Combining Open-domain and Biomedical Knowledge for Topic Recognition in Consumer Health Questions

Yassine Mrabet, PhD[1], Halil Kilicoglu, PhD[1], Kirk Roberts, PhD[2],
Dina Demner-Fushman, MD, PhD[1]
[1] Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD, USA
[2] University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

Determining the main topics in consumer health questions is a crucial step in their processing as it allows narrowing the search space to a specific semantic context. In this paper we propose a topic recognition approach based on biomedical and open-domain knowledge bases. In the first step of our method, we recognize named entities in consumer health questions using an unsupervised method that relies on a biomedical knowledge base, UMLS, and an open-domain knowledge base, DBpedia. In the next step, we cast topic recognition as a binary classification problem of deciding whether a named entity is the question topic or not. We evaluated our approach on a dataset from the National Library of Medicine (NLM), introduced in this paper, and another from the Genetic and Rare Disease Information Center (GARD). The combination of knowledge bases outperformed the results obtained by individual knowledge bases by up to 16.5% F1 and achieved state-of-the-art performance. Our results demonstrate that combining open-domain knowledge bases with biomedical knowledge bases can lead to a substantial improvement in understanding user-generated health content.

Introduction

Online resources are increasingly used by consumers to meet their health information needs [1]. According to recent surveys, one of three U.S. adults (35%) looks for information on a medical condition online[1], 30% found online health information helpful for them or someone they know, while only 3% found online information harmful[2]. The same study showed that 15% of Internet users posted questions, comments or information about health-related issues on the web.

The National Library of Medicine (NLM) receives health-related questions from consumers from all over the world. Most consumer health questions received are concerned with disease-related information, such as diagnosis, treatment, and prognosis. At NLM, we have been working towards a system that can automatically answer such questions using available resources. Our previous work focused on intermediate tasks such as question frame extraction [2], question decomposition [3], question type recognition [4], anaphora resolution [2], and spelling correction [5].

Much research in biomedical question answering has focused on answering well-formed clinical questions by professionals [6, 7]. However, consumer health questions differ widely in vocabulary and structure from professional questions. In a recent study, Roberts and Demner-Fushman [8] compared various collections of consumer health questions and professional questions and found that:

- Consumer health questions are closer to open-domain language models while professional questions are closer to medical language models
- Consumer health questions are more focused on medical problems than treatments and tests
- Consumer health questions tend to contain more sub-questions than professional questions.

These characteristics along with high rates of misspellings and ungrammatical sentences pose challenges for automatic question understanding, including tasks such as named entity recognition, and frame extraction. Consumer questions

[1] http://www.pewinternet.org/2013/01/15/health-online-2013/
[2] http://www.pewinternet.org/2011/05/12/social-media-in-context/

about medical problems are often centered on specific diseases, symptoms or treatments, which we refer to as the question topics in this paper. Identifying the topic of a question is a critical step in answering it. Consider the following question:

(1) *my question is this: I was born w/a esophagus atresia w/dextrocardia. While the heart hasn't caused problems,the other has. I get food caught all the time. My question is...is there anything that can fix it cause I can't eat anything lately without getting it caught. I need help or will starve!.*

Note that while two diseases are being mentioned in the question (*esophagus atresia* and *dextrocardia*), the topic of the question is only *esophagus atresia*. While the single disease assumption holds true to a large extent, there are questions with multiple topics, especially when the question is about the relationship between diseases. In the example below, both *megalocytic interstitial nephritis* and *malakoplakia* are question topics.

(2) *I'd like to learn more about megalocytic interstitial nephritis with malakoplakia.*

Question topic recognition can be viewed as an extension of named entity recognition, where the goal is to determine only the entities deemed central to answering the question. In this paper, we propose a novel approach to topic extraction that relies on multi-perspective and knowledge-based recognition of named entities. We use open-domain and biomedical knowledge bases to normalize named entities in consumer health questions. More precisely, we combine disambiguation results from DBpedia[9] and UMLS[10] to provide an initial list of candidate named entities. Starting from this set, we represent the problem of topic recognition as a binary classification task and extend the span of detected topic entities with generic rules. Our experiments show that our method improves the state-of-the-art performance and that the combination of knowledge bases outperforms substantially the individual results, demonstrating that open-domain knowledge bases can benefit information extraction from consumer health text.

Related Work

Duan et al. [11] designed a question analysis approach that tackles the recognition of both the question topic, defined as the main context or constraint of a question (e.g., "*Berlin*" in "*What is the population of Berlin?*") and the question focus, which they introduced as a descriptive feature of the question topic (e.g., "*population*" in the previous question). Their approach for the identification of question topic and focus is based on MDL (minimum description length) tree cut model. They first represent the question as a chain of base noun phrases, considered as topic terms. Each chain is ordered according to the term specificity in a set of questions and a tree of questions is built from the chains of topic terms. Their cut method then separates the tree into two sets of terms: less specific terms considered as focus and more specific terms considered as topic. While their method allows retrieving multiple topics for one question it requires a rich set of similar questions talking about the same topics, and small sets of questions or dissimilar questions won't lead to relevant specificity values for the topic terms. While such method could work with open domain questions, it is not adaptable to consumer health questions that tend to be highly heterogeneous in terms of topics, making it difficult to find rich sets of similar questions.

Other methods in open domain used search results to recognize the main entities in user questions. For instance, Carroll et al. [12] proposed a method where questions are first classified as *entity triggering questions* and then submitted to a search engine. The goal of this process is to find entities associated with the retrieved documents and compare them to the entities mentioned in the question. Candidate entities are then scored according to different factors including the proportion of retrieved documents associated with them or the number of their mentions in the snippets generated from these documents. The candidate entity with the best score is then selected as the main entity (topic) of the question and common features associated with the topic in the retrieved documents are presented to the user. The shortcoming of this method is that it is restricted to questions identified as entity-triggering, which are mostly short open-domain questions. Consumer health questions are more elaborate and often contain several named entities that do not represent the question topic (e.g., disease history or diagnostic tests used to identify the main disease), which makes it difficult to find documents that are relevant to the main topic.

Many studies tackled biomedical question answering [13]. Different approaches have been used for question analysis, ranging from keyword-based search [14] to the extraction of structured frames expressing the relations between entities [15]. Demner-Fushman and Lin[16] proposed an approach to recognize the primary medical problems in clinical questions using UMLS concepts belonging to the *Disorder* semantic group. This is done by ranking the list of concepts belonging to this category according to their occurrence/position in the analyzed text.

Roberts et al. [3] addressed topic recognition in the scope of the decomposition of consumer health questions. In their approach, candidate named entities were first identified with high recall using a lexicon-based method and then ranked using a support vector machine (SVM) classifier. A series of post-processing rules addressed the term boundaries. The method obtained good results (73.6 and 88.1 F1 score with exact and relaxed match, respectively); however, it did not address multi-topic questions, which represent a quarter of the questions in our dataset.

In general, few approaches tackled specifically the recognition of the main topic in biomedical question analysis. This can be explained by the fact that questions submitted by professionals tend to be concise and tend to have a regular structure which makes the task of topic recognition close to simple keyword extraction. This is not the case for non-expert consumer questions that are more complex to process due to the noise added by numerous peripheral entities.

Methods

Our approach to detect topics in consumer health questions is twofold. First, we extract candidate named entities with an unsupervised method. More precisely, we detect and normalize textual mentions with different knowledge bases then we perform named entity recognition using the semantic types associated with the knowledge bases concepts. Second, we classify the recognized entities as a question Topic or not. We call the overall normalization, named entity recognition and entity classification approach K_{map}.

A. Knowledge-based normalization

We use a concept normalization approach inspired by Mrabet et al. [17]. More precisely, we exploit TF-IDF search and optimization techniques to locate and disambiguate the textual mentions that refer to concepts in a knowledge base k_i. This modular process does not require learning corpora, which allows our approach to extend naturally with additional knowledge bases when needed. The implemented algorithm follows the steps outlined below:

1. Split the question into sentences and perform part-of-speech tagging[3].
2. For each sentence, select candidate textual mentions corresponding to a sequence of allowed part-of-speech tags (e.g., nouns, adjectives, adverbs).
3. Use each candidate textual mention as a keyword query to look up concepts in knowledge base k_i based on a TF-IDF search.
4. If no exact match is found between the candidate mention and the knowledge base concepts, the subsequence of words with best TF-IDF score is selected instead. The TF-IDF score is computed by the SolR indexer[4]. A minimum threshold is fixed to discard weak matches.
5. For ambiguous mentions (i.e., mentions that have more than one corresponding concept with the maximum TF-IDF score) disambiguation is performed according to global coherence: i.e., we select the knowledge base concept that has more relations with the concepts obtained from other textual mentions. This is done using integer linear programming and contextual search as described in Mrabet et al.[17].

[3] We used Stanford Core NLP for POS tagging.

[4] https://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html

Table 1: Example type mappings

Target types	DBpedia types	UMLS types
Disorder	dbo:Disease umbel-rc:AilmentCondition	Acquired Abnormality Cell or Molecular Dysfunction
Drug	umbel-rc:DrugProduct; dbo:Drug yago:Medicine103740161	Clinical Drug Pharmacologic Substance
Diagnostic	yago:DiagnosticTest105739043	Diagnostic Procedure

As a result of this process, using a different knowledge base will not necessarily lead to the same candidate textual mentions and will provide different perspectives for the same consumer question.

In the course of this study we are interested in the combination of open-domain knowledge bases and biomedical knowledge bases. We use DBpedia as an open-domain knowledge base [9]. DBpedia is built semi-automatically from Wikipedia infoboxes. For a given Wikipedia article, RDF[5] triples in the format <subject, predicate, object> are constructed using the article's title as subject and the infobox columns as a list of predicates and objects. In a final step, manual mappings were performed by the DBpedia community to normalize predicate names to RDF properties defined in a reference ontology[6]. The 2014 version of DBpedia includes more than 4 million concepts and 3 billion triples.

For the biomedical perspective we use UMLS[10] as a reference knowledge base. We leverage UMLS Semantic Network relations for concept-level disambiguation. In contrast with DBpedia, UMLS relations are not established facts but are defined between semantic types, indicating possible factual relations between the concepts belonging to the given semantic types. For example, the Semantic Network relation Disease or Syndrome-TREATS-Pharmacologic Substance licenses a relation between concepts Cold and Paracetamol. By expanding the subject and object using semantic types hierarchy, we obtain 4,895 relations linking 133 semantic types.

We indexed both knowledge bases with SolR[7] and used LpSolve[8] to solve the disambiguation problem by maximizing the number of relations between the selected concepts.

B. Named entity recognition

We consider all textual mentions successfully normalized in the previous step (i.e., mentions linked to only one knowledge base concept) as candidate named entities. The set of candidate concepts associated with a given mention m, according to a knowledge base k_i is denoted $C_{k_i}(m)$. In order to associate a category to a given named entity, we perform a mapping between the semantic types of its associated concept $c \in C_{k_i}$ and target recognition types (e.g., Disorder, Diagnostic Procedure). This mapping is based on simple regular expressions; for example, we look for the presence of the tokens disorder, disease, syndrome, illness, and pathology to recognize entities referring to medical problems in DBpedia. Table 1 presents some examples of exact mappings from DBpedia and UMLS. In the context of this study, we are primarily interested in the identification of entities referring to medical problems and use the corresponding regular expressions for type mapping.

In order to test the impact of the combination of dissimilar domain point-of-views, we merge the named entities obtained by normalizing the text with different knowledge bases. This merge extends simple union by merging candidate mentions m_i, m_j that have the same mapped semantic type and overlapping or contiguous text spans into a single mention m_{ij} such that: $C_B(m_{ij}) = \cup_{i=1}^{N} C_{k_i}(m)$. The text span of m_{ij} is the text segment covered by m_i and m_j. The main motivation of this procedure is to obtain higher precision, better text spans and to avoid partial matches.

For instance, in the question "*Are there any new research studies enrolling people with* ***carbamoyl phosphate*** ***synthase***

[5] https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/
[6] http://wiki.dbpedia.org/services-resources/ontology
[7] http://lucene.apache.org/solr/
[8] http://lpsolve.sourceforge.net/5.5/

1 deficiency?" the two candidate mentions "*carbamoyl phosphate synthase*" and "*synthase 1 deficiency*" are merged to form a single mention "**carbamoyl phosphate synthase 1 deficiency**?".

As we do not use learning corpora, the coverage of our approach is limited to the vocabulary defined in the knowledge bases. To overcome this limitation, we apply a rule-based method to obtain better text spans for the extracted entities as in Roberts et al. [3]. More precisely, we use three rules to extend the span of a named entity with (a) preceding adjectives and modifiers, (b) acronyms occurring directly after the entity and (c) generic keywords occurring directly after the entity (e.g., syndrome, disease, condition).

C. Entity Classification

In the last step of our processing, we approach the task of determining whether an entity is a question topic as a supervised classification task. Given an annotated corpus with gold topic mentions, we build automatically a training corpus for this classification task by considering each recognized named entity that overlaps a gold topic mention as a positive example and each entity with no overlap as a negative example. This translation can be applied to any given annotated corpus and does not restrict the scalability of our approach.

We used a support vector machine (SVM) classifier and experimented with various textual, lexical, and knowledge-based features. After empirical tests we kept only the features that have a minimum Pearson's correlation factor with the class label. The Pearson's factor indicates how much two variables are linearly correlated, its values range from -1 (total negative correlation) to 1 (total positive correlation) with 0 indicating no correlation. In our experiments we used a minimum absolute threshold value of 0.1. The first 16 features according to the Pearson's score are presented in Table 2, along with feature values for the classification of question (3). Disease entities recognized with DBpedia are highlighted, disease entities recognized with UMLS are underlined. In Table 2, candidate named entities are taken from the merge of both results.

(3) *Is there any evidence that* ***trauma*** *such as a* ***physical injury*** *i.e., neck* ***injury****, torn ligament, etc., can worsen McArdle's* ***disease****?*

In our experiments we used the LibLinear[9] implementation of support vector machines (SVM).

Evaluation and Discussion

A. Data

To evaluate our method we consider two corpora, one from the Genetic and Rare Disease Information Center (GARD) and the other from the U.S. National Library of Medicine (NLM). The GARD dataset consists of 1459 questions taken from a curated online question set[10] with answers created by NIH staff. Each question is associated with a genetic or rare disease (e.g., Beckwith-Wiedemann syndrome, SCOT deficiency, cold agglutinin disease), which is typically the topic of the question. This dataset is publicly available as part of the GARD question decomposition dataset[11].

The NLM dataset consists of 263 questions collected from questions submitted to NLM websites (e.g., MedlinePlus, PubMed) and manually classified as consumer health questions with at least one disease topic. The question writers are generally concerned with a particular medical problem, though not always a named disease (e.g., pain in right ear, difficulty walking), and are asking for general medical information (such as that available on MedlinePlus), examples (1), (2) and (3) were taken from the NLM dataset. The dataset contains identifying information and is thus not publicly available. A de-identified set of questions largely overlapping with this dataset is in preparation for public release. Table 3 presents more details on each corpus.

[9] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

[10] https://rarediseases.info.nih.gov/gard

[11] http://lhncbc.nlm.nih.gov/project/consumer-health-question-answering

Table 2: Main topic classification features for question 3.

Feature	**Description**	*trauma*	*physical injury*	*neck injury*	*McArdle's disease*
ST	UMLS semantic type (for mentions linked to UMLS entities).	Multiple types	Injury or Poisoning	Injury or Poisoning	Disease or Syndrome
$TFIDF1$	TF-IDF of entity in DBpedia.	11.63	18.33	11.38	9.33
$TFIDF2$	TF-IDF of entity in UMLS.	6.99	14.03	12.98	10.31
$Ambiguity$	relative number of entities sharing the maximum TF-IDF score from a target knowledge base. Average ambiguity is considered if an entity was detected by both DBpedia and UMLS.	25.65	0.00	12.50	0.00
$P-C1$	Relative character position of first occurrence.	0.21	0.34	0.51	0.86
$P-C2$	Relative character position of last occurrence.	0.21	0.34	0.51	0.86
$P-E$	Position in the set of entities.	0.25	0.5	0.75	1.00
$Freq.$	Entity frequency in the question.	1	1	1	1
Cl	Character length	6	15	11	17
N	Total nb. of entities in the question	4	4	4	4
T	Entity tokens number	1	2	2	3
W_1	Entity token 1 (W_1).	trauma	physical	neck	McArdle
L_1	Lemma of W_1.	trauma	physical	neck	McArdle
POS_1	Part-of-speech tag of w_1.	NNS	JJ	NNS	NNS
L_{-1}	Lemma of token preceding the entity (W_{-1}).	that	a	,	worsen
POS_{-1}	Part-of-speech tag of W_{-1}.	IN	DT	,	VB

Table 3: Corpus Statistics

Features	**GARD**	**NLM Requests**
Questions	1459	263
Multi-topic questions	46 (3.15%)	64 (24.33%)
Topic per Question	1.03	1.25
Tokens	46,964	17,848
Tokens per Question	32.18	67.86

On the NLM dataset, our NER process retrieved an average of 10.61 entities per question with DBpedia and 9.68 with UMLS. On the GARD dataset, our NER module retrieved 6.59 entities per question with DBpedia and 6.33 with UMLS.

B. Topic Recognition Results

To test our approach we perform a 10-fold cross validation to associate a class label (topic/none) to each candidate entity (the label is assigned to the entity when the latter is part of the 10% test split). Table 4 presents the results for topic recognition evaluated with 10-fold cross-validation on each corpus.

The merger of both knowledge bases performed substantially better than the individual knowledge bases (cf. Table 4) with an increase of 16.6% in F1 score on the NLM dataset and 2.1% on the GARD dataset for partial matches. For exact matches, the increase is of 3.5% on the GARD dataset and 10.6% on the NLM dataset. Besides the increased recall, this result can also be explained by the fact that our merge method is able to extend the spans of textual mentions from two annotation point-of-views; i.e., the open-domain perspective brought by DBpedia and the medical perspective from UMLS. The number of questions with multiple topics in the GARD dataset is not significant enough to draw reliable conclusions (only 3% of all questions), however we can observe that recall on these questions is above 70%. The NLM dataset contains more multi-topic questions (24.3%) and we can observe that recall on these questions

Table 4: Evaluation Results on Reference Topic Recognition (10-fold cross-validation)

Classifiers	GARD						NLM					
	Exact Match			Partial Match			Exact Match			Partial Match		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
$K_{map}(DBpedia)$	78.5	76.3	77.4	90.6	87.0	88.8	58.3	24.0	34.0	82.4	33.6	47.7
$K_{map}(UMLS)$	78.9	76.1	77.5	90.3	85.7	87.9	63.1	24.0	34.7	85.6	32.5	47.1
$K_{map}(Merge)$	**80.4**	**81.7**	**81.0**	**91.4**	**90.5**	**90.9**	**64.6**	**34.9**	**45.3**	**92.2**	**34.9**	**63.7**
Multi-topic only	90.7	71.0	79.6	95.9	73.0	82.8	76.4	45.2	56.8	83.6	49.5	62.2

is more than 10% better than the overall recall with a 12% improvement in precision for exact match, which shows that our approach has higher effectiveness on these questions. This may be explained by the fact that consumers who write multiple-topic questions are looking for a more-precise and fine-grained information requiring a more focused writing to express the link between the different topics, which eases the named entity recognition and classification process.

In Table 5, we compare our results on the GARD dataset with the the state-of-the-art approach, denoted QD[3], which uses a support vector machine over a set of semantic and word-level features to rank an initial set of candidate mentions provided by a lexicon constructed from UMLS. This approach also uses a set of rules (R_1) to extend the spans of the detected topic mentions (e.g., appending generic words such as “disease” or “syndrome” if they follow the detected entity in the text). The comparison should be taken with caution as we use a different set of rules to extend entity boundaries. In order to reach more accurate observations, we use both a subset of R_1 that can be applied to our system, denoted R'_1 and our full set of rules, denoted R_2. The K_{map} + R_2 results correspond to the full-method results presented in Table 4.

Table 5: Comparison with state-of-the-art approach on the GARD dataset.

Approach	GARD					
	Exact Match			Partial Match		
	P	R	F1	P	R	F1
QD	56.4	54.7	55.6	89.2	86.5	87.9
QD + R_1	74.8	72.5	73.6	89.5	86.8	88.1
K_{map}	70.0	68.3	69.1	**92.4**	84.9	88.5
K_{map} + R'_1	75.3	76.5	75.9	91.0	90.5	90.8
K_{map} + R_2	**80.4**	**81.7**	**81.0**	91.4	**90.5**	**90.9**

Using a the subset of rules R'_1 improved performance and the final set of rules R_2 led to an overall performance increase on both exact matching (+7.4% F1) and partial matching (+2.8% F1) over QD + R_1. The overall improvement with respect to QD can be explained by several factors:

- QD does not tackle multi-topic questions and extracts only one topic for a given question, which reduces recall. More generally, ranking-based methods do not provide a systematic solution to multi-topic questions. Even if a threshold can be fixed to get the N best entities, relevant extensions are needed in order to determine the number of topics in a question. Binary classification solves this problem as it processes each entity individually from both a local and contextual point of view regardless of the rank/importance of other entities in the question, which provides a natural way of selecting the number of entities to be considered as topic/important in a question.
- Vocabulary coverage: this factor is hard to estimate exactly. While our method used DBpedia as an open-domain knowledge base, it also did not have the same level of coverage on the UMLS part, as it annotates only nouns or noun phrases that correspond to UMLS concepts, while QD is based on an extended lexicon that is able to detect individual keywords and to annotate verbs and adjectives.

- Prior disambiguation: in K_{map} entities are disambiguated using knowledge base relations while QD uses only a lexicon-based method, which helps provide a more reliable set of candidate entities to the binary classifier.

We analyzed two random sets of 20 errors, respectively from the GARD dataset and from the NLM dataset on the basis of exact match evaluation. Table 6 presents the observed error types with percentages and examples.

Table 6: Error Types on Random subsets of GARD and NLM datasets

	GARD	
Error type	**Error %**	**Examples**
Entity Normalization	15%	Some acronyms like *APL* and *ARVC* were not disambiguated.
Entity not in the knowledge bases	15%	The system recognized only *agonizing disease* instead of *this very painful and agonizing disease*, or two foci: *hereditary neuropathy* and *liability to pressure palsy* instead of one single *hereditary neuropathy with liability to pressure palsy*.
Classification	45%	*autism* not classified as Topic in two questions.
Expansion Rules	15%	*multiple cerebral cavernous malformations (CCM)* instead of *cerebral cavernous malformations (CCM)*.
Misspelling	5%	The question mentions *Mycrobacterium fortuitum* instead of *Mycobacterium fortuitum*.
Benchmark Error	5%	*this disease* annotated as Topic instead of the two question topics *hypokalemic periodic paralysis* and *hyperkalemic periodic paralysis*
	NLM	
Entity not in the knowledge bases	25%	*sounding in my ear*, *liver is damaged*, *pain in my shoulder*, *growth on the neck*.
Classification	60%	*strokes*, *heart attacks*, *Nph*, *shingles* recognized as entities but not as Topic in some questions. *TX* wrongly classified as Topic.
Misspelling	15%	*canker*, *Thalassamia*, *hitiala hernia*.

Classification was the main cause of errors on both datasets, with a more noticeable impact on the NLM dataset, which is clearly harder to process due to the lack of training data (263 questions vs. 1459 in the GARD dataset) and to longer queries (NLM requests are two times longer than GARD requests in average, cf. Table 3). Other errors vary by dataset. In GARD, entity normalization was the second source of errors (15%) as some acronyms and expressions could not be disambiguated with our knowledge-based approach. At a similar rate, the lack of coverage (concepts) in the knowledge bases did not allow detecting some topics. The impact of coverage seemed to be more pronounced on the NLM dataset, where spatial expressions play an important role in problem descriptions (e.g., *pain in my shoulder*). Errors due to misspellings were also more frequent in the NLM dataset, which may be explained by the fact that these questions are not edited for spelling, as it was done for the GARD dataset.

C. Ablation study

In this section, we study the classification of every instance of recognized entities, as opposed to finding an overall topic of the question as described above. This internal evaluation is important as achieving high classification accuracy is required to assess the scalability of our approach in distinguishing topic entities from other entities. We also tested our classification method with each knowledge base individually, and then with the merger of the knowledge bases results. We report ROC Area and precision, recall and F1 score for the Topic class in Table 7.

Our classifier reached high ROC area on the GARD corpus (93.0%) but obtained a lower performance on the NLM consumer health questions dataset (76.2%). This is partly due to additional noisy entities in the NLM questions which are generally longer than GARD requests, and thus have a lower topic-to-named-entity ratio. NLM questions are also less grammatical and don't target genetic/rare diseases, which makes the topic term less likely to be a typical

Table 7: Evaluation Results for Binary Classification (10-fold cross-validation).

Classifiers	GARD				NLM Requests			
	ROC Area	P	R	F1	ROC Area	P	R	F1
$DBpedia$	92.6	92.8	**94.0**	**93.4**	73.5	65.6	61.3	63.4
$UMLS$	91.9	**93.0**	93.7	93.3	70.7	62.4	55.7	58.9
$Merge$	**93.0**	92.6	93.4	93.0	**76.2**	**70.3**	**61.5**	**65.6**

named entity (e.g., "*loss of vision*" vs. "*streiff syndrome*"). This makes it more difficult to detect Topic entities with position-based features and TF-IDF scores in the NLM dataset.

The combination of both knowledge bases obtained the best classification performance on both corpora. While higher recall was anticipated, higher precision can be partly explained by the merge heuristic that is performed at entity level which leads to better text spans. The different perspectives of the knowledge bases may also allow the SVM to reach more accurate hyperplanes.

To better understand the behavior of the classifier, we also studied the role of the main corpus features. Table 8 presents the Pearson's correlation score of the best features for both datasets.

Table 8: Pearson's correlation score for the best corpus features

Corpus Features	**GARD**	**NLM**
Character Position	.58	.31
Entity Position	.54	.37
Character Length	.45	.22
Ambiguity	.42	.15
Entity Nb.	.42	.27
Entity Token Nb.	.35	.15
Entity Word 1 POS	.32	.18
UMLS Semantic Type	.24	.10

Topic entities often occur at the beginning of consumer health questions which explains the impact of position-related entities on both datasets. Ambiguity was interestingly among the best features. High ambiguity values indicate that an entity is an overly general (open-)domain entity. If such entity is associated with the medical problem type it would be a relevant indicator for positive classification (like generic disease terms such as *syndrome*), if it is not associated with medical problems it can be a relevant indicator for negative examples. If we compare the features ranking between the two corpora, we note that the semantic type information is substantially more correlated with the class label in the GARD dataset, which suggests a more heterogeneous set of semantic types for topics in the NLM dataset. Ambiguity is also more highly correlated in the GARD dataset, which suggest a more regular use of generic disease terms. More generally, the tokens and lemmas are not among the best features which suggests that our approach may also be ported to other corpora as a baseline for topic recognition in corpora from other domains.

Conclusions

We presented a novel approach to recognition of topics in consumer health questions using both open-domain and biomedical knowledge bases. Our experiments showed that the combination of such knowledge bases leads to substantial improvement. Our binary classification approach was also able to tackle effectively multi-topic questions. Comparison on a standard benchmark also showed that our classification approach outperforms an SVM-based ranking approach. In future work, we plan to include additional knowledge bases to improve the recall of our method by taking into account spatial relations. We will also explore other semantic features such as WordNet synsets to improve the contextual knowledge for entity classification.

Acknowledgments

This work was supported in part by the Intramural Research Program of the NIH, National Library of Medicine and by an appointment of Y. Mrabet to the NLM Research Participation Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education. K. Roberts was supported by NLM grant 1K99LM012104.

References

[1] Tustin N. The role of patient satisfaction in online health information seeking. Journal of health communication. 2010;15(1):3–17.

[2] Kilicoglu H, Fiszman M, Demner-Fushman D. Interpreting consumer health questions: The role of anaphora and ellipsis. In: Proceedings of the 2013 Workshop on Biomedical Natural Language Processing; 2013. p. 54–62.

[3] Roberts K, Kilicoglu H, Fiszman M, Demner-Fushman D. Decomposing consumer health questions. BioNLP Workshop, ACL 2014. 2014;p. 29–37.

[4] Roberts K, Kilicoglu H, Fiszman M, Demner-Fushman D. Automatically Classifying Question Types for Consumer Health Questions. In: AMIA Annual Symposium Proceedings. vol. 2014. American Medical Informatics Association; 2014. p. 1018.

[5] Kilicoglu H, Fiszman M, Roberts K, Demner-Fushman D. An ensemble method for spelling correction in consumer health questions. In: AMIA Annual Symposium Proceedings; 2015. .

[6] Cairns BL, Nielsen RD, Masanz JJ, Martin JH, Palmer MS, Ward WH, et al. The MiPACQ clinical question answering system. In: AMIA Annu Symp Proc. vol. 2011; 2011. p. 171–80.

[7] Ni Y, Zhu H, Cai P, Zhang L, Qiu Z, Cao F. CliniQA: highly reliable clinical question answering system. In: MIE; 2012. p. 215–219.

[8] Roberts K, Demner-Fushman D. Interactive use of online health resources: A comparison of consumer and professional questions. Journal of the American Medical Informatics Association. 2016;15(5):1–41.

[9] Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes PN, et al. DBpedia-a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal. 2014;5:1–29.

[10] Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004;32(suppl 1):D267–D270.

[11] Duan H, Cao Y, Lin CY, Yu Y. Searching Questions by Identifying Question Topic and Question Focus. In: ACL; 2008. p. 156–164.

[12] Carroll MK, Lee JJ. Using an entity database to answer entity-triggering questions. Google Patents; 2015. US Patent 9,081,814.

[13] Athenikos SJ, Han H. Biomedical question answering: A survey. Computer methods and programs in biomedicine. 2010;99(1):1–24.

[14] Yu H, Cao Y. Automatically extracting information needs from Ad Hoc clinical questions. In: AMIA; 2008. .

[15] Fan RSJJY, Chua THCTS, Kan MY. Using syntactic and semantic relation analysis in question answering. In: Proceedings of the 14th Text REtrieval Conference (TREC); 2005. .

[16] Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. Computational Linguistics. 2007;33(1):63–103.

[17] Mrabet Y, Gardent C, Foulonneau M, Simperl E, Ras E. Towards Knowledge-Driven Annotation. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence; 2015. p. 2425–2431.