphi(\mathbf{W}^q \mathbf{q}^i + \mathbf{b}^q)^\top \cdot \varphi(\mathbf{W}^d \mathbf{d}^i + \mathbf{b}^d),$$
$$\mathbf{a} = \texttt{softmax}([f(\mathbf{q}^i, \mathbf{d}^i), \forall i]), \tag{1}$$

where $\mathbf{W}^q \in \mathbb{R}^{n \times l}$ and $\mathbf{W}^d \in \mathbb{R}^{n \times m}$ are weight matrices to transform the query and document representations into the same underlying space of dimension n, $\mathbf{b}^q \in \mathbb{R}^n$ and $\mathbf{b}^d \in \mathbb{R}^n$ are the trainable bias vectors for the query and document respectively and $\varphi(.)$ is the ReLU function. The

[4] Our model can incorporate any number of views; we only illustrate two views here for simplicity.

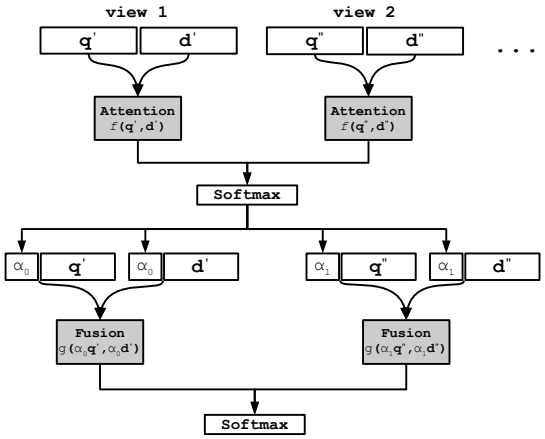

Figure 1: The architecture of our Attentive Multiview Neural Model (AMNM). For simplicity, we illustrate two views only, e.g. $(\mathbf{q}', \mathbf{d}')$ indicates representations of the texts of a query and a document, and $(\mathbf{q}'', \mathbf{d}'')$ indicates representations of the medical codes and concepts associated with the same query and document. $f(.)$ and $g(.)$ indicate attention and fusion functions respectively, and $\mathbf{a}_i$ indicates the attentive weight of the ith view estimated by the attention sub-network.

`softmax` activation function transforms the attention weights to $[0, 1]$ range. Assuming that the query-document pair of the more influential view are more similar in the underlying shared space (estimated by dot product in (1)), $\mathbf{a}$ captures attention weights of different views.

2.3 Fusion Model

Previous learning to rank approaches often concatenate query and document representations to combine their corresponding features (dos Santos et al., 2015; Amiri et al., 2016). There are a few approaches that explicitly capture feature interactions between queries and documents (Severyn and Moschitti, 2015; Echihabi and Marcu, 2003). We extend these fusion techniques and compare them.

Given the attention weights from (1), we develop a fusion sub-network, function $g(.)$ in Figure 1, to capture the intra-view feature interactions for query and document representations of each view. Our fusion network takes as input the *attentive embeddings* of each view, i.e. $(\alpha \times \mathbf{q}, \alpha \times \mathbf{d})$, and combines them through *one* of the following tensor fusion operations:

$$\begin{aligned} g^{dot}(\alpha\mathbf{q}, \alpha\mathbf{d}) &= \alpha^2 \times \varphi(\mathbf{W}^q\mathbf{q} + \mathbf{b}^q)^\top \cdot \\ &\quad \varphi(\mathbf{W}^d\mathbf{d} + \mathbf{b}^d), \\ g^{outer}(\alpha\mathbf{q}, \alpha\mathbf{d}) &= \alpha^2 \times \mathbf{q} \otimes \mathbf{d}, \\ g^{conv}(\alpha\mathbf{q}, \alpha\mathbf{d}) &= \alpha^2 \times \texttt{Conv1d}(\mathbf{q} \otimes \mathbf{d}), \end{aligned} \tag{2}$$

where g^{dot}, g^{outer}, and g^{conv} denote the dot product, outer product, and one-dimensional (1D) convolution with average pooling. In contrast to g^{dot}, g^{outer} and g^{conv} are considerably more expensive operations but may better encode feature interactions. The output of function g is flattened and considered as the *intra-view embedding*.

Finally, we obtain the overall fused representation for each view by concatenating its intra-view and attentive embeddings. The representations of all views are then fed into a `softmax` to estimate the relevance between queries and documents.

3 Experiments

Data: Our data includes medical histories of 257 patients provided by the the Undiagnosed Diseases Network (UDN[5]) (Gahl et al., 2015; Ramoni et al., 2017), as well as general descriptions (including clinical features) of more than 9K diseases available in the Online Mendelian Inheritance in Man (OMIM) dataset (Amberger et al., 2015). The UDN is a nationwide program that improves the level of diagnosis for individual patients (with severe clinical conditions) whose signs and symptoms have been intractable to diagnosis (Kobren et al., 2021; Amiri et al., 2021). To the best of our knowledge, this dataset is the largest available dataset for investigation on rare disease patients. The relevance judgment between patients and diseases is provided by a team of expert clinicians at the UDN. The total number of positive patient-disease pairs is $4,746$, where the number of unique diseases among these pairs is $1,131$; note that different patients can match with the same disease. We split the patients into training (80%), validation (10%), and test (10%) sets. In addition, for each positive pair in the training set, we create a negative pair for the same patient through random sampling of diseases. At test time, we create all the possible patient-disease pair combinations (more than 218K pairs) and use the estimated confidence scores of the classifier to rank all diseases against each test patient. In terms of views, we consider the texts of medical histories and diseases as the first view, and medical concepts and codes extracted from histories by QuickUMLS (Soldaini and Goharian, 2016) as the second view.

We note the concept and code view provides a higher level and more general semantic distinctions by grouping semantically-similar terms, while text view encodes other elements of semantics such as negation, hedging, etc.

Baselines: We consider the following baselines:

• **BM25** (Robertson et al., 1995): An unsupervised approach that effectively predicts relevance based on term frequency, inverse document frequency, and document length.

• **SVMs** (Cortes and Vapnik, 1995): We develop TF/IDF weighted ngrams (n=[1–2]) as features for the text and code/concept views, and conduct exhaustive search over hyperparameters for best performance on validation data. Such features were found effective on clinical texts by previous work (Howes et al., 2012; Reuber et al., 2009).

• **BERT** (Devlin et al., 2019): An attentive bidirectional language model that estimates the relevance between queries and documents by generating contextual representations, jointly conditioned on left and right contexts. We use BERT models developed for clinical text (Alsentzer et al., 2019).[6]

• **SVM**^{rank} (Joachims, 2002): An extension of SVMs to ranking problems which adaptively sorts documents based on their relevance to each query through empirical risk minimization. As features, we use relevance scores or probability predictions generated by the above baselines as well as additional features (unigram overlap and IDF-weighted unigram overlap) (Yu et al., 2014) to better establish the relevance between queries and documents.

• **PhenoTips** (Girdea et al., 2013): This *commercial* tool is currently used at the UDN to assist diagnostic efforts. It utilizes external sources such as the Human Phenotype Ontology (Köhler et al., 2017) and Orphanet data[7] to rank candidate diseases according to their ontology-based similarity to phenotypic descriptions of patients. PhenoTips employs advanced statistical modeling to differentiate candidate diseases, accounts for disorder frequencies in the general population according to Orphanet, supports negative phenotypes–symptoms that were not observed in the patient–and utilizes both code and text views.

[5]Access to phenotypic and genomic UDN data can be granted by submitting an online access request at dbGaP: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001232.v1.p1.

[6]We input medical concepts to BERT by replacing them with their "preferred" concept, determined by UMLS (Lindberg, 1990; Bodenreider, 2004), across all patient and disease descriptions. For example, "diabetes mellitus type 1," "type 1 diabetes," "juvenile diabetes" and "IDDM" are all converted to "juvenile diabetes" (as the preferred concept).

[7]http://www.orpha.net

Settings: Initial representations for patient and disease descriptions are obtained from clinical BERT (Devlin et al., 2019; Alsentzer et al., 2019), i.e. $d_1, d_2 = 768$. In (1) and (2), we set the dimension of the shared space between query and document representations to $n = 100$. In addition, for the CNN fusion model, see (2), we use 250 filters and kernel size of 3. Further details are provided in the supplementary materials.

Evaluation Metrics: We employ Mean Average Precision (MAP), Precision at rank K (P@K), and Precision-Recall curve implemented in `trec_eval`[8] to compare competing systems. We use t-test for significance testing and asterisk mark (*) to indicate significant difference at $\rho = 0.01$.

3.1 Experimental Results

We report the performance of single and multiview models separately to ease comparison between views. The overall MAP and P@K, $\forall$K $\in \{5, 10\}$, performance of baselines for each view are reported in Table 1. The results show that BERT outperforms the other baselines across almost all measures. We attribute the poor performance of BM25 and SVMs to considerable difference in the underlying word/concept distribution in query and document spaces which can't be effectively addressed through lexical features (Burgun and Bodenreider, 2001; Pedersen et al., 2007).[9] In addition, BERT (code view) shows lower performance than BERT (text view). We conjecture that this results could be explained through the following points: (a): BERT is a strong language model and is robust in retrieving noun hypernyms or in completions involving shared category or role reversal (Ettinger, 2020), and (b): replacing medical concepts in text with their preferred concepts (see footnote 6) makes the original text less coherent, which can adversely affect the performance of BERT.

Table 2 shows the performance of SVM^{rank} with combined features across views, PhenoTips, and our Attentive Multiview Neural Model (AMNM) with different fusion functions. AMNM combines traditional and recent representation learning techniques by using BERT representations for text view, and BERT and SVMs representations for code view. All model combinations except for AMNM$_{bert\text{-}svms}$ (g^{conv}) lead to significant improvement against the best performing baseline–BERT (text view) in Table 1. AMNM$_{bert\text{-}bert}$ (g^{dot}) improves the best baseline by 3.4, 1.7 and 5.8 points in MAP, P@5 and P@10 respectively; the corresponding improvement for AMNM$_{bert\text{-}svms}$ (g^{dot}) is 2.9, 5.8 and 6.2 points respectively. We note that AMNM$_{bert\text{-}svms}$ (g^{dot}) leads to considerably higher P@{5,10}, metrics that have a pivotal role in practical use of search systems. In addition, PhenoTips shows comparable MAP to BERT but has considerably lower P@{5,10}.[10]

The fusion functions g^{dot} (dot product) and g^{outer} (outer product) outperform the more expensive fusion function g^{conv} (one-dimensional convolution). The lower performance of g^{conv} could be attributed to average pooling, which assumes different input dimensions equally contribute to the final representation and relevance. As a result, it may fail to eliminate noisy features or prioritize important ones.

	Text View			Code View		
Model	**MAP**	**P@5**	**P@10**	**MAP**	**P@5**	**P@10**
BM25	4.1	5.0	3.8	6.5	8.3	6.3
SVMs	8.8	8.3	8.3	7.7	8.3	8.8
BERT	**15.5**	**12.5**	11.7	**10.8**	**13.3**	**10.8**
SVM^{rank}	12.1	9.2	**12.5**	8.5	8.3	8.6

Table 1: MAP, P@5 and P@10 performance of baselines (in percentages) on text and code views.

Model	**Fusion**	**MAP**	**P@5**	**P@10**
SVM^{rank}	text & code	12.9	12.5	12.9
PhenoTips	text & code	15.4	8.3	5.4
AMNM$_{bert\text{-}bert}$	g^{dot}	**18.9***	14.2	17.5
AMNM$_{bert\text{-}bert}$	g^{outer}	18.0*	16.7	17.5
AMNM$_{bert\text{-}bert}$	g^{conv}	16.0*	10.0	12.1
AMNM$_{bert\text{-}svms}$	g^{dot}	18.4*	**18.3**	**17.9**
AMNM$_{bert\text{-}svms}$	g^{outer}	17.1*	17.5	17.1
AMNM$_{bert\text{-}svms}$	g^{conv}	11.4	14.2	13.9

Table 2: Model performance across different fusion functions. The Model column shows the source of representations for text and code views respectively. * indicates significant improvement against best-performing baseline reported in Table 1.

[8] https://trec.nist.gov/trec_eval/

[9] For example, these models can't effectively match a query containing "congestive heart failure" to relevant documents containing "cardiac decompensation," "pulmonary edema," and "ischemic cardiomyopathy."

[10] We note that, in case of rare and undiagnosed diseases, any small improvement is crucial as it can lead to better diagnostic clues. Clinicians often look at the top K results for clues and potential matches for each patient. Therefore, compared to standard evaluation metrics, a more practical evaluation metric for our task is Hit@K, which measures the likelihood of observing "at least one" relevant disease in the ranked list of top K diseases. The Hit@K ($K = 20$) performance of our model is 0.49, while the corresponding value for our best performing baseline is 0.37.

3.2 Model Analysis

We discuss how and why AMNM achieves its improved performance through the following experiments; see supplementary materials for details:

Prediction Variance Across Views: The Pearson correlation between the Average Precision of BERT (text view) and BERT (code view) on individual test queries (patients) is 0.87, which indicates less performance variation across views at query level. This is while the corresponding correlation between BERT (text view) and SVMs (code view) is only 0.34. The lack of diversity in the performance of BERT across these views could be a source of improvement in AMNM$_{bert\text{-}svms}$.

Attention Function: Given test examples (more than 218K patient-disease pairs), our attention sub-network is expected to assign a higher attentive weight to the view that better estimates the corresponding relevance score. To estimate the accuracy of this sub-network, we separately apply the trained BERT (text view) and SVMs (code view) models to generate their corresponding ranked lists of diseases for test patients. Then, for each *relevant* patient-disease pair, we evaluate our attention function in AMNM$_{bert\text{-}svms}$ by measuring whether it assigns a higher attentive weight to the better view–the view that positions the relevant disease at a higher rank compared to the other view. The results show that (a): our attention sub-network is 57.7% accurate in prioritizing better views, (b): BERT (text view) outperforms SVMs (code view) on 64.7% of relevant patient-disease pairs in terms of relative ranks, and our attention network accurately assigns higher weight to BERT on 88.6% of these examples, and (c): on the remaining 35.3% of examples that SVMs (code view) outperforms BERT (text view) in terms of relative ranks, our attention network assigns higher weight to SVMs in only 0.9% of these examples. Improving this percentage could boost the performance of our model and is the subject of our future work.

4 Related Work

The National Institutes of Health established the Undiagnosed Diseases Network (UDN) (Gahl et al., 2015; Ramoni et al., 2017) to facilitate research on undiagnosed and rare diseases. The UDN is a network of 12 clinical sites, and application to the UDN is open to all individuals who complete the application form and submit a referral letter from a health care professional (Kobren et al., 2021). A committee of experts in a review session reviews each UDN application and makes admission decisions. Walley et al. (2018) investigated major factors that may determine application outcomes of the UDN, which has been found effective in developing computational models for predicting admission outcomes (Amiri et al., 2021). In (Dragusin et al., 2013), authors developed a search engine for rare diseases, named FindZebra[11], which was based on information retrieval techniques available in Indri search engine (Strohman et al., 2005). In addition, previous work developed experimental setup to evaluate and compare search engines such as Google or Bing in predicting relevant diseases to given phenotypes (Shenker, 2014), employed medical anthologies and information content techniques (Köhler et al., 2009), leveraged collaborative filtering (Shen et al., 2017) and ensemble techniques (Jia et al., 2018) for this purpose.

Our work departs from previous research by investigating a multiview approach to undiagnosed patients, where we show effective attention and fusion techniques lead to better pairwise ranking for differential diagnosis.

5 Conclusion and Future Work

Given electronic health records of patients, we develop an attentive multiview text representation model to assist clinical experts by ranking the most probable and relevant diseases. Accurate and timely diagnosis is especially important for critically ill patients as it assists specialists to distinguish, prioritize, and accelerate treatment for such patients. Our work can be improved by (a): enriching the feature space through patient- and disease-specific information such patient demographic information and clinical synopsis of diseases, (b): improving model's attention mechanism, and (c): tackling differences in word distributions across patients (queries) and diseases (documents).

Acknowledgments

Research reported in this manuscript was supported by the NIH Common Fund, through the Office of Strategic Coordination/Office of the NIH Director under Award Number U01HG007530. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

[11] https://www.findzebra.com/

Ethics and Broader Impact Statement

This investigation included a small cohort of diagnosed patients in the Undiagnosed Diseases Network (UDN). The UDN is a network of 12 clinical sites, and application to the UDN is open to all individuals who complete the application form and submit a referral letter from a health care professional; a committee of experts in a review session reviews each UDN application and makes admission decisions. We included all data with no exclusions during the data analysis and manual review, except for cases with missing data or formatting issues. The population will therefore reflect the gender, race, ethnicity, age, and health status of the participating patients. In addition, all results have been presented in aggregate and no attempt have been made to identify individuals or facilities. However, during the course of this research and beyond that, there is a potential risk of loss of patient privacy and confidentiality. We have made and will make every effort to protect human subject information and minimize the likelihood of this risk (all authors with access to the data have successfully completed an education program in the protection of human subjects and privacy protection). In addition, our work is transformational in nature and its broader impacts are first and foremost the potential to improve the well-being of individual patients in the society (individuals who often find themselves on a protracted journey from one specialist to another without diagnosis even in this era of genomic sequencing), and support clinicians in their diagnostic efforts.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Joanna S Amberger, Carol A Bocchini, François Schiettecatte, Alan F Scott, and Ada Hamosh. 2015. Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. *Nucleic acids research*, 43(D1):D789–D798.

Hadi Amiri, Isaac S Kohane, et al. 2021. Machine learning of patient characteristics to predict admission outcomes in the undiagnosed diseases network. *JAMA network open*, 4(2):e2036220–e2036220.

Hadi Amiri, Philip Resnik, Jordan Boyd-Graber, and Hal Daumé III. 2016. Learning text pair similarity with context-sensitive autoencoders. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1882–1892.

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

Anita Burgun and Olivier Bodenreider. 2001. Comparing terms, concepts and semantic classes in wordnet and the unified medical language system. In *Proceedings of the NAACL'2001 Workshop,"WordNet and Other Lexical Resources: Applications, Extensions and Customizations.*

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Richard S Dick, Elaine B Steen, Don E Detmer, et al. 1997. *The computer-based patient record: an essential technology for health care*. National Academies Press.

Radu Dragusin, Paula Petcu, Christina Lioma, Birger Larsen, Henrik L Jørgensen, Ingemar J Cox, Lars Kai Hansen, Peter Ingwersen, and Ole Winther. 2013. Findzebra: a search engine for rare diseases. *International Journal of Medical Informatics*, 82(6):528–538.

Abdessamad Echihabi and Daniel Marcu. 2003. A noisy-channel approach to question answering. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 16–23. Association for Computational Linguistics.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Carol Friedman, Thomas C Rindflesch, and Milton Corn. 2013. Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the national library of medicine. *Journal of biomedical informatics*, 46(5):765–773.

William A Gahl, Anastasia L Wise, and Euan A Ashley. 2015. The undiagnosed diseases network of the national institutes of health: a national extension. *Jama*, 314(17):1797–1798.

Marta Girdea, Sergiu Dumitriu, Marc Fiume, Sarah Bowdin, Kym M Boycott, Sébastien Chénier, David Chitayat, Hanna Faghfoury, M Stephen Meyn, Peter N Ray, et al. 2013. Phenotips: Patient phenotyping software for clinical and research use. *Human mutation*, 34(8):1057–1065.

Christine Howes, Matthew Purver, Rose McCabe, Patrick GT Healey, and Mary Lavelle. 2012. Predicting adherence to treatment for schizophrenia from dialogue transcripts. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 79–83. Association for Computational Linguistics.

Jinmeng Jia, Ruiyuan Wang, Zhongxin An, Yongli Guo, Xi Ni, and Tieliu Shi. 2018. Rdad: a machine learning system to support phenotype-based rare disease diagnosis. *Frontiers in genetics*, 9:587.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142.

Shilpa Nadimpalli Kobren, Dustin Baldridge, Matt Velinder, Joel B Krier, Kimberly LeBlanc, Cecilia Esteves, Barbara N Pusey, Stephan Züchner, Elizabeth Blue, Hane Lee, et al. 2021. Commonalities across computational workflows for uncovering explanatory variants in undiagnosed cases. *Genetics in Medicine*, pages 1–11.

Yllka Kodra, Bernardino Fantini, and Domenica Taruscio. 2012. Classification and codification of rare diseases. *Journal of clinical epidemiology*, 65(9):1026–1027.

Sebastian Köhler, Marcel H Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N Robinson. 2009. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *The American Journal of Human Genetics*, 85(4):457–464.

Sebastian Köhler, Nicole A Vasilevsky, Mark Engelstad, Erin Foster, Julie McMurry, Ségolène Aymé, Gareth Baynam, Susan M Bello, Cornelius F Boerkoel, Kym M Boycott, et al. 2017. The human phenotype ontology in 2017. *Nucleic acids research*, 45(D1):D865–D876.

C Lindberg. 1990. The unified medical language system (umls) of the national library of medicine. *Journal (American Medical Record Association)*, 61(5):40–42.

Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–299.

Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. 2019. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358.

Rachel B Ramoni, John J Mulvihill, David R Adams, Patrick Allard, Euan A Ashley, Jonathan A Bernstein, William A Gahl, Rizwan Hamid, Joseph Loscalzo, Alexa T McCray, et al. 2017. The undiagnosed diseases network: accelerating discovery about health and disease. *The American Journal of Human Genetics*, 100(2):185–192.

Markus Reuber, Chiara Monzoni, Basil Sharrack, and Leendert Plug. 2009. Using interactional and linguistic analysis to distinguish between epileptic and psychogenic nonepileptic seizures: a prospective, blinded multirater study. *Epilepsy & Behavior*, 16(1):139–144.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pages 232–241. Springer.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Cícero dos Santos, Luciano Barbosa, Dasha Bogdanova, and Bianca Zadrozny. 2015. Learning hybrid representations to retrieve semantically equivalent questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 694–699, Beijing, China. Association for Computational Linguistics.

Aliaksei Severyn and Alessandro Moschitti. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 373–382.

Feichen Shen, Sijia Liu, Yanshan Wang, Liwei Wang, Naveed Afzal, and Hongfang Liu. 2017. Leveraging collaborative filtering to accelerate rare disease diagnosis. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1554. American Medical Informatics Association.

Bennett S Shenker. 2014. The accuracy of internet search engines to predict diagnoses from symptoms can be assessed with a validated scoring system. *International journal of medical informatics*, 83(2):131–139.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR Workshop, the 39th international ACM SIGIR conference on research and development in information retrieval*.

Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. 2005. Indri: A language model-based search engine for complex queries. In *Proceedings of the international conference on intelligent analysis*, volume 2, pages 2–6. Citeseer.

Nicole M Walley, Loren DM Pena, Stephen R Hooper, Heidi Cope, Yong-Hui Jiang, Allyn McConkie-Rosell, Camilla Sanders, Kelly Schoch, Rebecca C Spillmann, Kimberly Strong, et al. 2018. Characteristics of undiagnosed diseases network applicants: implications for referring providers. *BMC health services research*, 18(1):1–8.

Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep learning for answer sentence selection. *Deep Learning Workshop, Neural Information Processing*.