Improving Rare Disease Classification Using Imperfect Knowledge Graph

Xuedong Li*, Yue Wang†, Dongwu Wang§, Walter Yuan§, Dezhong Peng*, Qiaozhu Mei‡
*College of Computer Science, Sichuan University, Chengdu, China
†School of Information and Library Science, University of North Carolina at Chapel Hill
§MobLab Inc., Pasadena, CA
‡School of Information, University of Michigan, Ann Arbor, MI
Email: *xuedong.li@stu.scu.edu.cn, †wangyue@email.unc.edu, §{dongwu.wang, walter.yuan}@moblab.com,
*pengdz@scu.edu.cn, ‡qmei@umich.edu

***Abstract*—Objective: Accurately recognizing rare diseases based on symptom description is an important task in patient triage, early risk stratification, and target therapies. However, due to the very nature of rare diseases, the lack of historical data poses a great challenge to machine learning-based approaches. On the other hand, medical knowledge in automatically constructed knowledge graphs (KGs) has the potential to compensate the lack of labeled training examples. This work aims to develop a rare disease classification algorithm that makes effective use of a knowledge graph, even when the graph is imperfect.**

Method: We develop a text classification algorithm that represents a document as a combination of a "bag of words" and a "bag of knowledge terms," where a "knowledge term" is a term shared between the document and the subgraph of KG relevant to the disease classification task. We use two Chinese disease diagnosis corpora to evaluate the algorithm. The first one, HaoDaiFu, contains 51,374 chief complaints categorized into 805 diseases. The second data set, ChinaRe, contains 86,663 patient descriptions categorized into 44 disease categories.

Results: On the two evaluation data sets, the proposed algorithm delivers robust performance and outperforms a wide range of baselines, including resampling, deep learning, and feature selection approaches. Both classification-based metric (macro-averaged F_1 score) and ranking-based metric (mean reciprocal rank) are used in evaluation.

Conclusion: Medical knowledge in large-scale knowledge graphs can be effectively leveraged to improve rare diseases classification models, even when the knowledge graph is incomplete.

***Index Terms*—rare disease diagnosis, knowledge graph, machine learning, text classification, extremely imbalanced data**

I. Introduction

A disease is defined as *rare* if it affects fewer than 1 in 2,000 people in Europe [1], or it affects fewer than 200,000 people in the United States (1 in 1,500) [2]. China has recently released its first national list of rare diseases [3]. Across the globe, hundreds of millions of people could be affected by one of about 6,000 known rare diseases [4].

Accurate diagnosis of rare diseases is an important task in patient triage, risk stratification, and targeted therapies. Rare disease symptoms often appear unfamiliar and atypical to a clinician, as the cases are too rare to encounter [5]. This brings significant challenge for clinicians to diagnose rare diseases timely, and calls for machine-assisted diagnosis methods.

Rare disease diagnosis is challenging to machine learning approaches as well. Machine learning algorithms often require a significant number of training examples to achieve a good generalization performance. However, by the very nature of rare diseases, the number of relevant clinical records is bounded by the size of population. To compensate the lack of training data for rare disease diagnosis, we need to make use of domain knowledge. Recent efforts in information extraction and knowledge engineering communities have created large-scale knowledge graphs [6]–[8], in which a large number of entities and relations are extracted from unstructured and semi-structured data, verified manually or semi-automatically, and then organized into a massive graph. Although many of these knowledge graphs are freely available as web-based services, most of them have limited coverage and accuracy. They are often built without considering downstream machine learning tasks, therefore imperfect from a task point of view. In this paper, we are interested in leveraging such knowledge resources in machine-assisted rare disease diagnosis.

We present a simple and effective statistical learning method that improves rare disease classification using an imperfect knowledge graph. We define a rare disease in its statistical sense, *i.e.* a disease that affects a small percentage (in this paper, less than 0.1%) of the population in a large disease diagnosis corpus. The proposed method is based on the intuition that if a rare disease has a corresponding entity in the knowledge graph, then we can use this piece of knowledge to guide the classifier on "where to focus" when examining a clinical document. This proves to be an effective strategy in classifying rare diseases when the training documents are too few for the algorithm to learn informative features. On two disease classification corpora, the proposed method demonstrates robust improvements over strong baseline methods on rare diseases diagnosis.

II. Related Work

Machine-assisted rare disease diagnosis. Machine-assisted diagnosis approaches have attracted various lines of research recently [5]. Svenstrup et al. developed a search system that, given symptoms as a search query, returns probable rare disease diagnosis [9]. MacLeod et al. applied gradient boosted decision tree classifiers on behavioral survey data to identify potential rare diseases. Shen et al. proposed a neighborhood-based collaborative filtering algorithm, where patients with similar phenotypes receive similar diagnosis [10]. Their follow-up work further incorporated phenotype-disease associations in biomedical literature [11] and biomedical ontology [12] to improve disease recommendation results. In the current work, we approach rare disease diagnosis in a multiclass classification formulation, which has been shown to deliver state-of-the-art performance in Web-scale applications like ranking and recommendation [13], [14].

Imbalanced data classification. From a machine learning perspective, rare diseases in a patient population can be viewed as rare classes in a data set, which is a typical example of imbalanced data set. We can therefore consider imbalanced learning techniques in rare disease classification [15]. Typical imbalanced learning techniques include resampling, cost-sensitive learning, and rare class data synthesis [16]. However, typical machine learning research deals with class imbalance ratios between 1:4 and 1:100, and few previous works tackle imbalance ratio as extreme as 1:1,000 or lower [17]. In this study, we only consider resampling as one of the potential methods, as its performance closely resembles that of cost-sensitive learning, and synthesizing text documents from rare classes is itself a challenging task.

Feature engineering. When training documents are too few to provide high-quality features, various feature engineering techniques can help enhance data representation. Feature selection methods can be used to identify informative features for the classification task and discard irrelevant features to alleviate overfitting, especially for high-dimensional data such as text [18]. Instead of reducing features, feature generation aims to add features using external knowledge [19]. The technique first identifies a set of knowledge concepts related to a given document, and then "appends" informative words in these concepts to the document. In between the above two strategies are feature labeling and highlighting, which originated from interactive machine learning literature [20]–[22]. These methods use domain knowledge to identify a subset of existing informative features, then incorporate them as certain type of informative prior in subsequent classifier training process. In this study, we evaluate various feature engineering methods for integrating domain knowledge into disease classification algorithm.

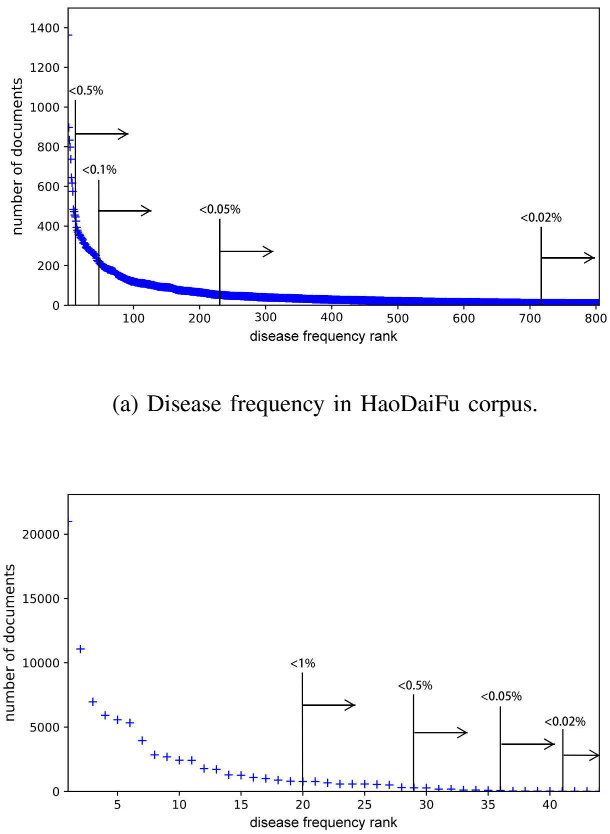
III. Data Description and Problem Formulation

We start by describing the two corpora and the knowledge graph used in our study, followed by our definition of rare diseases, all of which lead to our problem formulation.

TABLE I: Corpora statistics. A "knowledge terms" is a term appearing in medical knowledge graph (see Section IV-B).

	HaoDaiFu	ChinaRe
# of documents	51,374	86,663
# of classes (diseases)	805	44
vocabulary size	59,879	41,087
average # of words/doc	26.7	29.7
average # of knowledge terms/doc	10.8	4.0

Corpora: HaoDaiFu and ChinaRe. We use two Chinese patient diagnosis corpora. The first corpus, HaoDaiFu, contains 51,374 patient records categorized into 805 diseases. Each document contains the symptom description submitted by a patient to Haodf.com, a major Chinese online platform that connects patients to doctors. The second corpus, ChinaRe, contains 86,663 patient records categorized into 44 disease categories. Each document contains the symptom description of a patient written by an insurance professional in ChinaRe, which is one of the largest reinsurance groups in China. Table I summarizes basic statistics of the two corpora. For Chinese word segmentation, we use the Jieba package [23].



(a) Disease frequency in HaoDaiFu corpus.

(b) Disease frequency in ChinaRe corpus.

Fig. 1: Zipf's plots of disease frequency in the two corpora. The x-axis is the disease frequency rank; the y-axis is the disease frequency (number of documents in the disease category). Common diseases appear on the left; rare diseases correspond to the long tail on the right. We annotate cutoff ranks above which the diseases are rarer than the specified percentage.

Figure 1 shows disease distributions of the two corpora. We see that both distributions are highly skewed: a few diseases account for thousands of people, while many diseases affect a small percentage of the population.

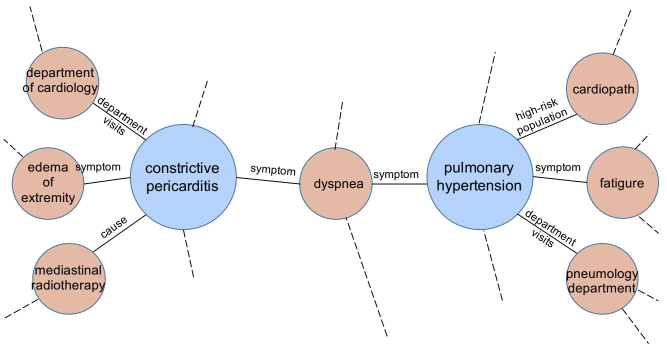


Fig. 2: An illustrative example of two disease entities and some of their attributes in a knowledge graph.

Knowledge graph: CN-DBpedia. A knowledge graph (KG), also known as an ontology, is a collection of entities and relations between entities. An entity has a set of attributes, some of which may itself be an entity. Figure 2 illustrates a small part of a medical KG.

In our study, it would be ideal to have a well-curated medical KG. Unfortunately there is no equivalent of English medical KG like the Unified Medical Language System (UMLS) in Chinese[1]. We therefore resort to a general Chinese knowledge graph, CN-DBpedia [24]. It aggregates knowledge from various resources and constructed in a similar manner as DBpedia. At the time of writing, it contains 16,892,423 entities and 223,137,127 relations. We use a web-based platform that provides RESTful API access to CN-DBpedia (Knowledge Works [25]). Given a textual query, the API returns matched entities. This allows us to perform entity linking relatively easily. Since CN-DBpedia is automatically constructed from Chinese equivalents of Wikipedia, it does not have perfect coverage over all medical entities, and its medical content quality may be inaccurate or incomplete. Not all diseases in the above two corpora have a corresponding entity in the current CN-DBpedia. We find an entity for 751 out of 805 diseases in HaoDaifu and 37 out of 44 diseases in ChinaRe.

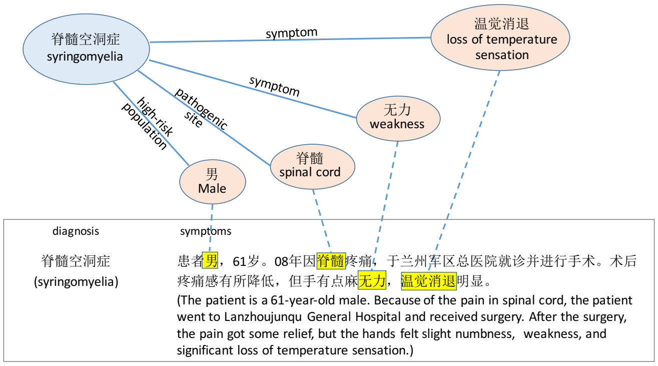
Rare disease definition. Since different countries and regions adopt different definitions of rare diseases [1], [2], and new rare diseases continue to be registered [3], there is no commonly accepted definition of rare diseases.

For the purpose of this study, we define a rare disease in its statistical sense: a disease is rare if it affects no more than a small percentage of the patient records in a large diagnosis corpus. We set the percentage to 0.1%, or 1/1,000, which is slightly higher than the 1/1,500 – 1/2,000 threshold used in the United States and Europe, since both corpora are biased samples of the entire population, *i.e.*, missing the healthy subpopulation. This definition allows us to develop and evaluate algorithms on a wide variety of statistically rare diseases observed in empirical data. In HaoDaiFu, 571 diseases have a percentage lower than 0.1% of all the records. In ChinaRe, 10 diseases have a percentage lower than 0.1% of all the records.

Problem formulation. Our goal is to build text classification algorithms that can automatically assign a disease label given the narrative description of a patient's symptoms. Besides a set of training documents, we also assume access to an existing knowledge graph that contains an entity for (at least a subset of) the diseases in question. In this paper, we specifically focus on classifying rare diseases, or diseases accounting for no more than 0.1% of records in a corpus.

IV. Knowledge Graph Enhanced Rare Disease Classification

This section describes the proposed method for KG-enhanced rare disease classification. The basic idea is to use external knowledge to "emphasize" existing features in the classifier. To illustrate, let us consider a concrete example in Figure 3. Suppose we want to detect the rare disease *syringomyelia* in text, but the training documents are extremely few (in the HaoDaifu corpus, 12 out of 41,105, or 1 out of 3,425 records). A text classifier essentially aims to identify important words among many irrelevant words that indicate *syringomyelia*. This is a difficult task given the very few training documents and a large vocabulary of words. How can we identify important features, assuming we have access to a KG? A natural strategy is to look up the entity *syringomyelia* in the KG, take the attributes that describe this entity, and "inform" the classifier that words mentioned in the attributes are important features. Figure 3 illustrates this idea.



diagnosis	symptoms
脊髓空洞症 (syringomyelia)	患者男，61岁。08年因脊髓疼痛，于兰州军区总医院就诊并进行手术。术后疼痛感有所降低，但手有点麻无力，温觉消退明显。 (The patient is a 61-year-old male. Because of the pain in spinal cord, the patient went to Lanzhoujunqu General Hospital and received surgery. After the surgery, the pain got some relief, but the hands felt slight numbness, weakness, and significant loss of temperature sensation.)

Fig. 3: An illustrative example of using knowledge graph to "emphasize" features (words) in a document. This is an ideal case, where the highlighted features are relevant to the diagnosis. In practice, all features that appear in diagnosis-related part of KG will be highlighted.

Below we describe our method in detail. It comprises of three steps: (1) To identify relevant KG entity (or entities) for each disease; (2) To extract important word features from a given KG entity; (3) to incorporate the importance of features into a text classifier.

[1] Only a small fraction of UMLS concepts has Chinese translation. As it is challenging to guarantee accurate translation of an English KG to Chinese using machine translation, we leave this direction for future work.

A. Mapping diseases to KG entities

In this step, we use the KG API to map a disease to the corresponding KG entity. The API performs entity linking and resolves different surface forms (or "mentions") to the same entity, e.g. mapping "cancer" and "malignancy" to the *cancer* entity. Some disease names may have ambiguous matches. For instance, *insomnia* matches both a health-related entity and a song. To filter out non-medical entities, we further check the category attribute of an entity. We call it the *matched entity* of a disease.

As discussed before, some diseases cannot be mapped to an entity due to the incompleteness of the KG in use. We devised a fall-back strategy to handle these cases. The goal here is to identify not the exact, but the most relevant, entity of a disease. To do so, we evaluate the content overlap between a disease (represented by high inverse document frequency words in all documents of a disease) and an entity (represented by words in its various attributes), and select the entity with the highest content overlap. We call it the *surrogate entity* of a disease.

B. Acquiring knowledge features from KG entities

To facilitate further discussion, we denote the native word features found in all training documents as V, where Chinese stop words are removed.

If a disease has a matched entity, we use words in its attributes and related entity names to form disease features. Accumulating over all diseases, we obtain a set of words K_1. K_1 has overlap with V but may also contain words not in V.

If a disease has a surrogate entity, we do not extract features as above because unlike a mapped entity, the attributes of a surrogate entity are highly likely to be irrelevant to the associated disease. We only extract words that appear at least once in any training document of the disease and appear in 0.01% of KG entities (to ensure specificity – similar to the idea of inverse document frequency). This gives us a set of words K_2. By construction, $K_2 \subset V$.

We call the union set $K = K_1 \cup K_2$ *knowledge features*, or *knowledge terms*.

C. Integrating knowledge features into text classifier

Choice of text classifier. We employed one-vs-rest support vector machine (SVM) classifier with linear kernel, with sparse bag-of-words (BOW) feature representation. In pilot experiments, we found that dense classification models such as long short-term memory (LSTM) perform comparably with linear SVM on frequent diseases but significantly worse on rare diseases. Indeed, deep learning models need a large quantity of training examples to perform well, which is unavailable for rare diseases. So we exluded deep learning models in later experiments.

Feature vector construction. Given BOW feature set V and knowledge feature set K, we construct the feature vector for a document d as follows (d is viewed as a set of words):

1) Construct a $|V|$-dimensional count vector for BOW features, then apply TF-IDF (term frequency-inverse document frequency [26]) transformation and L_2 length normalization;
2) Construct a $|V \cap K|$-dimensional count vector for knowledge features in $d \cap K$, then apply TF-IDF transformation and L_2 length normalization;
3) Concatenate the above two vectors to represent the document.

The first step constructs a feature vector for the original document. The second step constructs a feature vector for words in the document that are mentioned in KG ($d \cap K$). Concatenating feature vectors is also called *early fusion* in multimodal learning, where different vector segments correspond to different modalities of the same data [27].

If a document contains a word $w \in V \cap K$, then it will appear twice in the feature vector: one as a BOW feature, the other as a knowledge feature. Note that the two feature values will not be identical, since the two vectors will have different L_2 lengths before normalization. Such a word will receive a larger feature value in the second vector, since the "KG-mentioned part" ($d \cap K$) is shorter than the original document (d). Table I shows that each document in HaoDaiFu has 26.7 words on average, in which 10.8 words are knowledge features. The ratio is lower in ChinaRe (4.0/29.7). Therefore, the second feature vector can be understood as *emphasizing knowledge features in a document*.

V. Experimental Evaluation

In this section, we evaluate the effectiveness of proposed method and a suite of baseline settings on the rare diseases in the two corpora.

Train-test split. To reduce the variance of results due to a random train-test split, we average the results of 10 runs. In each run, we randomly split the corpus into 80% for training and 20% for test. To avoid the case where some classes do not appear in training or test set, the random split is applied on a per-class basis.

A. Compared methods

All compared methods use one-vs-rest linear SVM classifier, sparse feature representation, and default hyperparameter setting ($C = 1$).

Methods that do not make use of knowledge features:

1) **BOW**: only use BOW feature vector in Section IV-C.
2) **UpSample**: upsample the rare disease documents in the training set, so that each disease has equal number of documents. This is a standard method for imbalanced classification.
3) χ^2: use $|V \cap K_1|$ features selected by the χ^2 criterion.
4) **BOW+**χ^2: concatenate the BOW and χ^2 feature vectors in the same manner as in Section IV-C.

Methods that make use of knowledge features:

1) **KG**$_1$: only use $V \cap K_1$ as features;
2) **KG**$_{12}$: only use $(V \cap K_1) \cup K_2$ as features;
3) **BOW+KG**$_1^{\text{early-fusion}}$: concatenate BOW and KG$_1$ feature vectors as in Section IV-C;

TABLE II: Rare disease classification performance on HaoDaiFu corpus. The higher F_1 and MMR, the better. Each column's highest number is shown in **boldface**, second highest number shown with underline. Columns on the left side of the dividing line are rare disease bins; columns on the right are for comparison purposes.

Percentage Bins	(0, 0.02%] 89 diseases		(0.02%, 0.05%] 277 diseases		(0.05%, 0.1%] 205 diseases		(0.1%, 0.5%] 194 diseases		(0.5%, 1%] 32 diseases	
	F_1	MRR	F_1	MRR	F_1	MRR	F_1	MRR	F_1	MRR
BOW	34.10	45.86	40.80	49.91	49.48	58.81	**53.23**	**62.80**	**62.23**	**75.31**
UpSample	35.17	47.10	40.69	50.43	47.63	57.63	49.85	59.75	58.6	68.95
χ^2	34.04	46.75	40.81	50.66	49.15	58.53	51.74	61.38	61.55	74.05
BOW+χ^2	34.56	47.25	42.41	**51.84**	50.03	**59.33**	53.15	62.34	62.10	73.97
KG$_1$	33.66	44.98	38.25	47.45	45.17	53.97	48.07	57.55	59.21	71.29
KG$_{12}$	33.51	44.92	39.08	48.07	45.23	54.55	48.66	58.00	59.2	71.43
BOW+KG$_1^{\text{pseudo-doc}}$	31.91	42.81	37.51	46.08	44.08	53.22	47.01	56.94	55.91	69.47
BOW+KG$_1^{\text{pseudo-count}}$	34.87	46.14	41.74	50.14	49.31	57.94	52.56	61.59	61.65	74.19
BOW+KG$_1^{\text{late-fusion}}$	33.33	45.42	38.41	48.68	47.15	56.39	51.13	60.18	61.42	73.30
BOW+KG$_1^{\text{early-fusion}}$	36.87	**48.36**	**43.11**	51.79	**50.06**	58.99	52.86	61.90	61.90	73.57
BOW+KG$_{12}^{\text{early-fusion}}$	**36.94**	48.22	42.63	51.40	49.66	58.62	52.60	61.51	61.47	73.23

TABLE III: Rare disease classification performance on ChinaRe corpus. See the caption of Table II above for details.

Percentage Bins	(0, 0.02%] 5 diseases		(0.02%, 0.05%] 3 diseases		(0.05%, 0.1%] 2 diseases		(0.1%, 0.5%] 7 diseases		(0.5%, 1%] 9 diseases	
	F_1	MRR	F_1	MRR	F_1	MRR	F_1	MRR	F_1	MRR
BOW	91.58	93.36	29.76	53.97	90.49	93.49	88.69	92.64	92.6	95.09
UpSample	88.36	94.81	52.22	66.54	90.11	93.06	89.36	94.27	92.62	95.76
χ^2	91.38	95.83	47.97	65.12	90.40	93.68	91.92	95.41	93.84	**96.45**
BOW+χ^2	93.37	97.55	42.14	62.8	90.73	93.95	**92.01**	**95.55**	**94.05**	96.43
KG$_1$	91.06	97.47	22.63	43.64	48.52	48.11	80.54	86.67	74.32	77.33
KG$_{12}$	92.26	**97.70**	31.20	43.91	85.61	91.42	83.71	87.96	80.05	83.18
BOW+KG$_1^{\text{pseudo-doc}}$	75.68	82.49	34.86	52.08	83.20	87.84	78.79	85.57	88.34	91.86
BOW+KG$_1^{\text{pseudo-count}}$	88.14	91.02	30.04	52.62	89.02	93.61	85.54	88.64	90.8	93.34
BOW+KG$_1^{\text{late-fusion}}$	89.01	95.41	29.76	48.8	68.63	70.8	86.18	89.65	86.89	86.21
BOW+KG$_1^{\text{early-fusion}}$	92.30	97.66	**54.73**	**69.88**	90.27	92.54	91.00	95.05	93.59	95.92
BOW+KG$_{12}^{\text{early-fusion}}$	**93.43**	97.13	47.78	62.04	**91.68**	**95.41**	90.70	94.49	93.46	95.70

4) **BOW+KG$_{12}^{\text{early-fusion}}$**: concatenate BOW and KG$_{12}$ feature vectors as in Section IV-C.

Other variants that also make use of both BOW features and KG$_1$ features:

1) **BOW+KG$_1^{\text{late-fusion}}$**: the late fusion strategy (in multi-modal learning [27]): we combine two SVM predictions: one trained on BOW vectors, the other trained on KG$_1$ vectors. To combine the predictions for each document, we rank the predicted labels from most to least probable, and combine the two predicted lists using Borda's rank aggregation method [28].
2) **BOW+KG$_1^{\text{pseudo-count}}$**: the pseudo count strategy [20]: for each word in a given document that also appears in KG$_1$, we add k pseudo word counts to the BOW feature vector. We tuned $k = 1, 10, 100$ and set $k = 1$ as it gives the best performance.
3) **BOW+KG$_1^{\text{pseudo-doc}}$**: the pseudo document strategy: we view the mention of a knowledge feature in a training document as annotating the rationale of the label. We then use the rationale learning strategy to generate pseudo training documents [29].

B. Evaluation metrics

To evaluate the effect of different methods at different rarity levels, we bin the diseases by their percentage in a corpus. Three bins are below 0.1% (our definition of *rare* diseases):

- (0 – 0.02%]: no more than 1/5,000;
- (0.02% – 0.05%]: 1/5,000 to 1/2000;
- (0.05% – 0.1%]: 1/2,000 to 1/1,000.

For a comprehensive comparison, we also include two bins between 0.1% and 1%:

- (0.1% – 0.5%]: 1/1,000 to 1/200;
- (0.5% – 1%]: 1/200 to 1/100.

Machine-assisted diagnosis can be viewed both as a classification task (to assign a disease label to a document) and a ranking task (to sort disease labels by their relevance to a document). To evaluate the classification performance, we use macro-averaged F_1 score [30] as it balances precision and recall and is not biased by majority classes. To evaluate the retrieval performance, we use mean reciprocal rank (MRR) [31] since in both corpora, each document has only one associated disease. We report macro-averaged F_1 and MRR on the test data of each bin.

In Table II and III, we report macro-averaged F_1 and MRR results in each bin across different methods.

C. Discussion

First, we observe that for rare diseases (three bins under 0.1%), the proposed methods BOW+$\text{KG}_1^{\text{early-fusion}}$ and BOW+$\text{KG}_{12}^{\text{early-fusion}}$ delivers robust performance: they are almost always among the top two performers on both corpora. As the disease becomes less rare (two bins above 0.1%), simple BOW baseline and supervised feature selection work better. This is expected as the proposed methods can be viewed as doing feature selection using external knowledge. With more training data in each class, the knowledge inside training data allows us to select higher quality, more task-specific features than external knowledge.

In the disease-to-KG-entity mapping step (Section IV-A), including surrogate entities is sometimes beneficial to rare disease classification, but not always. The performance gain of having higher entity coverage (BOW+KG_{12} compared to BOW+KG_1) is the most salient when the disease is extremely rare (below 0.02%). This suggests that if we had a more complete KG, the rare disease classification performance could be even better.

The performance of upsampling is very unstable, which agrees with previous literature [16]. It dramatically improves classification performance in one specific case (ChinaRe, 0.02% – 0.05%). But in most other cases, upsampling does not help or even hurt performance compared to the BOW baseline. This suggests that resampling methods are not suitable for very imbalanced text classification tasks.

On rare diseases, concatenating the vectors of original BOW features and knowledge features tends to perform better than using either alone, for both χ^2-selected features and KG-selected features. We can understand this phenomenon as a type of regularization: the selected feature segment can be understood as “to put emphasis on these features”. Or equivalently, it can be understood as “to reduce attention (lower the weights) on the rest of the BOW features”. Conceptually, this is related to group-wise regularization: to apply different regularization strengths on two groups of features: $V \cap K$ and $V \backslash K$. The problem with this approach is that for each disease, we would need a different hyperparameter to balance the strength of regularization on two groups of features. In contrast, the proposed method does not have this problem.

Among different ways of using the KG feature information, we found that early fusion performs the best. Combining classification predictions (late fusion) is challenging at the global level, since the combination weights might be different for different diseases. The pseudo-count method has no significant effect, because increment the count of an existing term by 1 has a diminishing effect after TF-IDF transformation. On the other hand, a large pseudo-count makes the document vector as if containing only selected features. Instead, allocating additional dimensions for these features turns out to be more beneficial. It has been shown that text classification can benefit from having many redundant but not perfectly correlated features [32]. Finally, the pseudo-example method performs poorly because it generates more examples for large classes, making small classes even smaller.

D. Implication

One of the biggest challenges in applying machine learning techniques to healthcare is the lack of supervision signals in this domain. Unlike other domains (e.g., image, speech) where the availability of training labels is bounded by the annotation budget, in healthcare it is bounded by the availability of domain experts, and in the case of (rare) diseases, also bounded by the population of patients [33]. How to efficiently transfer domain knowledge into supervision signals for training machine learning models has been a heated debate in both the research community and industry of medical NLP. Under resource constraints, should the effort be spent on labeling additional training examples, or constructing knowledge graphs? Despite many potential advantages of knowledge graphs over unstructured annotations (e.g., precise and compact knowledge representation, extendable, reusable for different tasks [34]), there is always a concern that building a complete and accurate KG can be labor-intensive, if not impossible.

This work shows that a knowledge graph does not have to be perfect (in terms of coverage and accuracy) to be able to deliver desirable benefits for medical NLP tasks. Our use of a general-purpose KG also indicates that practitioners could start with customizing and refining an open domain KG for their tasks instead of building a medical KG from the scratch. Our results should resolve some of the concerns of building knowledge graphs in the practices of medical NLP.

VI. Conclusion

This paper studied the problem of rare disease classification, where rare diseases are defined by their presence in a large corpus (lower than 0.1%). We developed a text classification algorithm that represents a document in two as a combination of a “bag of words” and a “bag of knowledge terms”, where a “knowledge term” is a term shared between the document and the subgraph of knowledge graph relevant to the disease classification task. On two Chinese disease classification corpora, the algorithm delivers robust performance gain over feature selection methods on rare diseases.

In future work, we plan to explore a variety of methods for improving document representation. First, instead of “emphasizing” all words that appear in medical-related KG, we can do so more selectively. One way is to identify the most relevant KG entities to a specific document, and only emphasize words in those entities. We can use synonyms and word embedding methods to allow for fuzzy matching between KG entities and a document, to increase the coverage of knowledge features in a document. We can also consider “appending” words in relevant entities to a document, effectively performing feature generation. Finally, when medical experts are interacting with a list of predicted rare diseases or most similar patients, we can explore the opportunity of learning from experts feedback and improve the diagnosis algorithm continuously.

References

[1] European Commission, "Rare diseases," Accessed March 26, 2019). [Online]. Available: https://ec.europa.eu/health/non_communicable_diseases/rare_diseases_en

[2] United States Department of Health and Human Services, "National organization for rare disorders (nord)," Accessed March 26, 2019). [Online]. Available: https://www.nidcd.nih.gov/directory/national-organization-rare-disorders-nord

[3] J. He, Q. Kang, J. Hu, P. Song, and C. Jin, "China has officially released its first national list of rare diseases," *Intractable & rare diseases research*, vol. 7, no. 2, pp. 145–147, 2018.

[4] "Orphanet: The portal for rare diseases and orphan drugs," Accessed March 26, 2019). [Online]. Available: https://www.orpha.net/consor/cgi-bin/index.php

[5] D. Svenstrup, H. L. Jørgensen, and O. Winther, "Rare disease diagnosis: a review of web search, social media and large-scale data-mining approaches," *Rare Diseases*, vol. 3, no. 1, p. e1083145, 2015.

[6] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM, 2008, pp. 1247–1250.

[7] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," *Proceedings of the VLDB Endowment*, vol. 11, no. 3, pp. 269–282, 2017.

[8] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag, "Learning a health knowledge graph from electronic medical records," *Scientific reports*, vol. 7, no. 1, p. 5994, 2017.

[9] R. Dragusin, P. Petcu, C. Lioma, B. Larsen, H. L. Jørgensen, I. J. Cox, L. K. Hansen, P. Ingwersen, and O. Winther, "Findzebra: a search engine for rare diseases," *International journal of medical informatics*, vol. 82, no. 6, pp. 528–538, 2013.

[10] F. Shen, S. Liu, Y. Wang, L. Wang, N. Afzal, and H. Liu, "Leveraging collaborative filtering to accelerate rare disease diagnosis," in *AMIA Annual Symposium Proceedings*, vol. 2017. American Medical Informatics Association, 2017, p. 1554.

[11] F. Shen, S. Liu, Y. Wang, A. Wen, L. Wang, and H. Liu, "Utilization of electronic medical records and biomedical literature to support the diagnosis of rare diseases using data fusion and collaborative filtering approaches," *JMIR medical informatics*, vol. 6, no. 4, 2018.

[12] F. Shen and H. Liu, "Incorporating knowledge-driven insights into a collaborative filtering model to facilitate the differential diagnosis of rare diseases," in *AMIA Annual Symposium Proceedings*, vol. 2018. American Medical Informatics Association, 2018, p. 1505.

[13] R. Babbar and B. Schölkopf, "Dismec: Distributed sparse machines for extreme multi-label classification," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 2017, pp. 721–729.

[14] H. Jain, V. Balasubramanian, B. Chunduri, and M. Varma, "Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 2019, pp. 528–536.

[15] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge & Data Engineering*, no. 9, pp. 1263–1284, 2008.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[17] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[18] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Icml*, vol. 97, no. 412-420, 1997, p. 35.

[19] E. Gabrilovich and S. Markovitch, "Feature generation for text categorization using world knowledge," in *IJCAI*, vol. 5, 2005, pp. 1048–1053.

[20] B. Settles, "Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 1467–1478.

[21] G. Druck, G. Mann, and A. McCallum, "Learning from labeled features using generalized expectation criteria," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 595–602.

[22] H. Raghavan, O. Madani, and R. Jones, "Active learning with feedback on features and instances," *Journal of Machine Learning Research*, vol. 7, no. Aug, pp. 1655–1686, 2006.

[23] "Jieba chinese text segmentation," Accessed March 26, 2019). [Online]. Available: https://github.com/fxsjy/jieba

[24] B. Xu, Y. Xu, J. Liang, C. Xie, B. Liang, W. Cui, and Y. Xiao, "Cn-dbpedia: a never-ending chinese knowledge extraction system," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2017, pp. 428–438.

[25] "Knowledge works," Accessed March 26, 2019). [Online]. Available: http://kw.fudan.edu.cn/

[26] C. Manning, P. Raghavan, and H. Schütze, "Term frequency and weighting," in *Introduction to information retrieval*, 1st ed. Cambridge university press, 2008, ch. 6.2, pp. 117–119.

[27] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[28] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 613–622.

[29] O. F. Zaidan, J. Eisner, and C. Piatko, "Machine learning with annotator rationales to reduce annotation cost," in *Proceedings of the NIPS* 2008 workshop on cost sensitive learning*, 2008, pp. 260–267.

[30] Wikipedia, "F1 score," Accessed March 26, 2019). [Online]. Available: https://en.wikipedia.org/wiki/F1_score

[31] N. Craswell, "Mean reciprocal rank," *Encyclopedia of Database Systems*, pp. 1703–1703, 2009.

[32] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning*. Springer, 1998, pp. 137–142.

[33] Y. Wang, "Interactive machine learning with applications in health informatics," Ph.D. dissertation, University of Michigan, Ann Arbor, 2018.

[34] X. Wilcke, P. Bloem, and V. De Boer, "The knowledge graph as the default data model for learning on heterogeneous knowledge," *Data Science*, vol. 1, no. 1-2, pp. 39–57, 2017.