Automatic classification of doctor-patient questions for a virtual patient record query task

Leonardo Campillos Llanos **Sophie Rosset** **Pierre Zweigenbaum**
LIMSI, CNRS,
Université Paris-Saclay, Orsay, France
{campillos|rosset|pz}@limsi.fr

Abstract

We present the work-in-progress of automating the classification of doctor-patient questions in the context of a simulated consultation with a virtual patient. We classify questions according to the computational strategy (rule-based or other) needed for looking up data in the clinical record. We compare 'traditional' machine learning methods (Gaussian and Multinomial Naive Bayes, and Support Vector Machines) and a neural network classifier (FastText). We obtained the best results with the SVM using semantic annotations, but the neural classifier achieved promising results without it.

1 Introduction

Previous work on question classification has mostly been undertaken within the framework of question answering (hereafter, QA) tasks, where classification is but one step of the overall process. Other steps are linguistic/semantic question processing, answer retrieval and generation by integrating data; indeed, these make QA a different task to that of standard information retrieval. Biomedical QA (Zweigenbaum, 2003) has mostly focused on questions that aim to obtain knowledge to help diagnose or cure diseases, by medical doctors (Demner-Fushman and Lin, 2007) or by patients (Roberts et al., 2014b), or to obtain knowledge on biology (Neves and Leser, 2015). Clinical questions to obtain data from patient records have also been addressed (Patrick and Li, 2012).

Herein, we address a question classification task from a different perspective to existing research. Our task is set in a simulated consultation scenario where a user (a medical doctor trainee) asks questions to a virtual patient (hereafter, VP) (Jaffe et al., 2015; Talbot et al., 2016) during the anamnesis stage, i.e. the interview to the patient to obtain diagnostic information. Question types need accurate classification to search the data in the clinical record.

In this context, question classification has aimed at identifying detailed question types (Jaffe et al., 2015). In contrast, we consider a situation where we already have a rule-based question analysis system that classifies questions according to the semantic function or content (in order to restrict the search for data in the patient record and reply coherently). This strategy works well as long as questions remain within its specifications; other questions should be handled by a different strategy. What is needed in this context is a way to determine whether a given question should be transmitted to the rule-based system or to a fallback strategy. This is the goal of the present research, which is tackled as a binary classification task. Figure 1 is a schema of the processing steps we address in this work (note that we do not represent other stages such as dialogue management).

Guiding the processing of input questions is a common step in QA systems. Questions may be filtered through an upfront classifier based on machine-learning techniques, parsing (Hermjakob, 2001), regular expressions and syntactic rules, or hybrid methods (Lally et al., 2012). To achieve that, a question analysis process might precede, which may involve detecting lexical answer types, question targets or the question focus.

Our VP system relies on named entity recognition and domain semantic labels in the question analysis. The results we report seem to show that leveraging this semantic information was beneficial for the classification step. We also tested a neural method without the semantic information, and indeed did not achieve the best performance (despite having promising results). We suggest

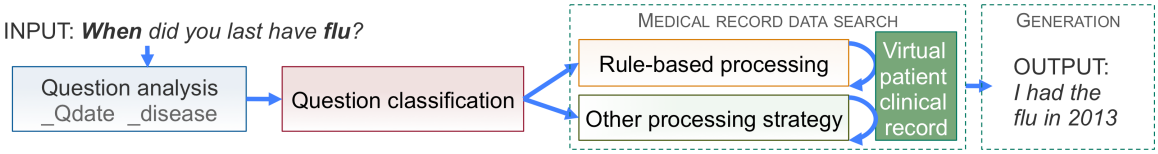

Figure 1: Schema of the question processing and search for data in the virtual patient record

that using a linear SVM classifier with the semantic information defined for the task (together with features such as token frequency and 3-grams) is a reliable technique for question triage in a rule-based system similar to the one we present.

We report results of the classification task and compare traditional machine-learning and a neural-network supervised classifiers (Bojanowski et al., 2016). We briefly review approaches to question classification (§2) and outline our task (§3). Then, we explain the sources of our data and describe them (§4). We present our methods (§5) and give our results (§6) then conclude (§7).

2 Related work

2.1 Question classification in medical QA

QA in medicine has extensively been researched. Approaches have addressed doctor questions on clinical record data (Patrick and Li, 2012), with the purpose of, among others, improving clinical decision support (Roberts et al., 2015; Goodwin and Harabagiu, 2016) or meeting the information needs of evidence-based medicine (EBM) practitioners (Demner-Fushman and Lin, 2007). EBM-focused approaches have relied on a specific knowledge framework, decomposing question topics into Problem/Population, Intervention, Comparison, and Outcome (PICO). Taxonomies of clinical question types already exist (Ely et al., 2000). (Patrick and Li, 2012) report an ontology and classification model for clinical QA applied to electronic patient notes.

Consumer health questions are another area of interest (Roberts et al., 2014b). Research has focused on classifiying the question according to the user (consumer or clinician) and question type (e.g focusing on the cause of a condition or the affected anatomic entity (Roberts et al., 2014a), or how consumer queries differ at the lexical, syntactic and/or semantic level (Slaughter et al., 2006; Roberts and Demner-Fushman, 2016).

We refer to (Athenikos and Han, 2010; Neves and Leser, 2015), respectively, for state-of-the-art reviews of QA for biomedicine and biology. Questions are generally classified into *Yes/No*, *Factoid/List* and *Definition/summary*.

Questions to a virtual patient have been addressed by mapping the user input to a set of predefined questions (Jaffe et al., 2015), as is done in a large subset of recent general-domain QA work which queries lists of frequently asked questions (FAQs) and returns their associated predetermined answers (Leuski et al., 2006; Nakov et al., 2016). Our setting is different in two ways: first, we do not rely on a FAQ but instead generate answers based on the question and on the contents of the virtual patient's record; second, we already perform fine-grained question classification with a rule-based system (Campillos et al., 2015), and aim to determine whether a given question should be referred to this rule-based strategy or deserves to be handled by a fallback strategy.

2.2 Approaches

Across the mentioned tasks, machine-learning methods for classifying questions range from hierarchical classifiers (Li and Roth, 2002) to linear support vector machines (SVM, hereafter) (Zhang and Lee, 2003). The benefit of using semantic features to improve question classification varies across experiments. For example, (Roberts et al., 2014a) reported improvements when classifying a dataset of consumer-related topics. They used an SVM with combinations of features including semantic information, namely Unified Medical Language System® (Bodenreider, 2004) Semantic Types and Concept Unique Identifiers. For their part, (Patrick and Li, 2012) used SNOMED categories. They reported improvements in classification through models including this type of feature, but not systematically. The type of the semantic information used in each task might explain these results. The impact of using semantic features is a point we explore in the present work in the context of questions to a virtual patient.

Neural network representations and classifiers are more and more applied to natural language

processing (Bengio et al., 2003; Collobert et al., 2011). Word embeddings—i.e. vector representations of words—allow the prediction of a word according to the surrounding context, and vice-versa. New research questions are being raised with regard to current architectures (Mikolov et al., 2013; Pennington et al., 2014; Goldberg, 2016), parameters (e.g. vector dimension or window size), hyperparameters or the effect of input data.

The latest models include subword information in word embeddings, encoding both n-grams of characters and the standard occurrence of words (Bojanowski et al., 2016). There is a growing interest in research on word embeddings for sentence classification (Kim, 2014; Zhang et al., 2016) and question classification (Mou et al., 2015). However, a far as we know, a neural network classifier using subword information has not yet been tested on a medical question classification task. This is another point we explore herein.

3 Task description

We classify questions into those that a rule-based dialogue system can process, and those needing a supplementary method. Table 1 gives examples of these two classes of questions, and shows the semantic annotation performed in our task. A rule-based system is to be favoured to maximize precision, but developing rules for any question type is not feasible in the long term. Thus, we need a classifier to distinguish which questions should be processed through our rules and which should resort to another strategy. Those *rule-based processing strategy* (*RBPS* hereafter) types of question are thought to have specific patterns (e.g. recurrent n-grams, question roots or domain semantic labels), which make it possible to formalise rules.

Example of questions	Strategy	Semantic annotation
Do you cough every day ?	Rule-based	SYMPTOM, FREQUENCY
Are your parents still alive ?	Other	FAMILYMEMBER

Table 1: Examples of questions and classes

In our system, rules are formalised based on the semantic annotation of questions.[1] For example, a rule processing the combination of SYMPTOM and FREQUENCY labels interprets the input as a query on the frequency of a symptom. Accordingly, the VP agent will answer with a fixed type of reply instantiated with the corresponding data in the record. We hypothesize that questions not fitting this scheme will require some *other processing strategy* (*OPS* hereafter), be it statistical, neural or machine-learning-based techniques, to search data in the record (or to reply adequately when data are not available).

[1]The semantic labels we use encode domain data (DISEASE), miscellanea (e.g. time or quantity) and question type or tense: e.g. QPASTYESNO (Campillos et al., 2016).

4 Data sources and preparation

4.1 Data sources

We collected French language questions from books aimed at medical consultation and clinical examination (Bates and Bickley, 2014; Epstein et al., 2015), as well as resources for medical translation (Coudé et al., 2011; Pastore, 2015).[2] We also collected questions from 25 transcribed doctor-patient interactions performed by human standardized patients (i.e. actors simulating medical consultations).

4.2 Additional data creation

The purpose of collecting the corpus is to train health dialogue systems aimed at simulating a consultation with virtual patients. There is a growing interest of research groups towards integrating Natural Language Interaction (NLI) features in medical education simulation systems (Hubal et al., 2003; Stevens et al., 2006; Kenny et al., 2008; Jaffe et al., 2015; Talbot et al., 2016).

Due to the lack of availability of questions, a subset of data was generated automatically by using question formation templates, semantic labels and resources from the UMLS. An example of template is *Do you suffer from* SYMPTOM *in your* ANATOMY*?*. There, the label SYMPTOM is replaced automatically with symptom terms (e.g. *pain* or *tingling*), and ANATOMY, with anatomic entities (e.g. *leg* or *arm*). We also generated automatically paraphrases of questions through a list of paraphrase patterns (e.g. *can you* → *are you able to*). These procedures allowed us to increase the corpus data, making up around 25% of the total number of questions. Of note is that we did not increase the corpus with more generated questions in order to avoid getting a too artificial dataset. Table 2 provides statistics on the experimental data.

4.3 Data preparation

We processed each question with our VP dialogue system (Campillos et al., 2015). Then, we manu-

[2]http://anglaismedical.u-bourgogne.fr/

Questions		RBPS	OPS	Total
Original		1,607	825	2,432
Generated		510	328	838
Total		2,117	1,153	**3,270**
Words	Tokens	15,276	10,299	**25,575**
	Types	3,470	2,624	**4,985**
	Mean	7.21	8.93	**7.82**
	Stdev	2.68	3.35	**3.04**
	Minimum	1	2	**1**
	Maximum	20	27	**27**
Sem. labels	Tokens	6,816	3,375	**10,291**
	Types	111	90	**119**
	Mean	3.22	3.01	**3.15**
	Stdev	1.30	1.59	**1.41**
	Minimum	0	0	**0**
	Maximum	11	11	**11**

Table 2: Distribution of experimental data (*stdev* = standard deviation)

ally labelled the output of question analysis, based on our knowledge of the dialogue system, into questions that should be processed by rule-based processing (*RBPS*) and questions requiring some other processing strategy (*OPS*). Specifically, we labelled as *RBPS* those questions with correct replies through the rule-based dialogue manager, or those questions for which the system has rules, but did not understand the questions or produced incorrect replies due to processing errors. We labelled as *OPS* the remaining questions that were not understood by the system or had wrong replies.

We split our corpus into 80% training and 20% test data (respectively, 2616 and 654 questions of both types). We performed 10-fold cross-validation on the training set for the non-neural classifiers, then applied the model to the test set.

5 Methods

We carried out tests with a linear support vector machine classifier and two Naive Bayes classifiers (Gaussian and Multinomial; from here on, respectively, Gaussian NB and Multinomial NB). We used Scikit-learn v0.18 (Pedregosa et al., 2011); the SVM used the LinearSVC implementation based on liblinear, one versus the rest scheme.

The combination of features used were inspired by (Roberts et al., 2014a). We used four sources of information:

1. The question Q itself, i.e., morphological and lexical features:

- Token and frequency in Q (TK)
- Question root (QR): the three first words of Q
- Three-character-grams (3CG) and frequency
- Three-grams (3G) and their frequency
- Number of words in Q (WC)
- Minimum, maximum and average word length in Q (WL)

2. The relation of Q to system knowledge, i.e., the term is found in the core system lexicon:

- Out-of-vocabulary words (NIL): terms in Q not found in system lexicon

3. Word representations computed from an external corpus:

- Average word embeddings of words in Q (WE). We used pre-trained word vectors (see below) with the best combination of parameters we tested (window=10, vector dimension=100, negative samples=10, learning rate=0.1, sampling threshold=1-e4). We only used this feature for the SVM classifier.

4. Annotations produced by the question analysis component of our dialogue system:

- Semantic annotation of Q (SEM)

We also tested the neural method implemented in FastText (Joulin et al., 2016). An extension of word2vec (Mikolov et al. 2013), FastText associates n-grams of words and/or characters to learn word vectors. It is a shallow neural model relying on a hidden layer, where a sentence is represented by averaging the vector representations of each word. This text representation is then input to a linear classifier (a hierarchical softmax, which reduces computational complexity). As our data were scarce, we used word vectors pretrained in a large domain corpus from the European Medicines Agency,[3] which amounts to more than 16 million tokens after tokenization. Several parameter values were tested: window size of 2, 4, 6, 8 and 10, vector dimension of 50, 100 and 300, use of 3-grams or 3-character-grams, number of negative samples (5, 10 or 20), learning rate (0.1 and 0.05) and sub-sampling threshold (1e-3 and 1e-4). We only tested the skip-gram architecture since it has

[3] http://opus.lingfil.uu.se/EMEA.php/

been observed to yield better results (Chiu et al., 2016). The minimum word count was fixed to 1, given the scarcity of our labelled data. We did not use semantic annotation to create word vectors.

6 Results and discussion

Table 3 breaks down our results (reported as F1-score) in the training set (top of the table) with different parameter combinations and non-neural classifiers. The weighted average F1-score was computed based upon both F1-scores of classifying *RBPS* and *OPS* types of questions. The best combinations of parameters found in the training set were applied to the test set; their results are placed at the bottom of the table. Note that a baseline method making a majority class decision would categorize each question as *RBPS*: since the proportion of *RBPS* is 0.647, its weighted average F1-score would be $0.647^2 = 0.419$.

The SVM classifier outperformed the other classifiers and the neural classifier. In all combinations of features used with non-neural methods, the use of semantic labels improved question classification. Multinomial NB obtained better results than Gaussian NB. Results with the best combinations of features and Multinomial NB gave similar results to those yielded by the neural method.

In such a small dataset and constrained task, the use of word embeddings as feature did not improve classification performance. This could be due to the data used for pre-training word embeddings. Despite being related to the domain, the nature of texts used for pre-training vectors is different to that of a clinical consultation context. Using the combination of token/frequency and semantic annotation together with another feature provided the highest results (or almost the highest). The use of 3-character-grams, word length or word count contributed to good classification, but their benefit was not strong, nor is it clear which feature was more relevant. Using 3-grams seems to be the exception: the best combination of parameters—as it improved results in all models—is 3-grams, semantic labels and token/frequency. Not shown in the table, when semantic labels are not used, the other features did not improve classification in our task (except 3-grams with Gaussian NB).

We note that the F1-scores obtained on the test set are similar to that obtained by cross-validation on the training set: the system did not overfit the training data.

The fact that we used a subset of generated questions from patterns could be argued as a bias. However, we tested the above models in a subset of 2,282 questions without any generated sentence, and the models and classifiers had similar results (but lower F1-scores). We again obtained the best results (avg. F1=0.81) with Linear SVC, with models using semantic features with or without all other parameters (e.g. QR+TK+WL+WC+SEM+3G+NIL and TK+SEM+QR+WC). We also tested the same combinations of features in Linear SVC with and without computing term frequency-inverse document frequency (tf-idf), and also a Logistic Regression classifier (with and without tf-idf). For each group of parameters, results were similar to those yielded by Linear SVC (which does not use tf-idf).

As for the neural method, Table 4 reports our results. The F1-score was computed based on precision and recall of the top ranked label (precision and recall @1). The best result was an average F-score of 0.812 (window of 10, vector dimension of 100, negative sampling of 10, learning rate of 0.1 and sampling threshold of 1-e4). We achieved similar results by modifying parameters (e.g. window of 6 or 8, vector dimension of 50, or use of 3 grams). Interestingly, using both 3 grams and 3-character-grams tended to lower performance.

We can draw two observations from our results. First, we find it beneficial leveraging the semantic information used for question analysis at the classification step. This could be a hint for developping QAs in a similar task and restricted domain to the one here presented. That is, the question analysis and classification steps for a similar rule-based system would need to build on a comprehensive semantic scheme permeating both rule development, entity type annotation and question triage. This is what seems to explain our lower results obtained when semantic features were not used in with machine-learning classifiers and the neural method. Indeed, (Jaffe et al., 2015) also reported an error reduction in question classification when domain concept-based features were used in the question classifier for their VP system.

Second, we found necessary to complement the neural approaches in this restricted task with natural language processing techniques to raise the classification performance. We trained a large amount of data for generating word embeddings (to use them as features for the LinearSVC classi-

TRAINING			
Parameters	**Linear SVC**	**Gaussian NB**	**Multinomial NB**
TK	**0.798**	0.573	0.783
3G	0.766	**0.678**	**0.787**
3CG	0.752	0.539	0.751
SEM	0.746	0.389	0.676
QR	0.679	0.427	0.670
WE	0.616	-	-
WC	0.611	0.554	0.612
WL	0.519	0.467	0.576
TK+SEM	**0.839**	0.595	**0.805**
TK+3G	0.814	**0.729**	0.794
TK+SEM+3G	**0.861**	**0.741**	**0.815**
TK+SEM+QR	0.844	0.657	0.802
TK+SEM+WC	0.841	0.596	0.803
TK+SEM+NIL	0.839	0.595	0.805
TK+SEM+3G+NIL	**0.862**	0.741	**0.815**
TK+SEM+3G+WC	**0.858**	**0.742**	0.809
TK+SEM+QR+WC	0.843	0.659	0.797
TK+SEM+3G+3CG	0.834	0.756	0.796
TK+SEM+QR+WC+WL	0.844	0.693	0.800
TK+SEM+QR+WC+WL+NIL	0.844	0.693	0.800
TK+SEM+3G+QR+WC+WL+NIL	**0.860**	**0.763**	**0.811**
TK+SEM+3CG+QR+WC+WL+NIL	0.816	0.701	0.781
TK+SEM+3G+QR+WC+WE+WL+NIL	**0.862**	-	-
TK+SEM+3G+QR+WC+WL+3CG+NIL	0.840	0.764	0.795
TEST			
TK+SEM+3G	**0.866**	**0.765**	**0.817**
TK+SEM+3G+WC	**0.871**	**0.766**	0.806
TK+SEM+3G+NIL	**0.866**	0.765	**0.817**
TK+SEM+3G+QR+WC+WL+NIL	**0.870**	0.759	**0.810**

TK: token; SEM: semantic labels; WL: maximum, minimum and average word length; WC: word count; QR: question root (3 first words); 3G: 3-grams; 3CG: 3-character-grams; NIL: word not in lexicon

Table 3: Avg. F1 of non-neural classifiers with the best tested features in training and test sets

WS	DIM	GR	CHGR	NEG	LR	SAMP	Avg F1
10	100	0	0	10	0.1	1-e4	**0.812**
8	50	3	0	20	0.1	1-e4	0.804
8	100	0	3	20	0.1	1-e4	0.803
6	50	0	3	10	0.1	1-e4	0.803
10	50	0	3	10	0.1	1-e4	0.800
2	50	3	3	20	0.1	1-e4	0.800
4	300	0	0	20	0.05	1-e3	0.792
10	300	0	3	10	0.05	1-e4	0.789

WS: window size; DIM: vector dimension; GR: n-grams; CHGR: character-grams; NEG: number of negative samples; LR: learning rate; SAMP: sampling threshold

Table 4: Results of the best tested models (neural approach)

fier) and also used a neural model to classify questions. However, our results agree with the observation that restricted-domain QA is less affected by data-intensive methods, but depend on refined language processing methods (Mollá and Vicedo, 2007)—in this type of system, accurate semantic annotation. On the other hand, the neural method seems promising in this kind of classification task, and how to use domain semantic information with it requires further exploration, in line with current works (Yu et al., 2016). We also need to pretrain vectors on domain data of different nature (e.g. clinical records) to confirm our results. Finally, other methods for computing vector representations of sentences deserve to be explored.

7 Conclusions

For the task of optimizing question processing in a VP natural language system, we reported the improvement of using the semantic information in the question analysis step as a feature for question classification. This is likely due to the idiosyncrasy of our task, where the dialogue system makes use of semantic rules for processing input questions. We are nonetheless interested in confirming to which extent reusing semantic information from the question analysis would benefit the classification step in QA systems for other tasks and domains. Anyhow, the neural method here tested yielded promising results for similar classification tasks. Other approaches to test might be including semantic annotation to generate vector representations of questions, pretraining word vectors on clinical record data, and using information from the VP clinical record as another source of features for classification.

Acknowledgments

The Société d'Accélération de Transfert Technologique (SATT) Paris-Saclay funded this research. The authors kindly appreciate the anonymous reviewers' comments and thank Julien Tourille for the helpful discussions on ScikitLearn.

References

Sofia J Athenikos and Hyoil Han. 2010. Biomedical question answering: A survey. *Computer methods and programs in biomedicine* 99(1):1–24.

Barbara Bates and Lynn S Bickley. 2014. *Guide de l'examen clinique-Nouvelle édition 2014*. Arnette-John Libbey Eurotext.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3(Feb):1137–1155.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research* 32(Database issue):D267–270.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *CoRR* abs/1607.04606. http://arxiv.org/abs/1607.04606.

Leonardo Campillos, Dhouha Bouamor, Éric Bilinski, Anne-Laure Ligozat, Pierre Zweigenbaum, and Sophie Rosset. 2015. Description of the PatientGenesys dialogue system. In *Proceedings of the SIGDIAL 2015 Conference*. Association for Computational Linguistics, pages 438–440.

Leonardo Campillos, Dhouha Bouamor, Pierre Zweigenbaum, and Sophie Rosset. 2016. Managing linguistic and terminological variation in a medical dialogue system. In *Proceedings of LREC 2016, Portoroz, Slovenia, 24-27 May 2016*. pages 3167–3173.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. *ACL 2016* pages 166–174.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

Claire Coudé, Franois-Xavier Coudé, and Kai Kassmann. 2011. *Guide de conversation médicale - français-anglais-allemand*. Lavoisier, Médecine Sciences Publications.

Dina Demner-Fushman and Jimmy Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics* 33(1):63–103.

John W Ely, Jerome A Osheroff, Paul N Gorman, Mark H Ebell, M Lee Chambliss, Eric A Pifer, and P Zoe Stavri. 2000. A taxonomy of generic clinical questions: classification study. *British Medical Journal* 321(7258):429–432.

Owen Epstein, David Perkin, John Cookson, and David P. de Bono. 2015. *Guide pratique de l'examen clinique*. Elsevier Masson, Paris.

Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* 57:345–420.

Travis R Goodwin and Sanda M Harabagiu. 2016. Medical question answering for clinical decision support. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, pages 297–306.

Ulf Hermjakob. 2001. Parsing and question classification for question answering. In *Proceedings of the Workshop on Open-domain Question Answering - Volume 12*. Association for Computational Linguistics, ODQA '01, pages 1–6.

Robert C Hubal, Robin R Deterding, Geoffrey A Frank, Henry F Schwetzke, and Paul N Kizakevich. 2003. Lessons learned in modeling virtual pediatric patients. *Studies in Health Technology and Informatics* pages 127–130.

Evan Jaffe, Michael White, William Schuler, Eric Fosler-Lussier, Alex Rosenfeld, and Douglas Danforth. 2015. Interpreting questions with a log-linear ranking model in a virtual patient dialogue system. In *Proc. of the 10 Workshop on Innovative Use of NLP for Building Educational Applic.*. pages 86–96.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* http://arxiv.org/abs/1607.04606.

Patrick Kenny, Thomas D Parsons, Jonathan Gratch, and Albert A Rizzo. 2008. Evaluation of Justina: a virtual patient with PTSD. In *International Workshop on Intelligent Virtual Agents*. Springer, pages 394–408.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the EMNLP 2014, October 25-29, 2014, Doha, Qatar*. pages 1746–1751.

Adam Lally, John M. Prager, Michael C. McCord, Branimir Boguraev, Siddharth Patwardhan, James Fan, Paul Fodor, and Jennifer Chu-Carroll. 2012. Question analysis: How watson reads a clue. *IBM Journal of Research and Development* 56(2:1):1–14.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, Stroudsburg, PA, USA, SigDIAL '06, pages 18–27.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pages 1–7.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* https://arxiv.org/abs/1301.3781.

Diego Mollá and José Luis Vicedo. 2007. Question answering in restricted domains: An overview. *Computational Linguistics* 33(1):41–61.

Lili Mou, Hao Peng, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2015. Discriminative neural sentence modeling by tree-based convolution. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *EMNLP*. Association for Computational Linguistics, pages 2315–2325.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. Semeval-2016 task 3: Community question answering. *Proceedings of SemEval* pages 525–545.

Mariana Neves and Ulf Leser. 2015. Question answering for biology. *Methods* 74:36–46.

Flicie Pastore. 2015. *How can I help you today? Guide de la consultation mdicale et paramdicale en anglais*. Ellipses, Paris.

Jon Patrick and Min Li. 2012. An ontology for clinical questions about the contents of patient notes. *J. of Biomedical Informatics* 45(2):292–306.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12(Oct):2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–1543.

Kirk Roberts and Dina Demner-Fushman. 2016. Interactive use of online health resources: a comparison of consumer and professional questions. *Journal of the American Medical Informatics Association* 23(4):802–811.

Kirk Roberts, Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. 2014a. Automatically classifying question types for consumer health questions. In *AMIA Annual Symposium*.

Kirk Roberts, Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. 2014b. Decomposing consumer health questions. In *BioNLP Workshop*. pages 29–37.

Kirk Roberts, Matthew Simpson, Dina Demner-Fushman, Ellen Voorhees, and William Hersh. 2015. State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track. *Information Retrieval Journal* .

Laura A Slaughter, Dagobert Soergel, and Thomas C Rindflesch. 2006. Semantic representation of consumer questions and physician answers. *International journal of medical informatics* 75(7):513–529.

Amy Stevens, Jonathan Hernandez, Kyle Johnsen, Robert Dickerson, Andrew Raij, Cyrus Harrison, Meredith DiPietro, Bryan Allen, Richard Ferdig, Sebastian Foti, et al. 2006. The use of virtual patients to teach medical students history taking and communication skills. *The American Journal of Surgery* 191(6):806–811.

Thomas B Talbot, Nicolai Kalisch, Kelly Christoffersen, Gale Lucas, and Eric Forbell. 2016. Natural language understanding performance & use considerations in virtual medical encounters. *Medicine Meets Virtual Reality 22: NextMed/MMVR22* 220:407.

Zhiguo Yu, Trevor Cohen, Elmer V Bernstam, and Byron C Wallace. 2016. Retrofitting word vectors of MeSH terms to improve semantic similarity measures. *EMNLP 2016* page 43.

Dell Zhang and Wee Sun Lee. 2003. Question classification using support vector machines. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, pages 26–32.

Ye Zhang, Stephen Roller, and Byron C. Wallace. 2016. MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification. In *HLT-NAACL*.

Pierre Zweigenbaum. 2003. Question answering in biomedicine. In Maarten de Rijke and Bonnie Webber, editors, *Proc Workshop on Natural Language Processing for Question Answering, EACL 2003*. ACL, Budapest, pages 1–4.