Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5408–5416
Marseille, 11–16 May 2020

An Empirical Comparison of Question Classification Methods for Question Answering Systems

Eduardo Gabriel Cortes[1], Vinicius Woloszyn[2], Arne Binder[3], Tilo Himmelsbach[2], Dante Barone[1], and Sebastian Möller[2, 3]
[1]Federal University of Rio Grande do Sul, Porto Alegre, Brazil
[2]Technische Universitat Berlin, Belin, Germany
[3]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
egcortes@inf.ufrgs.br, woloszyn@tu-berlin.de, arne.binder@dfki.de,
tilo.himmelsbach@tu-berlin.de, barone@inf.ufrgs.br, sebastian.moeller@tu-berlin.de

Abstract

Question classification is an important component of Question Answering Systems responsible for identifying the type of an answer a particular question requires. For instance, "Who is the prime minister of the United Kingdom?" demands a name of a PERSON, while "When was the queen of the United Kingdom born?" entails a DATE. This work makes an extensible review of the most recent methods for Question Classification, taking into consideration their applicability in low-resourced languages. First, we propose a manual classification of the current state-of-the-art methods in four distinct categories: low, medium, high, and very high level of dependency on external resources. Second, we applied this categorization in an empirical comparison in terms of the amount of data necessary for training and performance in different languages. In addition to complementing earlier works in this field, our study shows a boost on methods relying on recent language models, overcoming methods not suitable for low-resourced languages.

Keywords: Question Answering, Question Classification, Linguistic Resource

1. Introduction

Question Answering (QA) is a field of Computer Science that aims autonomously to answer in a precise way a question posed by a user in natural language. A typical architecture of a QA system is composed by three fundamental elements (Jurafsky and Martin, 2014; Dulceanu et al., 2018; Kalouli et al., 2018): I) Question Processing: which intents to extract the meaning of the question; II) Information Retrieval: that retrieves relevant documents and sentences from a knowledge base; and finally III) Answer processing: that formulates the final answer.

One important component in Question Processing is Question Classification, which determines the type of answer a question needs (Cortes et al., 2018). For example, the question "When will be the Easter holiday?" requires a date, while "How to make a carrot cake?" expects a recipe. The class of answer plays an important role in Questions Answering Systems since it defines what type of information must be recovered from a knowledge base.

The linguistics resources applied to Question Classification, as well as in other Machine Learning tasks are typically monolingual and commonly available for a handful different languages. However, according to Ethnologue [1], nowadays there are more than 7000 languages in the world, of which 80% have fewer than a million speakers, which means that approximately six in ten people on Earth use low-resource languages. In this paper, we present an updated review of the state-of-the-art methods addressed to Question Classification, taking into consideration the applicability for low resourced languages. We first propose a manual classification of the methods according to the degree of dependency on external linguistic resources. Second, we perform an analysis of applicability and performance in low-resourced languages. In short, the main contributions of this paper are:

1. an updated review of the state-of-the-art methods for Question Classification that complements the previous works (Loni, 2011; Sangodiah et al., 2015).
2. a classification and analyses of the methods in terms of dependency on language resources.
3. an empirical analyses of performance in different languages and also in terms of volume of the training data necessary for each method during the learning process.

The rest of this paper is organized as follows. The next section introduces a classification of state-of-the-art methods in four distinct categories according to the dependency of external resources. Section 3. presents current methods of Question Classification from literature and its classification according to the dependency level of external resources. Section 4. describes the experiments performed with the implemented methods. Section 5. presents the results and analysis. Finally, Section 6. summarizes our conclusions and presents future research directions.

2. Dependency of Language Resource

This work proposes an updated review of the state-of-the-art methods for Question Classification, taking into consideration its applicability for low-resourced languages. We defined four levels of dependency based on the effort of building a particular resource, considering the human effort and not the computational one. For instance, a task that relies on a labeled corpus has a higher level of dependency since it requires an expensive manual annotation typically using experts. On the other hand, methods that need

[1]www.ethnologue.com/statistics/size

only computer effort have a lower level of dependency, for instance, the unsupervised task of creation of a Language Model (dos Santos et al., 2017; Santos et al., 2018). We define four classes of dependency for Question Classification as follow:

- *Very High*: approaches that need a set of rules manually created by humans, for instance, a set of handcrafted rules based on morphological, lexical and syntactic characteristics.
- *High*: approaches that need labeled data or require a knowledge base - for example, a Syntactic Parser or WordNet.
- *Medium*: approaches that employ unsupervised strategies - e.g., a language representation model trained using an unlabeled corpus.
- *Low*: approaches that directly extract features from the training corpus, independent of an external resource. For example, term frequency–inverse document frequency (TF-IDF) vectors that can be easily applied in any low resource language.

3. Question Classification Methods for Question Answering

This section presents recent methods for Question Classification that complements previous surveys (Loni, 2011; Sangodiah et al., 2015). We have conducted a manual categorization of those works in four different classes: Very High, High, Medium and Low level of dependency according to definition in section 2.. Additionally, we have implemented at least one method of each category (except by Very High level) to analyze their performance in different settings - the experiments performed is further described in the upcoming Sections. We have tagged as "[tested]" the methods that were implemented and used in our experiments.

3.1. Low Level of Dependency

This section presents recent methods classified as *Low* dependency level of external linguistic resources.

1. **CNN [tested]** (Oswal, 2016) use a Convolutional Neural Network (CNN) model composed of an embedding input layer, three convolutions layers, three max pooling layers, a dropout layer, and a dense output layer for classification of questions. In the proposed architecture, the input question for the CNN model is represented through a list of indexes according to the vocabulary. Then, each word of each question is transformed into a vector of float values, through the embedding input layer. After that, these embedding vectors are sent to three distinct convolutional and max pooling layers. Following, the output of these three layers is concatenated, reshaped, and sent to the last dense layer.

2. **Aouichat** (Aouichat et al., 2018) employ a Support Vector Machine (SVM) and Term Frequency - Inverse Document Frequency (TF-IDF) for Question Classification in Arabic - which was translated from the original UIUC collection (Li and Roth, 2002). The experiments showed a performance of 93% F1-Score.

3.2. Medium Level of Dependency

This section presents recent methods classified as *Medium* dependency level of external linguistic resources.

3. **Kiros** (Kiros et al., 2015) proposes an unsupervised learning of a generic, distributed sentence encoder that can learn sentence representations without any labeled data. Using texts from books, they train an encoder-decoder model that tries to reconstruct the surrounding sentences of an encoded passage. Using the UIUC collection, this method reaches 92.2% of accuracy for Question Classification.

4. **AdaSent** (Zhao et al., 2015) proposes a self-adaptive hierarchical sentence model for Question Classification. It makes a hierarchy of representations from words to sentences through a recursive gated local composition of adjacent segments. Experiments using the UIUC collection shows that the method reaches 92.4% of accuracy.

5. **C-LSTM** (Zhou et al., 2015) proposes a model that combines two different deep learning architectures: CNN with Long Short-Term Memory (LSTM). This model employs CNN to extract a sequence of higher-level sentence representations, which is fed into an LSTM to obtain the sentence representation. The experiments using UIUC collection showed promising results achieving an accuracy of 94.6%.

6. **MGNC-CNN** (Zhang et al., 2016a; Zhang et al., 2016b) proposes the CNN architecture - multi-group norm constraint CNN (MGNC-CNN) that employs multiple sets of word embeddings for sentence classification. It extracts features from sets of input embeddings independently and then joins these at the last layers in the network to form a final feature vector. Using the UIUC collection, it achieved 95.5% of accuracy on Question Classification.

7. **LSTM [tested]** In (Zhou et al., 2016) proposes a integration of Bidirectional LSTM with Two-dimensional Max Pooling. The experiments addressed to a Question Classification task with the UIUC collection shows that this methods achieves 96.1% of accuracy.

8. **Li** (Li et al., 2017) proposes a novel Dropout Mechanism Integrated to avoid overfitting by randomly dropping units from the neural networks during training. This method is employed to improve neural networks for text classification. The experiments using the UIUC collection show an accuracy of 94.4%.

9. **TWEE** (Li et al., 2018) presents a neural network framework for Question Classification that employs

topic modeling, word embedding, and entity embedding. The proposed model incorporates global topical structures for a comprehensive representation of sentences in the learning process. The experiments using the UIUC collection show that the method reaches 96.5% of accuracy.

10. **Zhang** (Zhang et al., 2017) proposes a generalized multi-task learning architecture with four recurrent neural layers. Multi-task learning leverages potential correlations among related tasks to extract common features and yield performance gains. The results using the UIUC collection show that the method reaches 92.3% of accuracy in Question Classification.

11. **Zhao** (Zhao et al., 2018) proposes a method called capsule network for hierarchical multi-label text classification. This method has three strategies to stabilize the dynamic routing process of the capsule network in order to ease the disruption of noise, which may contain information such as irrelevant words. It reaches 92.8% of accuracy using the UIUC collection.

12. **CNN BERT [tested]** a bidirectional Encoder Representation from Transformer (BERT) (Devlin et al., 2018) is a new method of language representation model proposed by researchers at Google AI Language. The model is pre-trained with a deep directional representation from an unlabeled corpus. BERT utilizes self-attention to produce contextualized representations of textual input. This method has achieved state-of-the-art in different classification tasks, like reading comprehension and text classification.

13. **Yang** (Yang et al., 2019) propose a method for transferring capability of neural networks for text classification. The capsule networks allow capturing the intrinsic spatial part-whole relationship between the source and target domains. The authors demonstrate that this method is capable of transferring learning applications like single-label to multi-label text classification and cross-domain sentiment classification. The experiments with a CNN classifier within the UIUC collection show that the method reaches 92.8% of accuracy.

3.3. High Level of Dependency

This section presents recent methods classified as *High* dependency level of external linguistic resources.

14. **TBCNN** (Mou et al., 2015) proposes a tree-based convolutional neural network (TBCNN) for programming language processing, in which a convolution kernel is designed over a program's abstract syntax trees to capture structural information. Programming language processing is a topic in the field of software engineering. Different from a natural language sentence, a program contains rich, explicit, and complicated structural information. This model extracts different parsing trees from the sentences, as constituency and dependency trees. Therefore, the model aims to use a set of tree feature detectors that are applied to the parsing trees in a sliding window manner. These extracted feature vectors are aggregated by a max-pooling layer. Experiments with the UIUC collection show that the approach reaches 96.0% of accuracy.

15. **SVM [tested]** (Xu et al., 2016) proposed an SVM classifier that employs bag-of-words, POS-tag, synonyms and entity type. The results using bag-of-words and dependency word features show that the method reaches 93.4% of accuracy with the UIUC collection.

16. **Van-Tu** (Van-Tu and Anh-Cuong, 2016) focuses on how to select an efficient set of features corresponding to different groups of questions. To select the best features, the author uses an algorithm that tests each feature individually and concatenates the best in a vector. Using the UIUC collection, the method reaches 95.2% of accuracy using mainly lexical and syntactic features as wh-words and head-words.

17. **ATICM** (Hao et al., 2017) proposes a hybrid approach for Question Classification that employs both syntactic and semantic analysis. For syntactic analysis, it uses a dependency relation parsing while for the semantic analysis it employs a WordNet-based feature expansion method. The experiments using an SVM classifier and the UIUC collections show that the method reaches 95% of accuracy.

18. **SVMSR** (Mohd and Hashmy, 2018) proposes a semantic knowledge base based on WordNet to compute semantic similarity between sentences. The experiments using the UIUC collection show that the SVM model using the SR kernel achieved 91.9% accuracy.

19. **Liu** (Liu et al., 2018) proposes a hybrid method that employs information gain, word similarity and frequent lexical patterns for avoiding the use of features with a high computational cost. The experiments with the UIUC collection show that the approach reaches 96% of accuracy.

3.4. Very High Level of Dependency

This section presents recent methods classified as *Very High* dependency level of external linguistic resources.

20. **Madabushi** (Tayyar Madabushi and Lee, 2016) presents a rule based method for Question Classification. First, it creates a syntactic map using a parse tree. Second, the headword is extracted using possessive unrolling, preposition rolling, and entity identification. Finally, it checks the existence of a pattern that matches the wh-word, auxiliary verb, and headword. Once the pattern is found, the question class is returned. The experiment utilizes the UIUC dataset, and the approach hits the state-of-the-art with 97.2 % of accuracy in Question Classification.

Figure 1 presents the volume of work addressed to Question Classification over the time. The complete list of works

used in this picture is attached in Table **??**. We can see that most of the works addressed to Question Classification rely on methods with a *High* level of dependency. Additionally, since 2014, there is a substantial improvement in the number of works with a *Medium* dependency level.

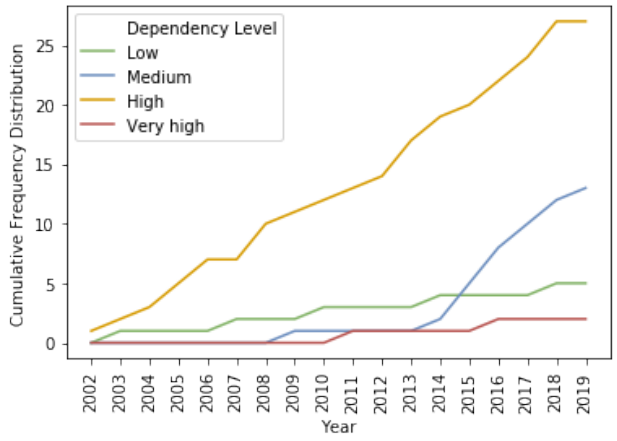


Figure 1: Cumulative Distribution of works over the years

4. Experiments Design

For a better understanding of the applicability and performance of the methods presented before, we performed an empirical comparison taking into consideration the level of dependency, language, and different sizes of training sets. Next, we provide more details about the data sets employed.

4.1. Datasets

We use UIUC (Li and Roth, 2002) and DISEQuA (Magnini et al., 2004) in our experiments since they are usually employed by most of the recent works and also available in different languages which makes possible the execution of our experiments.

4.1.1. UIUC

The UIUC is a monolingual dataset for Question Classification task (Li and Roth, 2002). It has about 6,000 questions in English based on Text Retrieval Conference (TREC), along with their classes. Additionally, it provides a two-layered taxonomy that organizes the questions by a semantic hierarchy.

Beyond the English version, it is also available translated for Spanish (Cumbreras et al., 2006) and Portuguese (Costa et al., 2012). The collection UIUC consists of two separate sets of 5500 and 500 questions. The first one is used as a training set, while the second one is used as an independent test set. Table 1 is presenting the coarse classes distribution for training and test collection, respectively. Also, the UIUC Spanish version has a distinct test collection, in that way, its class distribution is disposed on the column "Test Spanish".

4.1.2. DISEQuA

The DISEQuA provides questions and answers in Dutch, English, Italian, and Spanish (Magnini et al., 2004). A collaboration between three groups built the collection. It allowed a cross-verification of the labels and the distinct language versions. Different from UIUC collection, that split the questions into train and test sets, the DISEQuA do not provide a division. In Table 2 the class distribution is presented. Therefore, DISEQuA provides 450 questions and their classes for training and tests experiments.

Although the dataset is not current and little used by works on literature, it is relevant for our work once it provides the same questions in distinct languages. Therefore, it allows us to compare the results of approaches among languages in a fairer experiment environment.

Table 1: Class distribution of the UIUC collection.

Class	Training	Test	Test Spanish
ABBREVIATION	86	9	0
DESCRIPTION	1162	138	249
ENTITY	1250	94	144
HUMAN	1223	65	307
LOCATION	835	81	245
NUMBER	896	113	404

Table 2: Class distribution of the DISEQuA dataset.

Class	#
DATE	64
LOCATION	85
MEASURE	103
OBJECT	12
ORGANIZATION	41
OTHER	54
PERSON	91

There are few other datasets for QA, e.g., QA&CLEF (Forner et al., 2009) and (Gupta et al., 2018), However, most of them are available only in English. That way, it is not suitable for experiments with multiple languages. The next Section describes the parameterization of the implemented models.

4.2. Parameterization of the implemented models

In order to make it possible to reproduce the results from this work, all of our experiments are public available[2]. This Section presents details about the parameterization of our models.

For the methods that employ unsupervised language models, we use the models from MUSE (Conneau et al., 2017)[3], since we perform tests in different languages and MUSE made available different language models in its repository. Each item of the list represents a tested method with respective parameters:

- **CNN**: The maximum of words considered in a question is 12 (padding size). The dropout percentage used in dropout layers is 50%. The Adam optimizer is applied. The loss function used was the *Categorical*

[2]https://github.com/eduardogc8/simple-qc

[3]public available at https://github.com/facebookresearch/MUSE

Crossentropy. It employs 100 epochs for training with a learning rate equal to 0.001.

- **CNN Bert**: The maximum of words considered in a question is 12 (padding size). The Bert pre-trained model employed was the *bert-base-multilingual*. It uses in maximum 5 epochs for training, with patience equal to 2, annealing factor equal to 0.5, mini-batch size equal to 32 and learning rate equal to 0.001.

- **LSTM**: The maximum of words considered in a question is 12 (padding size). The dropout percentage used in dropout layers is 20%. Its models have two LSTM layers with 256 and 128 neurons respectively. The Adam optimizer is applied. The loss function used was the *Categorical Crossentropy*. It employs 100 epochs for training with a learning rate equal to 0.001.

- **SVM**: The maximum of words considered in a question is 12 (padding size). The SVM kernel used was the Linear with C parameter equal to 1.0. The word embedding used was taken from MUSE Repository, and it has 300 dimensions. The POS-tag and entities type is extracted from questions using the Spacy library.

Finally, the following section presents the results of the implemented approaches, its analysis, and a comparison among literature works, introduced in Section 3..

5. Results and Discussion

First, we present the performance of the most recent methods for Question Classification taking into consideration the dependency level; second, we compare the performance of the selected methods in different languages; last, we analysed the performance of the selected models using different levels of dependency and different sizes of the training set.

5.1. Inter-comparison between different levels of dependency

Figure 2 shows the evolution of the state-of-the-art methods for Question Classification since 2002. Regarding the levels of dependency, methods with a *Very High* level of dependency have improved their accuracy from 95% (Silva) to 97.2% (Madabushi) in 2.2 percentage points (pp). Methods with a *High* level of dependency have improved their accuracy from 91% (approach Li & Roth) to 96% (TBCNN) in 5 pp. Methods with a *Medium* level of dependency showed greater improvement of accuracy from 90% (Tomás) to 96.5% (TWEE) in 6.5 pp. Methods with a *Low* level of dependency have improved their accuracy from 87% (Zhang) to 93% (DCNN) in 6 pp.
Table 3 presents the text classification works from literature that perform experiments with the UIUC collection since 2015. The current difference between methods are: from *Very High* to *High*, *Medium*, and *Low*, are 1.2, 0.7 and 4.2 pp, respectively. From *High* to *Medium* and *Low* are -0.5 and 3 pp, respectively. Finally, from *Medium* to *Low* is 3.5 pp.

Figure 2: Question Classification methods for English UIUC collection considering the published date and their accuracy.

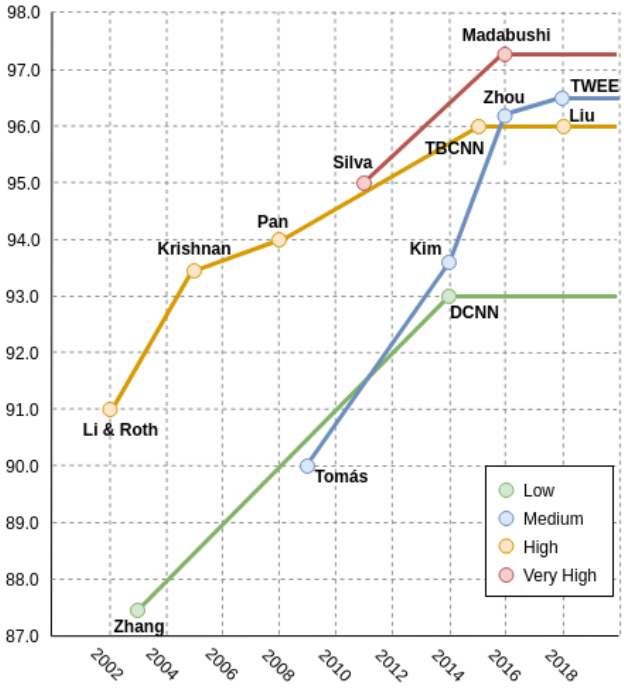


Table 3: Comparison of approaches from literature using UIUC English collection since 2015.

Method	Year	Dependency	Accuracy
Aouichat	2018	Low	93.0
TWEE	2018	Medium	96.5
Zhou	2016	Medium	96.1
MGNC-CNN	2016	Medium	95.5
C-LSTM	2015	Medium	94.6
Li	2017	Medium	94.4
Zhao	2018	Medium	92.8
Yang	2019	Medium	92.8
AdaSent	2015	Medium	92.4
Zhang	2017	Medium	92.3
Kiros	2015	Medium	92.2
TBCNN	2015	High	96.0
Liu	2018	High	96.0
Van-Tu	2016	High	95.2
ATICM	2017	High	95.0
Xu	2016	High	93.4
SVMSR	2018	High	91.9
Madabushi	2016	Very High	97.2

Since 2015, approaches using *Medium* dependency have achieved better results than the ones using *High* dependency. This can be correlated with the improvement of language models. Currently, the difference between *Medium* and *High* is 0.5 pp.

5.2. Performance in Different Languages

In order to make a comparison of the performance in different languages, we have implemented at least one approach of each level of dependency (except by *Very High*) and then

tested using UIUC and DISEQuA collections.

Table 4 and 5 present the results for UIUC collection and DISEQuA collection, respectively. *CNN Bert* presented the best results over the two analyzed collections among all languages tested. Also, CNN-Bert presents a significant difference in performance compared with the runner up methods on the English language. It has a difference of 11.4 pp compared with LSTM on DISEQuA collection and 1.7 pp with SVM on UIUC collection. In short, Bert pre-trained model presents a better performance for English than other languages. Finally, the Dutch and the Spanish present the worst results with CNN Bert, with 80.2 pp in the Spanish UIUC version and 85 pp on the Dutch DISEQuA version.

Table 4: A comparative analysis among the performance of the methods in distinct languages with the UIUC collection.

Method	Dependency	English	Portuguese	Spanish
CNN	Low	89.7	88.1	79.9
CNN Bert	Medium	**94.3**	**90.1**	**80.2**
LSTM	Medium	92.1	90.0	79.0
SVM	High	92.6	87.8	77.9

Table 5: A comparative analysis among the performance of the methods in distinct languages with the DISEQuA collection.

Method	Dependency	English	Italian	Spanish	Dutch
CNN	Low	78.6	80.2	78.8	80.6
CNN Bert	Medium	**92.0**	**91.4**	**91.4**	**85.0**
LSTM	Medium	80.6	84.2	80.2	80.0
SVM	High	75.4	73.6	77.2	75.2

CNN method presents results similar among languages, except for Spanish in UIUC collection. We believe that this difference is caused by the different test sets. The most significant difference was 1.8 pp between Spanish and Dutch in DISEQuA collection. Also, this approach is the only one that does not employ external resources, and therefore its performance does not depend whether the linguistic resource was well trained or build for a target language. In that way, the difference performance among languages can reflect on the difference between the particularities of each language.

Finally, it is possible to observe that approaches that employ external linguistic resources present a significant difference in performance among languages. For instance, the more significant difference between language on CNN approach (*Low* level of dependency) was 2 pp while SVM (*High* level of dependency) was 4.8 pp. It is mainly due to the different quality of these external resources for each language. Additionally, most of the approaches using the English version present better performance compared with the other languages, even if it was created in the same way.

5.3. Performance in different Sizes of Training Set

Figures 3 and 4 present the results of the tested approaches for each language varying the size of the training set of UIUC and DISEQuA collections.

Figure 3: Results with different size of the training set on UIUC collection.

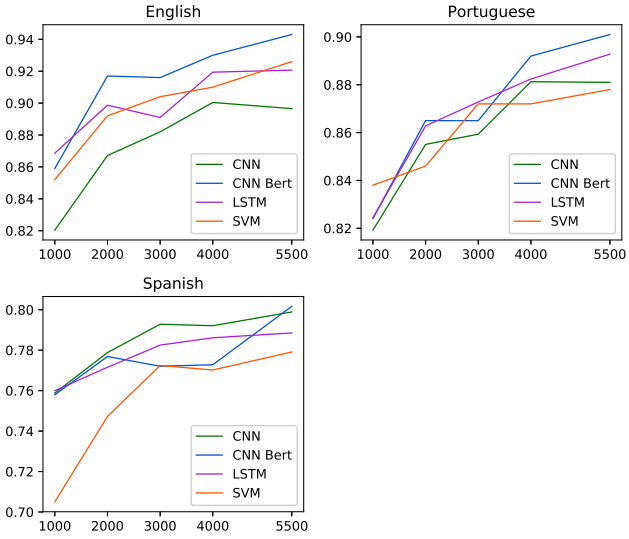
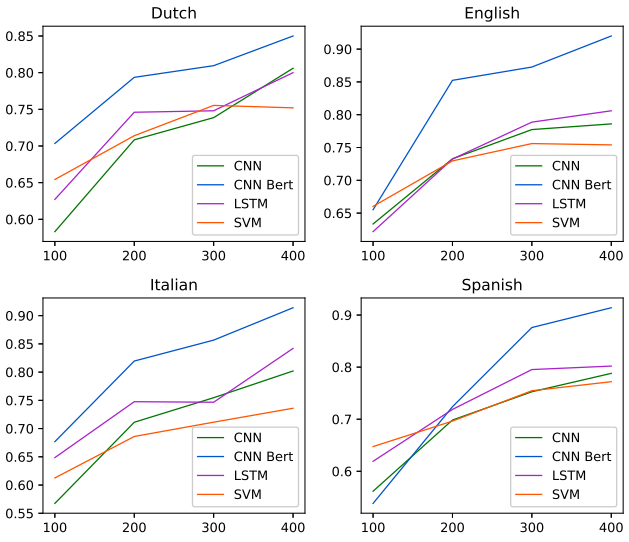


Figure 4: Results with different size of the training set on DISEQuA collection.



CNN Bert method reaches the best results in most of our tests. For the two collections and all the languages, the approach got the best result with the large size of training instance. In the UIUC collection, even it has a fixed training and test set, the results are noisier than the ones from DISEQuA. The different characteristics of each language may cause it, and especially by the peculiarities when creating the other versions, since Portuguese and Spanish one were separately translated from the original English version.

A disadvantage of deep learning models is that they require more training data to reach significant results. In Figure 4, it is possible to observe that most of the time LSTM and CNN reach the best performance when the training set has more than about 200 instances, while the SVM got the best results with few data. Nevertheless, for most of the languages, CNN Bert achieved better results when compared with other models, even under less than 200 instances of

training. Therefore, it can indicate that a pre-trained language representation model of BERT improves the performance of classifier models for Question Classification with little data for training.

6. Conclusion

In this work, we put forward an empirical comparison between recent methods for Question Classification. Additionally, we performed a manual classification of works on four levels based on language resource dependency. A quantitative analyses shows that most of the methods addressed to Question Classification relies on high-level of dependency of linguistic resources. On the other hand, we have observed a substantial rising (since 2014) of the volume of works that rely on *Medium* level of dependency.

A qualitative analysis of the literature shows that rule-based methods manually created by humans (classified as *Very high* level of dependency) still have better results. However, these methods require a broad set of rules based on lexical and syntactic patterns from questions. Therefore, its applicability to other languages requires a great human effort. On the other hand, methods that rely on pre-trained language models, which in our classification has a *Medium* level of dependency has substantially improved its performance, outperforming methods that rely on a *High* level of dependency. A considerable advantage of those models is the ability to learn how to represent relevant features in each layer of the model. In short, features like grammar class and entity type of each word can be learned in the hidden layers of these models, without the need to use an external tool to represent them.

Finally, we carry out experiments using two different datasets in five different languages using different sizes of training data. The experiments showed that it is possible to reach significant results in Question Classification without having to use complex features to build for low resource languages. For instance, CNN Bert, classified with a Medium Dependency level, reaches results superior in almost all languages and sizes of the training set.

7. Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

8. References

Aouichat, A., Hadj Ameur, M. S., and Geussoum, A. (2018). Arabic question classification using support vector machines and convolutional neural networks. In Max Silberztein, et al., editors, *Natural Language Processing and Information Systems*, pages 113–125, Cham. Springer International Publishing.

Cortes, E. G., Woloszyn, V., and Barone, D. A. (2018). When, where, who, what or why? a hybrid model to question answering systems. In *International Conference on Computational Processing of the Portuguese Language*, pages 136–146. Springer.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

dos Santos, H. D., Woloszyn, V., and Vieira, R. (2017). Portuguese personal story analysis and detection in blogs. In *Proceedings of the International Conference on Web Intelligence*, pages 709–715.

Dulceanu, A., Le Dinh, T., Chang, W., Bui, T., Kim, D. S., Vu, M. C., and Kim, S. (2018). PhotoshopQuiA: A corpus of non-factoid questions and answers for why-question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).

Forner, P., Peñas, A., Agirre, E., Alegria, I., Forăscu, C., Moreau, N., Osenova, P., Prokopidis, P., Rocha, P., Sacaleanu, B., Sutcliffe, R., and Tjong Kim Sang, E. (2009). Overview of the clef 2008 multilingual question answering track. In Carol Peters, et al., editors, *Evaluating Systems for Multilingual and Multimodal Information Access*, pages 262–295, Berlin, Heidelberg. Springer Berlin Heidelberg.

Gupta, D., Pujari, R., Ekbal, A., Bhattacharyya, P., Maitra, A., Jain, T., and Sengupta, S. (2018). Can taxonomy help? improving semantic question matching using question taxonomy. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 499–513, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Hao, T., Xie, W., Wu, Q., Weng, H., and Qu, Y. (2017). Leveraging question target word features through semantic relation expansion for answer type classification. *Knowledge-Based Systems*, 133:43 – 52.

Jurafsky, D. and Martin, J. H. (2014). *Speech and language processing*, volume 3. Pearson London.

Kalouli, A.-L., Kaiser, K., Hautli-Janisz, A., Kaiser, G. A., and Butt, M. (2018). A multilingual approach to question classification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In C. Cortes, et al., editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.

Li, S., Zhao, Z., Liu, T., Hu, R., and Du, X. (2017). Initializing convolutional filters with semantic features for text classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1884–1889, Copenhagen, Denmark, September. Association for Computational Linguistics.

Li, D., Zhang, J., and Li, P. (2018). Representation learning for question classification via topic sparse autoencoder and entity embedding. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 126–133, Dec.

Liu, Y., Yi, X., Chen, R., Zhai, Z., and Gu, J. (2018). Feature extraction based on information gain and sequential pattern for english question classification. *IET Software*,

12(6):520–526.

Loni, B. (2011). A survey of state-of-the-art methods on question classification.

Mohd, M. and Hashmy, R. (2018). Question classification using a knowledge-based semantic kernel. In Millie Pant, et al., editors, *Soft Computing: Theories and Applications*, pages 599–606. Springer Singapore, Singapore.

Mou, L., Peng, H., Li, G., Xu, Y., Zhang, L., and Jin, Z. (2015). Discriminative neural sentence modeling by tree-based convolution. *arXiv preprint arXiv:1504.01106*.

Oswal, B. V. (2016). Cnn-text-classification-keras. `https://github.com/bhaveshoswal/CNN-text-classification-keras`.

Sangodiah, A., Muniandy, M., and Heng, L. E. (2015). Question classification using statistical approach: A complete review. *Journal of Theoretical & Applied Information Technology*, 71(3).

Santos, H., Woloszyn, V., and Vieira, R. (2018). Blogsetbr: A brazilian portuguese blog corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Tayyar Madabushi, H. and Lee, M. (2016). High accuracy rule-based question classification using question syntax and semantics. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1220–1230, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Van-Tu, N. and Anh-Cuong, L. (2016). Improving question classification by feature extraction and selection. *Indian Journal of Science and Technology*, 9(17):1–8.

Xu, S., Cheng, G., and Kong, F. (2016). Research on question classification for automatic question answering. In *2016 International Conference on Asian Language Processing (IALP)*, pages 218–221, Nov.

Yang, M., Zhao, W., Chen, L., Qu, Q., Zhao, Z., and Shen, Y. (2019). Investigating the transferring capability of capsule networks for text classification. *Neural Networks*, 118:247 – 261.

Zhang, R., Lee, H., and Radev, D. R. (2016a). Dependency sensitive convolutional neural networks for modeling sentences and documents. *CoRR*, abs/1611.02361.

Zhang, Y., Roller, S., and Wallace, B. C. (2016b). MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification. *CoRR*, abs/1603.00968.

Zhang, H., Xiao, L., Wang, Y., and Jin, Y. (2017). A generalized recurrent neural architecture for text classification with multi-task learning. *CoRR*, abs/1707.02892.

Zhao, H., Lu, Z., and Poupart, P. (2015). Self-adaptive hierarchical sentence model. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Zhao, W., Ye, J., Yang, M., Lei, Z., Zhang, S., and Zhao, Z. (2018). Investigating capsule networks with dynamic routing for text classification. *CoRR*, abs/1804.00538.

Zhou, C., Sun, C., Liu, Z., and Lau, F. C. M. (2015). A C-LSTM neural network for text classification. *CoRR*, abs/1511.08630.

Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., and Xu, B. (2016). Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. *CoRR*, abs/1611.06639.

9. Language Resource References

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Costa, Â., Luís, T., Ribeiro, J., Mendes, A. C., and Coheur, L. (2012). An English-Portuguese parallel corpus of questions: translation guidelines and application in SMT. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2172–2176, Istanbul, Turkey, May. European Languages Resources Association (ELRA).

Cumbreras, M. A. G., López, L. A. U. n., and Santiago, F. M. (2006). Bruja: Question classification for spanish. using machine translation and an english classifier. In *Proceedings of the Workshop on Multilingual Question Answering*, MLQA '06, pages 39–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

Li, X. and Roth, D. (2002). Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F., and de Rijke, M. (2004). Creating the disequa corpus: A test set for multilingual question answering. In Carol Peters, et al., editors, *Comparative Evaluation of Multilingual Information Access Systems*, pages 487–500, Berlin, Heidelberg. Springer Berlin Heidelberg.

Table 6: Question classification methods from the literature that perform experiments with the UIUC collection.

Reference of the Question Classification work	Label	Dependency Level
Li, X. and Roth, D. (2002). Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics	Li & Roth	High
Aouichat, A., Hadj Ameur, M. S., and Geussoum, A. (2018). Arabic question classification using support vector machines and convolutional neural networks. In Max Silberztein, et al., editors, *Natural Language Processing and Information Systems*, pages 113–125, Cham. Springer International Publishing	Aouichat	Low
Hacioglu, K. and Ward, W. (2003). Question classification with support vector machines and error correcting codes. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–short Papers - Volume 2*, NAACL-Short '03, pages 28–30, Stroudsburg, PA, USA. Association for Computational Linguistics		High
Li, X., Roth, D., and Small, K. (2004). The role of semantic information in learning question classifiers. In *Proceedings of the International Joint Conference on Natural Language Processing*		High
Metzler, D. and Croft, W. B. (2005). Analysis of statistical question classification for fact-based questions. *Information Retrieval*, 8(3):481–504, Jan		High
Krishnan, V., Das, S., and Chakrabarti, S. (2005). Enhanced answer type inference from questions using sequential models. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 315–322. Association for Computational Linguistics	Krishnan	High
Blunsom, P., Kocik, K., and Curran, J. R. (2006). Question classification with log-linear models. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 615–616. ACM		High
Cumbreras, M. A. G., López, L. A. U. n., and Santiago, F. M. (2006). Bruja: Question classification for spanish. using machine translation and an english classifier. In *Proceedings of the Workshop on Multilingual Question Answering*, MLQA '06, pages 39–44, Stroudsburg, PA, USA. Association for Computational Linguistics		High
Merkel, A. and Klakow, D. (2007). Improved methods for language model based question classification. In *Eighth Annual Conference of the International Speech Communication Association*		Low
Li, F., Zhang, X., Yuan, J., and Zhu, X. (2008). Classifying what-type questions by head noun tagging. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 481–488, Stroudsburg, PA, USA. Association for Computational Linguistics		High
Pan, Y., Tang, Y., Lin, L., and Luo, Y. (2008). Question classification with semantic tree kernel. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 837–838. ACM	Pan	High
Huang, Z., Thint, M., and Qin, Z. (2008). Question classification using head words and their hypernyms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 927–936. Association for Computational Linguistics		High
Tomas, D. and Giuliano, C. (2009). A semi-supervised approach to question classification. In *ESANN*. Citeseer	Tomas	Medium
Huang, Z., Thint, M., and Celikyilmaz, A. (2009). Investigation of question classifier in question answering. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 543–550, Stroudsburg, PA, USA. Association for Computational Linguistics		High
Li, B. and Vogel, C. (2010). Improving multiclass text classification with error-correcting output coding and sub-class partitions. In *Canadian Conference on Artificial Intelligence*, pages 4–15. Springer		Low
Ray, S. K., Singh, S., and Joshi, B. P. (2010). A semantic approach for question classification using wordnet and wikipedia. *Pattern Recognition Letters*, 31(13):1935–1943		High
Loni, B., Van Tulder, G., Wiggers, P., Tax, D. M., and Loog, M. (2011). Question classification by weighted combination of lexical, syntactic and semantic features. In *International Conference on Text, Speech and Dialogue*, pages 243–250. Springer		High
Silva, J., Coheur, L., Mendes, A. C., and Wichert, A. (2011). From symbolic to sub-symbolic information in question classification. *Artificial Intelligence Review*, 35(2):137–154	Silva	Very High
Waltinger, U., Breuing, A., and Wachsmuth, I. (2012). Connecting question answering and conversational agents. *KI - Künstliche Intelligenz*, 26(4):381–390, Nov		High
Mishra, M., Mishra, V. K., and Sharma, H. (2013). Question classification using semantic, syntactic and lexical features. *International Journal of Web & Semantic Technology*, 4(3):39		High
Post, M. and Bergsma, S. (2013). Explicit and implicit syntactic features for text classification. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 866–872		High
Hardy, H. and Cheah, Y.-N. (2013). Question classification using extreme learning machine on semantic features. *Journal of ICT Research and Applications*, 7(1):36–58		High
Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. *CoRR*, abs/1404.2188	DCNN	Low
Kim, Y. (2014). Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882	Kim	Medium
Le-Hong, P., Phan, X.-H., and Nguyen, T.-D. (2015). Using dependency analysis to improve question classification. In Viet-Ha Nguyen, et al., editors, *Knowledge and Systems Engineering*, pages 653–665, Cham. Springer International Publishing		High
Liu, L., Yu, Z., Guo, J., Mao, C., and Hong, X. (2014). Chinese question classification based on question property kernel. *International Journal of Machine Learning and Cybernetics*, 5(5):713–720, Oct		High
Zhou, C., Sun, C., Liu, Z., and Lau, F. C. M. (2015). A C-LSTM neural network for text classification. *CoRR*, abs/1511.08630	C-LSTM	Medium
Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In C. Cortes, et al., editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc	Kiros	Medium
Zhao, H., Lu, Z., and Poupart, P. (2015). Self-adaptive hierarchical sentence model. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*	AdaSent	Medium
Mou, L., Peng, H., Li, G., Xu, Y., Zhang, L., and Jin, Z. (2015). Discriminative neural sentence modeling by tree-based convolution. *arXiv preprint arXiv:1504.01106*	TBCNN	High
Zhang, R., Lee, H., and Radev, D. R. (2016a). Dependency sensitive convolutional neural networks for modeling sentences and documents. *CoRR*, abs/1611.02361	MGNC-CNN	Medium
Zhang, Y., Roller, S., and Wallace, B. C. (2016b). MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification. *CoRR*, abs/1603.00968	MGNC-CNN	Medium
Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., and Xu, B. (2016). Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. *CoRR*, abs/1611.06639	LSTM	Medium
Xu, S., Cheng, G., and Kong, F. (2016). Research on question classification for automatic question answering. In *2016 International Conference on Asian Language Processing (IALP)*, pages 218–221, Nov	SVM	High
Van-Tu, N. and Anh-Cuong, L. (2016). Improving question classification by feature extraction and selection. *Indian Journal of Science and Technology*, 9(17):1–8	Van-Tu	High
Tayyar Madabushi, H. and Lee, M. (2016). High accuracy rule-based question classification using question syntax and semantics. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1220–1230, Osaka, Japan, December. The COLING 2016 Organizing Committee	Madabushi	Very High
Li, S., Zhao, Z., Liu, T., Hu, R., and Du, X. (2017). Initializing convolutional filters with semantic features for text classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1884–1889, Copenhagen, Denmark, September. Association for Computational Linguistics	Li	Medium
Zhang, H., Xiao, L., Wang, Y., and Jin, Y. (2017). A generalized recurrent neural architecture for text classification with multi-task learning. *CoRR*, abs/1707.02892	Zhang	Medium
Hao, T., Xie, W., Wu, Q., Weng, H., and Qu, Y. (2017). Leveraging question target word features through semantic relation expansion for answer type classification. *Knowledge-Based Systems*, 133:43 – 52	ATICM	High
Aouichat, A., Hadj Ameur, M. S., and Geussoum, A. (2018). Arabic question classification using support vector machines and convolutional neural networks. In Max Silberztein, et al., editors, *Natural Language Processing and Information Systems*, pages 113–125, Cham. Springer International Publishing	Aouichat	Low
Li, D., Zhang, J., and Li, P. (2018). Representation learning for question classification via topic sparse autoencoder and entity embedding. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 126–133, Dec	TWEE	Medium
Zhao, W., Ye, J., Yang, M., Lei, Z., Zhang, S., and Zhao, Z. (2018). Investigating capsule networks with dynamic routing for text classification. *CoRR*, abs/1804.00538	Zhao	Medium
Mohd, M. and Hashmy, R. (2018). Question classification using a knowledge-based semantic kernel. In Millie Pant, et al., editors, *Soft Computing: Theories and Applications*, pages 599–606. Springer Singapore, Singapore	SVMSR	High
Liu, Y., Yi, X., Chen, R., Zhai, Z., and Gu, J. (2018). Feature extraction based on information gain and sequential pattern for english question classification. *IET Software*, 12(6):520–526	Liu	High
Yang, M., Zhao, W., Chen, L., Qu, Q., Zhao, Z., and Shen, Y. (2019). Investigating the transferring capability of capsule networks for text classification. *Neural Networks*, 118:247 – 261	Yang	Medium