Generic Question Classification for Dialogue Systems

Marine Troadec, Matthis Houlès, Philippe Blache

Generic Question Classification for Dialogue Systems

Marine Troadec, Matthis Houlès, and Philippe Blache

LPL-CNRS, ILCB, Aix-en-Provence, France
blache@ilcb.fr

Abstract. We present in this paper a new classification approach for identifying questions during human-machine interactions and more specifically in dialogue systems. The difficulty in this task is first to be domain-independent, reusable whatever the dialogue application and second to be capable of a real time processing, in order to fit with the needs of reactivity in dialogue systems. The task is then different than that of question classification usually addressed in question-answering systems. We propose in this paper a hierarchical classifier in two steps, filtering first question/no-question utterances and second the type of the question. Our method reaches a f-score of 98% for the first step and 97% for the second one, representing the state of the art for this task.

Keywords: Question classification · Dialogue systems · Hierarchical classification.

1 Introduction

Question classification remains a difficult task depending on the application. A lot of work have been done in the domain of *question answering*, in particular in the framework of TREC [23]. In this case, the problem consists in finding a correspondence between questions and answers that are usually textual elements (sentences, summaries, etc.). Machine learning is very efficient for this type of problem [12, 10, 25], even though some rule-based approaches have also shown interesting results [14, 6, 15]. In this type of application, the problem consists in identifying fine-level class labels (and as a consequence a large number of classes) on a semantic basis, such labels being used as *keys* for accessing the answers [10].

The problem is different in the context of *dialogue systems*, in particular applications offering the possibility of a free conversation, not limited to question answering. In this case, recognizing questions during the dialogue becomes crucial: whatever the domain, the system needs to identify whether the user's input is a question or not, in order to answer in real time. Questions are therefore to be identified in a very accurate and robust way (including in comparison with other dialogue acts) [2, 21]. Moreover, dialogue systems require techniques as independent as possible from the context, in particular because of lack of resources: dialogues are usually task-oriented and in most of the cases, no adequate dataset with questions and answers relevant to the task (and/or to the language) exists.

This is the reason why dialogue system architectures usually rely on separate understanding and generation modules. In this case, instead of identifying the label of the question corresponding more or less directly to the index of the answer, the problem consists in analyzing more precisely the question by determining its type and its contents. This information makes it possible in a second step to access the dialogue knowledge database from which the answer is generated. As a consequence, the classification task may focus on the question's types. Only few works exist proposing a generic recognition of questions and their types.

This paper first presents an overview of the different approaches to question classification. We describe in particular the different questions taxonomies (identifying the classes to be classified), the datasets, the features and the methods proposed in the literature. We present then our approach aiming at building a generic classifier based on a linguistically motivated question typology. A hierarchical classification in two steps is proposed and experimented with different feature combinations, reaching the state of the art. This work is applied to French (even though the methods are language-independent).

2 Question taxonomies

Different typologies are used for question classification. The works focusing on question answering generally rely on the 2-levels taxonomy proposed in [10], making it possible to develop a hierarchical classification. First, a *coarse* level of the taxonomy comprises 6 classes corresponding to a typology of the answer: *abbreviation, entity, description, human, location* and *numeric value*. Second, the *fine* level proposes a set of 50 fine-grained classes. This taxonomy is based on the semantic type of the answer, obtained from the 500 questions of TREC10 [10]. This taxonomy is very efficient, including for comparable datasets and tasks, even though its reusability for more generic purposes could be questionable: in this type of approach, it is necessary to develop specific taxonomies for each different semantic domain.

Our work focuses on question classification for generic dialogue systems. Such applications need first to identify whether or not the user's utterance is a question and second the type of the question. Here, the answer is not given (at the difference with question-answering systems), but built thanks to specific techniques using a knowledge base. On the opposite, question-answering systems relies on the identification of the type of the semantic content of the question, which will be used as a key for retrieving the answer. Needs are then totally different depending on the application.

We propose, instead of such a semantic classification, a question typology based on morpho-syntactic characteristics more adequate for answering to the needs of a generic question recognition. The first typology level relies on the classical distinction between "*yes-no*" and "*wh-*" questions (corresponding to a distinction between closed-ended and open questions). In a dialogue system, *yes-no questions* typically concern the state of the user ("*Do you feel good this morning?*"), comprehension ("*Do you have questions?*"), social interaction ("*Do*

you want us to call somebody?"). In terms of morpho-syntax, such questions are formally marked by specific verbs (auxiliary, modal) and/or specific constructions (subject-verb inversion), depending on the language. In the perspective of classification, this means that morpho-syntactic features will play an important role. Note that we do not take into account in this study other modalities also playing an role in marking questions, typically prosody, because of the real time constraint: reducing the number of features and modalities to take into account also reduces the processing time.

On their side, *wh-questions* are strongly associated with a closed class of interrogative markers such as, for French, *qui, que/quoi, quand, où, pourquoi, comment, quel, à qui, combien*. In this case, lexical features will also play an important role for the classification. Also, the interrogative particles provide an important information about the type of the answer (see table 2), which is of great help in the response generation. Note that a partial correspondence exists between these question types and the coarse level classes proposed in [10], as illustrated in table 2.

Interrogative particle	**Question type**	**Correspondence with *Li et al.***
qui; à qui (who)	`Person`	Human
que/quoi; quel (what)	`Object`	Entity
quand (when)	`Time`	Description
où (where)	`Location`	Location
pourquoi (why)	`Event`	Description
comment (how)	`State`	Description
combien (how much/many)	`Quantity`	Numeric value

Table 1. *Correspondence between interrogative particles and question types in French*

3 Related Work

Question classification in the context of question answering requires specific datasets made of question/answer pairs. Most of such resources has been elaborated for English (even though different projects and evaluation campaigns such as TREC also have a multilingual goal). On the opposite, in the context of dialogue systems (which are mainly task-oriented), finding adequate question/answer corpora for the application domain is more complex or even impossible. We are faced in this case with the question of resources and when they exist, of their size, requiring specific processing methods. We propose in this section to review the main approaches used for question classification, independently from the application, in order to identify techniques that could be domain-independent.

Most of question classification works refer to the seminal approach developped in [10, 9, 11]. This approach relies on a specific taxonomy and a hierarchical classification based on a two-level organization (6/50 classes). The dataset

contains 5,500 questions, the feature space made of 6 features among which 4 are syntactic (*words, pos, chunks, head chunks*) and 2 semantic (named entities, semantically related words). The classification accuracy reaches 98.80% precision for coarse classes with all the features and 95% for the fine classes. This experiment also shows the importance of the last semantic feature (related words), in comparison with the others. Interestingly, this work also shows that there is almost no difference between the hierarchical and the flat classifier.

On their side, [17] considered two classes: *open* and *closed* questions. They use a set of syntactic rules to determine the *closed* question class and a classifier for the *open-ended*. This thematic classification was also used by [7] for the development of a chatbot for medical students. The questions are classified into 359 different labels, which offers a straightforward method for identifying the answers. Interestingly, only 8 of these labels alone represent 20% of the themes.

In general, the most frequent features used in these approaches are based on word forms, encoded as bag of word, n-grams and tf-idf [8]. On top of them, we can also find morpho-syntactic and semantic features such as POS n-grams, chunks, named entities, related words, wordnet synsets [9, 11, 18, 24, 1]. Table 2 summarizes different works with the type of the classifier, the features and the results.

Work	Classifier	Features	Accuracy	
			Coarse	Fine
Li & Roth (2002)	SNoW	U+P+HC+NE+R	91.0%	84.2%
Li & al. (2008)	SVM+CRF	U+L+P+H+HY+NE+S	-	85.6%
Loni & al. (2011)	Linear SVM	U+B+WS+H+HY+R	93.6%	89.0%
Mishra & al. (2013)	Linear SVM	U+H+HY+WS+QC	96.2%	91.1%
Margolis & al. (2011)	SCL	U+B+T+H	-	85.9%
Duggenpudi & al. (2019)	LSTM	NE	-	99.32%
Kumar & al. (2016)	SVM	U+B+T+H+P+HC	-	87.65%
Madabushi & al. (2002)	None	W	-	97.2%
O'Sheah & al. ()	RFC	SFWC	-	98.51%
Somnath & al. (2013)	DT+AdaBoost	WH+L+H+WS+P+NE+R	-	89.12%
Zhang & al. (2015)	CNN	GloVe	-	91.57
Suzuki & all ()	HDAG-SVM	W+N+S	88.2%	-

Table 2. *Review of the literature. Features are: U (unigrams), B (bigrams), T (trigrams), NG (n-grams), WH (wh-word), WS (word forms), L (question length), P (POS-tags), H (headword), HC (head chunk), IS (informer span), HY (hypernyms), IH (indirect hypernyms), S (synonyms), NE (name entities), R (related words), D (dependency structure), BOW (bag of word), W (wordnet synsets)*

4 Data, features

Our work has been done in the context of a project aiming at developing a task-oriented dialogue system, training professionals in improving their communication skills **_anonymous citation_**In this case, conversations are free and all types of question can occur. The dataset as well as the features to be involved in the model have therefore to fit with the requirements of a generic approach to question classification.

4.1 Dataset description

A specific dataset has been created for this project, gathering two types of existing resources. A first part is made of two corpora of conversations: the *Acorformed* corpus [19], containing dialogues between doctors and patients, and the *Corpus of Interactional Data* [3] made of free conversations between relatives. The second part of the dataset is the *FQUAD* database [5] containing a set of *open questions* completed with a subset of *yes-no questions*, in fewer number. The dataset contains in total 35,853 questions (*Acorformed*: 2,934; *FQUAD*: 19,776; *CID*: 13,143).

The dataset has been cleaned by removing signs and punctuation, as well as questions marked by prosody only, and not with any morphology mark (these constructions correspond to affirmative sentences onto which an interrogative prosodic pattern has been applied). Only 50 such questions have been removed.

The utterances have been then labeled as `open`, `yes-no` or `not a question` (hereafter `NaQ`) by automatically identifying interrogative particles, subject-verb inversion, etc. and corrected/completed manually. The resulting figures are presented table 3.

Label	*Open question*	*Yes-no question*	*Not a question*
Number	19,816	193	15,563

Table 3. *Question type distribution*

In a second step, open questions have been labeled in 6 subclasses (`object, person, state, location, quantity, time`) corresponding to the main *wh*-particles. This annotation fits well with the needs of task-oriented dialogue systems by being generic. Note that the type "`event`", corresponding to the particle "*pourquoi* (why)" being almost absent from our dataset, we did not included it in our taxonomy. Table 4 summarizes the distribution of the subclasses.

The classes at both steps are quite unbalanced. As described in the next section, we will applied over/undersampling methods in order to partly rebalance them.

Label	*Object*	*Person*	*State*	*Location*	*Quantity*	*Time*
Number	12,042	4,733	1,189	897	718	233

Table 4. *Label distribution*

4.2 Features description

As underlined in [13], features used in question classification falls in three types: *lexical* (word n-grams, wh-words, word shapes, question length), *syntactic* (POS n-grams, head words, head chunk) and *semantic* (hypernyms, related words and named entities). Word n-grams are used in almost all studies, and more generally, word forms and POS are broadly used. On their side, semantic features have shown to be efficient for question classification in a question-answering perspective, which is an expected effect.

The feature set chosen in our study has been elaborated first by taking into account the specificity of question classification in a dialogue system. In this case, we need to analyze the question by identifying its type, the type of the expected answer and the focus of the question. The classes into which the question have to be classified are then more generic than those used in question-answering systems. This has a consequence on the set of features to be chosen: it is for example unlikely that semantic features play an important role there as it is the case for Q/A. Moreover, we also need in dialogue systems to have feature acquisition done in real time, in order to ensure systems reactivity. This excludes some features that would require a complex or even offline computation, for example prosodic features rather difficult to extract with a good accuracy in real time [16].

We use in this study features calculated from the word forms, lemmas and part-of-speech. We propose to encode them in different manners: bag of words, n-grams and tf-idf. On top of these features, we also encoded n-grams of characters, that capture some morphological information and could therefore play a role in our study. Finally, we also decided to explore the role of syntactic information by adding dependency relations (also encoded as bow and tf-idf). In this case, our hypothesis is that the syntactic structure can play a role in the detection of questions (which is the first step of our classification mechanism).

We keep in this work only features with an occurrence greater than 2. The part-of-speech and lemmas are acquired thanks to the `Marsatag` platform [22], word dependencies with the `spaCy` library. The feature values have been calculated with the *CountVectorizer* and *TfidfVectorizer* functions provided by *Scikit-learn*.

5 Experiments

As explained above, dialogue systems need first to identify as quickly as possible whether the user's utterance is a question or not in order to adjust its strategy

in real time. We propose therefore to distinguish two tasks (and two different classifiers): one for determining whether the utterance is a question and a second for specifying its type.

5.1 Hierarchical classification

Following our data description and the specific needs of dialogue systems, we have developed a hierarchical classification in two steps. The first step is very useful in dialogue systems by determining in real time the type of processing to apply to the user's utterance, that can be either answering a question or updating the dialogue knowledge base. This is of deep importance for the system efficiency: question identification (i.e. the first step) being rapid and robust, this task will constitute the first processing mechanism applied to the user's input, guiding the system towards a question answering module or a knowledge updating one.

The second step of the classification consists in determining the type of the question (and as a consequence that of the answer). We decided for this second step not to distinguish as it is usually done between *open* and *yes-no* questions, but to apply directly a flat classification with the six open question subclasses (`object, person, state, location, quantity, time`) plus a `yes-no` question type.

In this experiment, we build several models by means of different algorithms and feature combinations. We have used *Random forests* (hereafter RFC) that also offers an interesting method to study feature importance. Besides this reference, we applied *Gaussian Naive Bayes* (GNB), *Logistic regression* (LGR), *k-nearest neighbors* (KNN) and *multilayer perceptron* (MLP). The dataset set has been divided into 77% for train and 33% for test.

We present in the following the main results with different evaluation metrics. We applied for the evaluation a 10-fold cross validation. We also calculated the Cohen's kappa in order to complete the figures, taking into account the problem of class imbalance.

5.2 First step classification

The first step aims at classifying user's utterances in two types: *question* and *not-a-question*. The training set contains 10,000 utterances, after applying SMOTE in order to take into account imbalanced classes. In a first experiment, we used the entire set of 13,775 features. As shown in table 5, the best results are obtained with random forest and multilayer perceptron (this last one obtaining the best balanced accuracy).

It is interesting to compare these results with those obtained in generic dialogue act classification consisting in classifying the main dialogue acts, including questions. This type of classification is usually used in dialogue systems in order to identify the *intent* associated with the user's utterance. Among the possible techniques, a two-step classification, reaching the state of the art, has been proposed in [2]. In this approach, the first step classifies 5 main classes including questions, reaching 0.94 accuracy (0.89 of balanced accuracy). The best results

Model	Accuracy	Kappa	F1-score	Balanced Accuracy
RFC	0.967	0.953	0.98	0.975
GNB	0.970	0.960	0.98	0.980
LGR	0.857	0.764	0.88	0.882
KNN	0.801	0.802	0.90	0.899
MLP	**0.971**	0.957	**0.98**	0.987

Table 5. *Results for question/no-question classification using the entire set of features*

here show an accuracy at 0.97 (0.98 F1 score) with MLP (almost at the same level as for RFC). This result is interesting and means that using the question/no-question classifier as a first step for processing user's utterance makes it possible to reduce the question error rate. This constitutes a major improvement for dialogue systems, taking into account that answering questions is considered as their most important behavior: a system not capable of recognizing a question is considered as very bad, much more than a system giving a wrong answer to a question.

In order to address the question of dimensionality and more generally feature selection, we evaluated the respective contribution of the different features. For doing that, we calculated feature importance based on the results of the random forest. By analyzing this ranking, we extracted a set of 1,976 important features, out of the initial set of 13,775. The best result, reported table 5.2, has been obtained by random forest. It show a same F1-score but a slight improvement in comparison with the use of all features for accuracy and balanced accuracy.

Model	Accuracy	Kappa	F1-score	Balanced Accuracy
RFC	0.983	0.967	**0.98**	0.983

Table 6. *Results for question/no-question classification using the 1,976 most important features*

Interpreting feature ranking in this case is complicated because of the great variability of the features: some of them are transparent such as forms (covering typically the interrogative particles) of POS ngrams (e.g. verb/subject inversion) but many of them remains difficult to understand. Note that character n-grams, encoding partially the morphology of interrogative constructions also appear in the main features. However, a global interpretation remains difficult, as it is usually the case with a large set of features. We have therefore implemented a second experiment in which we aggregated features into different clusters according to their type (*form, lemma, pos* and *word dependency*) and their encoding (*bigram, trigram, tf-idf*). As a result, by applying a threshold arbitrarily set to 0.019, we have selected the 7 most important aggregated features (in this order): TF-IDF

of form bigrams and of lemma bigrams, bigrams of forms, TF-IDF of POS, bow of forms, TF-IDF of forms and of lemmas, bow of POS, TF-IDF of POS bigrams.

It is interesting to note first the importance of forms in their different encodings. This effect could come from the fact that interrogative constructions rely on a closed class of words, but also often use specific verb types such as *stative* (also in a closed class). It is also interesting to note the fact that in most of the cases, bigram encodings appear important (more than other n-grams). This can also be explained by the fact that interrogative constructions are often associated with short sequences of words or syntactic disposals (such as subject/verb inversion).

We have used these 7 aggregated features to build a new model on the complete dataset. This modeling relies therefore on a reduced number of features applied to a large number of examples.

After training the 5 classifiers, we observed an improvement in the performance, reaching 0.99 of F1-score (0.98 of balanced accuracy) both for random forest and multilayer perceptron, as shown table 7. When comparing with other approaches, these results can be considered as the state of the art for the identification of questions (vs. not-a-question). Remind that this results answers our first need: proposing to dialogue system a fast and robust classifier for identifying questions in real time, based on a simple set of features.

Model	Accuracy	Kappa	F1-score	Balanced Accuracy
RFC	0.984	0.975	**0.99**	0.986
GNB	0.947	0.891	0.95	0.946
LGR	0.907	0.818	0.91	0.911
KNN	0.904	0.772	0.89	0.899
MLP	0.988	0.977	**0.99**	0. 989

Table 7. *Results for question/no-question classification using the 7 most important features*

5.3 Second step classification

This step consists in building a classifier for the 7 classes of questions we have identified: `object, person, state, location, quantity, time, yes-no`. In this perspective, we built a dataset with 10,000 questions from the full dataset, systematically including *yes-no questions* in order to compensate their low number. As for the first step, we trained 5 classifiers with different algorithms: random forest, naive bayes, logistic regression, KNN and multilayer perceptron.

Table 8 summarizes the results. As for 1st step, the best results are obtained with random forest, reaching a very good F1 score in spite of the number of classes and their distribution. Also note in this case the good level for Kappa and balanced accuracy.

Model	Accuracy	Kappa	F1-score	Balanced Accuracy
RFC	0.956	0.945	**0.97**	0.949
GNB	0.752	0.544	0.77	0.482
LGR	0.596	0.289	0.68	0.891
KNN	0.855	0.443	0.58	0.764
MLP	0.935	0.852	0.91	0.865

Table 8. *Results for 7 classes of questions using all features*

In the same way as we did for step 1, we applied similar feature selection and trained the different classifiers with a reduced set of features first by selecting first the 4,000 best features for this step. Results are shown in table 9. At the difference with step 1, feature aggregation does not lead to a significant improvement. In particular, there is only a slight improvement in the accuracy of random forest, but a loss in F1 score.

Model	**Feat. selection**	**Accuracy**	**Kappa**	**F1-score**	**Balanced Accuracy**
RFC	*4000*	0.965	0.928	0.96	0.965
	Aggreg.	0.969	0.931	0.96	0.968
GNB	*4000*	0.57	0.320	0.56	0.496
	Aggreg.	0.528	0.620	0.77	0.633
LGR	*4000*	0.605	0.705	0.81	0.888
	Aggreg.	0.592	0.672	0.78	0.874
KNN	*4000*	0.871	0.382	0.50	0.786
	Aggreg.	0.858	0.383	0.50	0.791
MLP	*4000*	0.91	0.948	**0.97**	0.921
	Aggreg.	0.941	0.903	0.94	0.925

Table 9. *Results for 7 classes of questions using feature selection. For each model, the first line indicates the result for the 4000 best features, the second for the 7 best aggregated features, as described for step 1.*

6 Conclusion

Question classification is a central task for dialogue systems, for several reasons. First, the system's behavior and its credibility mostly relies on its capacity to answer questions in an efficient way. But at the difference with question answering applications, the answer generally has to be generated instead of retrieved in an existing database. Moreover, dialogue has to answer in real time, which means to analyze the user's input very rapidly. These two constraints renders the task specific: first, the classification should not be related to a specific domain, which

means that the set of classes has to be generic. Moreover, for efficiency reasons, the features to be used by the classifier has to be acquired in real time.

The solution for question classification presented in this paper is generic, based on a high-level question taxonomy that could be applied independently from the domain. It relies on a hierarchical classifier in two steps. The first stage consists in identifying whether the user's input is a question or not. This task is central for dialogue systems. The originality in our approach is that instead of classifying user's input into a set of different dialogue acts, we only focus on question identification. We obtain for this first level much higher results than with dialogue acts: accuracy, balanced accuracy and F1-score of 0.98 with a random forest using feature selection. These results constitute the state of the art for this task. This first level is of deep importance for the evaluation of a dialogue system: almost all questions from the user can be identified, the system knows it has to generate an answer, which greatly improves the evaluation of the system's behavior by the user. The second level of our system classifies into 7 different classes, corresponding to generic question types. We also obtain for this task very good results, F1-score reaching 0.97 with random forest (0,95 balanced accuracy). In conclusion, our proposal presents a generic question classifier, independent from the application domain, and representing for this specific task the state of the art.

This classifier will be implemented in a dialogue system [4], part of a human-machine interaction platform for training medical doctors to break bad news by interaction with a virtual patient [20].

References

1. Banerjee, S., Bandyopadhyay, S.: An empirical study of combing multiple models in Bengali question classification. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing. pp. 892–896 (2013)
2. Blache, P., Abderrahmane, M., Rauzy, S., Ochs, M., Oufaida, H.: Two-level classification for dialogue act recognition in task-oriented dialogues. In: COLING'20 (2020)
3. Blache, P., Bertrand, R., Ferré, G., Pallaud, B., Prévot, L., Rauzy, S.: The corpus of interactional data: a large multimodal annotated resource. In: Handbook of Linguistic Annotation (2014)
4. Blache, P., Houlès, M.: Common Ground, Frames and Slots: Understanding Doctors Interacting with a Virtual Patient. In: International Conference on Natural Language Computing (NATL 2021). London, United Kingdom (Nov 2021), https://hal.archives-ouvertes.fr/hal-03397028
5. d'Hoffschmidt, M., Belblidia, W., Brendlé, T., Heinrich, Q., Vidal, M.: Fquad: French question answering dataset. Tech. rep., arXiv:2002.06071 (2020)
6. Duggenpudi, S.R., Varma, S., Mamidi, R.: Samvaadhana : A telugu dialogue system in hospital domain. In: Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo) (2019)
7. Jin, L., White, M., Jaffe, E., Zimmerman, L., Danforth, D.: Combining cnns and pattern matching for question interpretation in a virtual patient dialogue system. In: Workshop on Innovative Use of NLP for Building Educational Applications (2017)

8. Kumar, A., Aurisano, J., Eugenio, B.D., Johnson, A.: Towards a dialogue system that supports rich visualizations of data. In: Proceedings of SIGDIAL-2016 (2016)
9. Li, X., Huang, X.J., de Wu, L.: Question classification using multiple classifiers. In: Workshop on Asian Language Re-sources and First Symposium on Asian Language Resources Network (2005)
10. Li, X., Roth, D.: Learning question classifiers. In: COLING'02 (2002)
11. Li, X., Roth, D.: Learning question classifiers: The role of semantic information. Natural Language Engineering **12**(3) (2006)
12. Lifeng Jin, D.K., Hussein, A., White, M., Danforth, D.: Using paraphrasing and memory-augmented models to combat data sparsity in question interpretation with a virtual patient dialogue system. In: Workshop on Innovative Use of NLP for Building Educational Applications (2018)
13. Loni, B., van Tulder, G., Wiggers, P., Loog, M., Tax, D.: Question classification with weighted combination of lexical, syntactical and semantic features. In: Proceedings of the 15th international conference of Text, Dialog and Speech (TSD11) (2011)
14. Madabushi, H.T., Lee, M.: High accuracy rule-based question classification using question syntax and semantics. In: Coling-16 (2016)
15. Manna, R., Das, D., Gelbukh, A.: Question classification in a question answering system on cooking. In: Martínez-Villaseñor, L., Herrera-Alcántara, O., Ponce, H., Castro-Espinoza, F.A. (eds.) Advances in Computational Intelligence. pp. 103–108. Springer International Publishing, Cham (2020)
16. Margolis, A., Ostendorf, M.: Question detection in spoken conversations using textual conversations. In: Proceedings of ACL-11 (2011)
17. McAuley, J., Yang, A.: Addressing complex and subjective product-related queries with customer reviews. In: International World Wide Web Conference Committee (IW3C2) (2016)
18. Mishra, M., Mishra, V.K., Sharma, H.: Question classification using semantic, syntactic and lexical features. International journal of Web and Semantic Technology **4**(3) (2013)
19. Ochs, M., Donval, B., Blache, P.: Virtual patient for training doctors to break bad news. In: proceedings of WACAI-2016 (2016)
20. Ochs, M., Mestre, D., de Montcheuil, G., Pergandi, J.M., Saubesty, J., Lombardo, E., Francon, D., Blache, P.: Training doctors' social skills to break bad news: evaluation of the impact of virtual environment displays on the sense of presence. Journal on Multimodal User Interfaces **13**(1), 41–51 (2019)
21. O'Shea, J., Bandar, Z., Crockett, K.: A Multi-Classifier Approach to Dialogue Act Classification Using Function Words, p. 119–143. Springer-Verlag (2012)
22. Rauzy, S., Grégoire, M., Blache, P.: Marsatag, a tagger for french written texts and speech transcriptions. In: Second Asia Pacific Corpus Linguistics Conference (2014)
23. Voorhees, E., Tice, D.: The trec-8 question answering track. In: Proceedings of LREC-2000 (2000)
24. Zhang, D., Lee, W.S.: Question classification using support vector machines. In: International conference on Research and development in information retrieval (SIGIR'03) (2003)
25. Zulqarnain, M., Khalaf Zager Alsaedi, A., Ghazali, R.and Ghouse, M.G., Sharif, W., Aida Husaini, N.: A comparative analysis on question classification task based on deep learning approaches. PeerJ. Computer science (7) (2021)