Towards Multilingual Neural Question Answering

Ekaterina Loginova, Stalin Varanasi, and Günter Neumann

DFKI, Saarbrücken, Germany
dfki.de
ekaterina.d.loginova@gmail.com

Abstract. Cross-lingual and multilingual question answering is a critical part of a successful and accessible natural language interface. However, many current solutions are unsatisfactory. We believe that recent developments in deep learning approaches are likely to be efficient for question answering tasks spanning several languages. This work aims to discuss current achievements and remaining challenges. We outline requirements and suggestions for practical parallel data collection and describe existing methods and datasets. We also demonstrate that a simple translation of texts can be inadequate in case of Arabic, English and German languages (on InsuranceQA and SemEval datasets), and thus more sophisticated models are required. We hope that our findings will ignite interest in neural approaches to multilingual question answering.

Keywords: question answering · multilingual natural language processing · neural natural language processing · deep learning

1 Introduction

The focus in natural language processing, and especially in question answering, has always been on the English language, in spite of others (such as Spanish, Hindi and Chinese) having equal or larger number of native speakers. Nonetheless, cross and multilingual language technologies have many applications in the field of intelligent assistance, information retrieval and question answering (QA). Cross-lingual (essentially, bilingual) techniques are concerned with the retrieval of information in a language different from that of a query, while multilingual work with information sources in several languages simultaneously. With the development of deep learning approaches, which do not require manual feature engineering, we strongly feel it is time to revisit the research area and broaden the language coverage.

In this work, we focus on QA task and provide an overview of the current state of the field for multilingual (MLQA) and cross-lingual (CLQA) subtasks. We also include a preliminary attempt to analyse the performance of a simple deep learning model in multiple language setting. Our research question is how well does the same model perform on original Arabic (English) texts and their English (German) translations, and if there is a difference, then why.

This paper is divided into five sections. The first section provides a brief overview of existing datasets for cross-lingual and multilingual QA and discusses the collection and analysis of such linguistic resources. The second section examines approaches to the problem and reports state-of-the-art results on several shared tasks. In the third section, a case study is presented for which we compare the performance of a deep learning model before and after translating a corpus of non-factoid questions and answers. Possible directions for future research are outlined in the fourth section. Our conclusions are drawn in the final section.

2 Datasets

2.1 Existing datasets

Parallel Previous work has generally been limited to machine translation of questions [11], [13], as interrogative structures are likely to be a challenge for systems trained mainly on declarative. Apart from that, Cross-Language Evaluation Forum (CLEF) provided the following multilingual datasets: Multisix corpus [27], (200 questions, 6 languages), the DISEQuA corpus [26] (450 questions, 4 languages), the Multieight-04 corpus [28] (700 questions, 7 languages), and the Multi9-05 [42] (900 questions, 9 languages). Some corpora also include question type in their annotation [5]. A small parallel corpus for Japanese and English was constructed by the NTCIR organisers [37].

However, even though such a corpus can be relatively easily augmented with crowd-sourced answers to become a fully functional question-answering corpus, it still would have significant flaws. Firstly, it does not allow one to perform answer rankings, only answer selection. Secondly, its small size limits the use of deep learning approaches. At the time of publication, to the best of our knowledge, there are no parallel corpora for QA sufficiently large to to take advantage of deep learning techniques fully.

Lack of resources (such as parallel corpora or sufficiently large dictionaries) for many languages is also a significant problem. Not only does it hinder the machine translation methods' development, but also the generation of cross-lingual embeddings, effectively leaving little choice in approaches. We hypothesise that using knowledge databases and semantic parsing might be helpful. However, if the method relies on heavy preprocessing of a text, it might be difficult to transfer to languages with scarce tools. We suggest using a multi-level (tokenisation, chunking, translation, answering) evaluation process during the shared tasks to facilitate the development.

Code-Mixed Code-mixing or code-switching is a linguistic phenomenon frequently occurring in multilingual communities. It results in texts where words of two languages are used simultaneously within a single sentence. Code-mixing is particularly noticeable in India, where native speakers of Telugu, Hindi and Tamil are often using English words without translating them. As an example, consider the following Hinglish (Hindi + English) sentence: Bhurj Khalifa kaha

located he? (Where is Burj Khalifa located?). Due to the morphological richness and non-latin alphabets of many languages, it can be even more complicated. For instance, a sentence can include combinations of transliterated English stems with native affixes or switching alphabets (cross-script).

Despite extensive use of code-mixing in informal conversations, there are not many datasets present. The existing datasets are mainly concerned with machine translation, language detection and sentiment analysis, but not QA. Among recent developments in this area, we can note [4]. Along with the corpus, a novel annotation scheme and evaluation strategy specific for QA have been proposed. Another potentially useful dataset is CMIR, described in detail in [7]. This dataset consists of 1959 code-mixed tweets.

2.2 Dataset collection and analysis

As stated above, one of the key problems in multilingual QA is a lack of large parallel corpora. Presumably, such a corpus would have to satisfy the following requirements:

1. Not only questions and answers should align, but also the contexts. This alignment should be taken into account when working with existing multilingual collections such as Wikipedia, as the articles on the same topic might differ significantly across the languages. Otherwise, comparison of results for different languages might be affected.
2. Annotators should be bilingual. Proof of language knowledge needs to be provided to ensure quality translation. It should be noted that interrogative structures can sometimes present a bigger problem for non-native speakers. It is thus preferable to check how confident a crowd-source worker is with advanced grammar constructions.
3. The annotation scheme should include at least tokenisation and chunking to help researchers elicit a step in the preprocessing pipeline causing the most errors. In addition, question and answer type labels might be useful.

One issue that needs to be raised is which languages should be a priority? We can either choose according to the quality of machine translation or by the number of bilingual native speakers. Language pairs also ought to include languages with profoundly different grammar rules and preferably from several alphabets. Hence, we can separate four main types of parallel corpora (listed in increasing complexity):

- closely related languages with similar alphabets (Italian, Spanish)
- distant languages with similar alphabets (Danish, French)
- closely related languages with different alphabets (Polish, Russian)
- distant languages with different alphabets (English, Chinese)

Possible sources of parallel data include Trivia and other worldwide question answering games, as well as multilingual countries' exam sheets. Nevertheless,

current techniques to collect QA pairs are time-consuming. Another possibility is to generate question-answer pairs from existing parallel MT corpora. The recent success of deep learning question generation [12] is promising for doing so in an end-to-end (semi-) automated fashion, which is important for languages with scarce resources. A hybrid system can be considered when a neural network generates question-answer pairs, and a human annotator further refines them.

The properties of multilingual QA datasets have not been dealt with in depth. We argue that the following characteristics need to be considered:

1. diversity and balance of answer and question types
2. reasoning type for questions along with the difficulty score
3. whether the reasoning over multiple sentences or documents is required
4. the degree of syntactic and lexical divergence between the question and the answer

We surmise that more attention should be paid to the properties of texts potentially useful for deep learning models. Among them is the perplexity of the dataset. It indicates how patterns are repeating in the dataset. The higher the perplexity, the more unlikely it is to see patterns repeating and hence the more difficult it is to learn a model.

Another question is, what can be seen as a universal baseline? It should satisfy the following conditions: applies to a wide variety of languages, easy to use, and freely available. The most widespread baseline at the moment is to translate texts to English with Google Translate [22]. A major drawback is an unequal quality of translation for different languages, which should be taken into account during a comparison. Besides, an appropriate metric should be chosen to evaluate the code-mixed complexity of a corpus, such as Complexity Metric proposed in [17].

3 Methods

3.1 Related work

In general, one can distinguish the following approaches:

- annotate data for each input language and then train separately
- use machine translation directly beforehand or as part of a QA system; and then to work in the monolingual setup of a target language
- use a universal cross-language representation (such as cross-lingual embeddings)
- map terms in several languages with a multilingual knowledge base or a semantic graph, such as Wikipedia or BabelNet ([6])

In the following paragraphs, we will discuss some recent developments in the field.

Regarding the latest machine learning models, [20] considers the use of semantic parsing for multilingual QA over linked data. One of the main issues

there is the lexical gap, which is present even in a monolingual case due to paraphrasing diversity but is even further pronounced for a multilingual case. The authors propose a model that utilises DUDES (Dependency-based Underspecified Discourse Representation Structures) [10] universal dependencies to tackle it. Experiments were carried out on the QALD-6 dataset covering English, German and Spanish language. Although the results are behind state-of-the-art, it is quite likely that semantic parsing might be helpful for fully exploiting information from several languages. [43] describes a combination of a Maximum Entropy model for keyword extraction and an SVM for answer type classification to find an answer in a knowledge base. One of the main advantages is language independence except for the use of a chunker. The paper reports an F1 score for Spanish data of 54.2 as compared to the 32.2 baseline score obtained by translating question into English with Google Translate. Lastly, it is worth to note that the findings stated in [41] do not support the hypothesis that learning a custom classifier for each language would outperform the single classifier baseline.

Despite its long use, machine translation has some problems. Due to practical constraints, we have not yet performed an in-depth error analysis. However, a loss or corruption of named entities has frequently been observed during translation. Code-mixed texts are even more challenging as illustrated in [22]. Besides, in the traditional approach, the machine translation is performed independently of QA as a part of input preparation. Recently, there has been a trend to blend the two components. [41] draws our attention to the problem of joint training for machine translation and QA components. They propose an answer ranking model that learns the optimal translation according to how well it classifies the answer. This novel approach achieves 0.681 MAP (Mean Average Precision) on a collection of English, Arabic and Chinese forum posts, which outperforms the baseline approach of translating everything into English. [18] reports on a novel method to incorporate response feedback to the machine translation system. The response is received based on performance in an extrinsic task. For instance, one might generate the translation of a question and define a successful response as receiving the same answer for both translation and the original question. Finally, [38] calls into question the correspondence between human assessment of translation, machine translation metrics and cross-lingual QA quality. They create a dataset and investigate the relationship between translation evaluation metrics and QA accuracy. According to sentence-level correlation analysis, the NIST score is the most indicative one due to it treating content words as more important than question ones. This effect is especially noticeable for named entities, but the correct translation of question words is also crucial. Moreover, the authors claim that the conversion of entities into logical forms, typical for methods utilising a knowledge base, can be heavily affected by a translation. Another potential issue is a change in the word order, which might harm the performance of predicate construction and merging. Overall, the authors conclude that QA system and humans do estimate the translation quality in a very different way.

Recent developments in deep learning have led to considerable interest from natural language processing community. However, these models are just enter-

ing the field of multilingual QA. In [29], authors compare a tree-kernel-based system with a feed-forward neural network which uses cross-lingual embeddings. The data is in Arabic and English. As a baseline, they translate questions into English and train a monolingual system. They assume that a large parallel corpus is available. They also point out that since many models utilise syntactic and semantic structures, it might be difficult to adapt them to a new language. In addition to it, they demonstrate the robustness of standard similarity features. Another exciting result has been reported by [23], who developed a new method for community QA based on the Domain Adversarial Neural Network model from [16]. They have adapted it to a cross-lingual task by coupling a detection network with a question-answering one. They achieve a MAP of 0.7589 on Arabic data and 0.7593 on English. Finally, a promising direction is to use cross-lingual embeddings in deep learning models for CLQA. For a detailed review, see [35].

3.2 Benchmarks

As mentioned above, one of the most popular MLQA challenges is the CLEF campaigns. It includes 33 cross-language sub-tasks across 10 European languages. The corpus was combined out of ELRA/ELDA news and Wikipedia articles. During the challenge, the participating systems had to answer factoid, definition and list questions and provide supporting evidence in the form of text snippets. The performance metric was top-1 accuracy and, in most cases, MRR. In case of the German language as a target, the best performing system scored 37% [36]. More details on this topic can be found in [15].

Another noticeable shared task is NTCIR-6 [37], which concentrates on cross-lingual QA. The target languages are English, Chinese and Japanese. The corpus is also based on newspaper articles. There can be only one or no answer, and its type is restricted to a named entity. The performance metric is top-1 accuracy and MRR. The organisers note that some questions can be answered correctly in CLQA but not in MLQA and the named entity identification modules, as well as translation, significantly influence the performance.

Concerning code-mixed texts, there is a surge of interest from multilingual communities in India, specifically for Hindi, Telugu and Tamil. However, currently, the work is mainly limited to question type classification and information retrieval, encouraged by the recently shared tasks of MSIR and FIRE [3]. In the question classification task, [33] report an accuracy of 45.00%. [8] achieve an MRR of 0.37 and 0.32 for Hinglish and Tenglish respectively, using lexical translation and SVM-based question classification. In information retrieval, machine learning methods such as Naive Bayes and RF classifiers dominate, with the best MAP score being 0.0377.

4 Experiments

We have performed our experiments on two datasets: InsuranceQA (version 2) [14] and SemEval 2017 (subtask D) [31]. The approach involves training the

same deep learning model on original texts and their translations to compare the performance. For both datasets, questions are non-factoid and can have multiple correct answers. The task is to rank the set of answers based on their relevance to the question. InsuranceQA texts are originally in English, while the SemEval ones are in Arabic. The texts were translated to German and English respectively. The Google Translate neural machine translation system for English - German language pair achieves a BLEU score of 24.60 [22]. For Arabic to English translation, the average precision has been reported to be 0.449 [19].

The deep learning model we use is an attentional Siamese Bidirectional LSTM. The method is essentially the same as that introduced by [39] with some adjustments in hyper-parameters and loss function. More details will be given below in the corresponding sections. We chose this model because it performed the best over multiple runs for SemEval 2017 Subtask A in our previous experiments. There it has obtained a MAP of 0.8349 (the IR baseline is 0.7261, the best result is 0.8843, and the best only deep learning result is 0.8624 [31]). The model accepts a question, its correct answer and an incorrect answer from the pool as an input. The goal is to project them in such a way that correct answers are closer to their corresponding questions than the incorrect ones.

The motivation for using deep learning is a significant gap in parallel resources and advanced tools for many languages. Our current approach only requires a collection of texts to train monolingual word embeddings on, a translation system and a tokeniser.

We limit the word sequence length to 200 tokens during training. For all languages, we use FastText word embeddings pre-trained on Wikipedia [24]. We choose these embeddings because they are widely used in the community, are available for several languages and trained on the same dataset for each language, which minimises the possible difference in performance. For the English language for InsuranceQA, we also tried custom word2vec [30] embeddings pre-trained on Wikipedia[1] with similar results.

The number of units per two layers of Shared BiLSTM is 96 and 64 respectively. Parameters are randomly initialised, and the initial state is set to zero. The models are implemented in Keras [9] with Tensorflow [1] backend for SemEval and PyTorch [32] for InsuranceQA. The optimiser is Adam [25] with a learning rate of 0.001 and batch normalization is used in Keras implementation. For PyTorch, optimiser is SGD with a learning rate of 1.1 (following the original implementation [39]). Training on a single GPU (NVIDIA TITAN Xp) takes approximately 30 and 15 minutes per epoch for PyTorch and Keras respectively.

4.1 InsuranceQA

Statistics This dataset contains non-factoid questions and answers from the insurance domain. It consists of a training set, a validation set, and two test sets, which in practice can be combined. Table 1 presents the statistics of the dataset. There are two versions of the dataset available, the main difference between them

[1] The parameters are as follows: skip-gram, window 5, negative-sampling rate -1/1000.

Table 1. Dataset statistics (InsuranceQA v2)

	Train	Validation	Test
# Questions	12889	2000	2000
# Answers	21325	3354	3308

being the construction of the wrong answers pool: it is either sampled randomly or retrieved with SOLR. We use the texts from the second version, as they are not lemmatised and as such are better suited for machine translation, but keep the pools random as in the first version, as such setup is better studied. Besides, the SOLR setup appears to be much more challenging. More specifically, the model trained on random pools achieves a validation accuracy score of 0.6241, and test scores of 0.6223 and 0.5987 respectively. In spite of this, it only obtains less than 0.1 accuracy when tested on SOLR pools. The pool size is 50 for the training set to make computations feasible and 500 for validation and test. The texts have been translated from English to German with Google Translate.

The preprocessing step for English is limited to lowercasing words. For German, we additionally apply compound nouns splitting and compare the performance of [40] and [34]. The tokenisation was already performed by the dataset authors, and we have also tried the SpaCy tokeniser [21]. As can be seen from the Table 2, the correct choice of splitter and tokenisation is crucial for reducing the number of out-of-vocabulary words. While there exist several strategies for handling such words, we chose to omit them completely. Studying the influence of different strategies is reserved for the future. We also opted to use fixed word vectors, as we empirically found that training the embeddings resulted in poorer performance.

Table 2. Number of out-of-vocabulary words depending on a splitter and text embeddings used

Method		Embeddings	
Splitter	Tokenisation	Polyglot	FastText
[40]	SpaCy	27714	17057
	split	52596	40706
[34]	SpaCy	22387	5304
	split	30983	15501

Table 3. Performance on InsuranceQA v2 (accuracy). Texts are originally in English and translated into German.

Language	Validation	Test
English	0.6361	0.6448
German	0.5435	0.5654

Performance We compare the performance of the system on original English texts with that on translations to German. The loss function is margin ranking loss. The performance metric is top-1 accuracy. On English texts, we obtain the following scores: validation set - 0.6380, test set - 0.6297. On German texts the scores are significantly lower: validation is 0.5435 and test is 0.5654. Re-

search into classifying errors is already underway. Our first hypothesis is that the quality of machine translation might be the main source of errors. More specifically, translation changes the word order and might rephrase the salient content words. It is also known for omitting or incorrectly translating named entities and affecting the sentiment. Another possible error-introducing step is compound splitting, which affects the number of out-of-vocabulary words and can be crucial for a correct understanding of the question. Apart from that, we expect different languages to influence the LSTM performance because of varying long-range dependencies length, as well as the number of function words.

4.2 SemEval

Table 4. Dataset statistics (SemEval 2017 Task 3 D)

	Train	Development	Test
# Questions	1031	250	1400
# Answers	30411	7384	12600

Table 5. Performance on SemEval 2017 (MAP) Texts are originally in Arabic and translated into English.

Language	Test
English	0.4997
Arabic	0.4939

Statistics The SemEval-2017 Task 3 [31] is concerned with community QA. Its different subtasks cover question-comment, question-external comment and question-question similarity. The subtask D focuses on the Arabic language, and the task is to rank new answers for a given question. The set of 30 related questions retrieved by a search engine is given, and each is supported by one correct answer. The resulting set of answers should be ranked based on their relevance to the given question. There are three possible labels - Direct, Relevant and Irrelevant - for an answer, but during the evaluation Relevant and Direct are grouped as a single label. The dataset is divided into training, a development and a test set. Their corresponding statistics are reported in Table 4.

Arabic is believed to be one of the most challenging languages [2] for automated processing, because of its morphological richness, free word order, and the mix of dialect and standard spelling. One of the problems we encountered was a large number of out-of-vocabulary words, which can be connected to the informal nature of the texts (slang, code mixing, typing errors, etc.). We have created an additional dictionary mapping out-of-vocabulary (OOV) words to their synonyms. Synonyms were obtained by translating an Arabic word into English and back into Arabic with Google Translate. Theoretically, such procedure should return the most common meaning and form, thus allowing us to reduce the vocabulary gap. In practice, we first preprocessed 62 161 OOV words to exclude numbers and cases when a word was concatenated with a number. After this, 53 344 OOV were left. Next, we successfully extracted 23 445 synonyms. 29 899 words were still not present in the FastText vocabulary. The number of OOV words is still relatively large and might be critical for the performance if

the important content words are not present in the vocabulary. As a possible solution, one can train custom word embeddings on a corpus with texts closer in spirit to the online fora.

Performance We compare the performance of the system trained and test on original Arabic texts with the one using translations to English. The loss function is cross-entropy, and the performance metric is a MAP. The evaluation is carried out with the official SemEval script. For the original Arabic texts, we obtain a test score of 0.4997. It is noticeably lower than the strong Google baseline of 0.6055, and we are now in the process of establishing the exact reasons for that. Contrary to expectations, for translated texts, the test MAP score is 0.4939, which is remarkably similar. We are investigating the reasons at the moment.

5 Discussion

Considering the challenges mentioned above, we suggest that further research should be undertaken in the following areas:

1. (Semi-) Automated collection of multilingual QA corpora. A procedure outlined in section 2 of this paper might be adopted by other research groups in most widely spoken languages, such as Chinese, Arabic and Hindi.
2. Incorporation of response-based machine translation.
3. Interpretation and comparison of cross-lingual QA deep learning models with monolingual ones. It may be assumed that the features and the behaviour of the model will change with respect to the language, and thus it is of interest to find what aspects stay universal and what change, as well as an explanation for this.
4. Code-mixed language detection and translation as a part of QA pipeline. Further investigation is required to assess whether including these components in joint training with QA model is beneficial.

More broadly, there are many research questions in need of further study. Among some of them are:

- Do some classes of languages require fewer data and less time for deep learning models to reach a specified performance? What are the properties of the languages that might affect the performance? Is there a universal neural architecture for all languages? Are some languages more suitable for LSTM-based architectures and others for CNN-based ones?
- How does a translation to English affect performance? How far can a system go without machine translation? Can we efficiently transfer a QA model from one language to another just by machine-translating texts? Can it be done per some categories of questions better than for others? Are some machine translation metrics more suitable in this setting?
- How well do cross-lingual embeddings work in QA setup? Are some types of cross-lingual embeddings better suited for particular language pairs?

6 Conclusion

In conclusion, cross-lingual and multilingual QA has attracted significant attention in the past. However, in the classical approach, techniques were mainly limited to machine translation of input texts and manual feature engineering. In the recent years, deep learning techniques for natural language processing have been developed which allow us to approach the problem in a new way. Nonetheless, it currently remains a challenging, yet neglected area. It is quite likely that the lack of research in the area may hinder the usage of more advanced dialogue systems and machine-human interfaces, if not addressed.

We have demonstrated that simply translating texts is not a sufficient solution, as it results in a significant drop in performance in some cases, and is not applicable to code-mixing or cross-script scenario. This paper also has highlighted existing problems with resources for multi- and cross-lingual applications. Moreover, it provides an agenda for collecting parallel QA corpora and gives an account of recent promising developments in the field.

We are currently in the process of collecting parallel data and investigating the effect of different preprocessing steps. Our future work will also concentrate on neural approaches. In particular, we are working on joint training of a machine translation and QA components, as well as experiments with cross-lingual embeddings for the code-mixed scenario.

Our work is still in progress. Nevertheless, we believe it could be a starting point, and we hope to attract more attention to the discussed area.

7 Acknowledgements

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) through the project DEEPLEE (01IW17001).

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org/, software available from tensorflow.org
2. Almarwani, N., Diab, M.: Gw_qa at semeval-2017 task 3: Question answer reranking on arabic fora. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 344–348 (2017)
3. Banerjee, S., Chakma, K., Naskar, S.K., Das, A., Rosso, P., Bandyopadhyay, S., Choudhury, M.: Overview of the mixed script information retrieval (msir) at fire-2016. Organization (ORG) **67**, 24 (2016)

4. Banerjee, S., Naskar, S.K., Rosso, P., Bandyopadhyay, S.: The first cross-script code-mixed question answering corpus. In: MultiLingMine@ ECIR. pp. 56–65 (2016)
5. Boldrini, E., Ferrández, S., Izquierdo, R., Tomás, D., Vicedo, J.L.: A parallel corpus labeled using open and restricted domain ontologies. In: International Conference on Intelligent Text Processing and Computational Linguistics. pp. 346–356. Springer (2009)
6. Bouma, G., Kloosterman, G., Mur, J., Van Noord, G., Van Der Plas, L., Tiedemann, J.: Question answering with joost at clef 2007. In: Workshop of the Cross-Language Evaluation Forum for European Languages. pp. 257–260. Springer (2007)
7. Chakma, K., Das, A.: Cmir: A corpus for evaluation of code mixed information retrieval of hindi-english tweets. Computación y Sistemas **20**(3), 425–434 (2016)
8. Chandu, K.R., Chinnakotla, M., Black, A.W., Shrivastava, M.: Webshodh: A code mixed factoid question answering system for web. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 104–111. Springer (2017)
9. Chollet, F., et al.: Keras. https://github.com/fchollet/keras (2015)
10. Cimiano, P.: Flexible semantic composition with dudes. In: Proceedings of the Eighth International Conference on Computational Semantics. pp. 272–276. Association for Computational Linguistics (2009)
11. Costa, Â., Luís, T., Ribeiro, J., Mendes, A.C., Coheur, L.: 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-2012) (2012)
12. Du, X., Shao, J., Cardie, C.: Learning to ask: Neural question generation for reading comprehension. arXiv preprint arXiv:1705.00106 (2017)
13. Espana-Bonet, C., Comas, P.R.: Full machine translation for factoid question answering. In: Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra). pp. 20–29. Association for Computational Linguistics (2012)
14. Feng, M., Xiang, B., Glass, M.R., Wang, L., Zhou, B.: Applying deep learning to answer selection: A study and an open task. In: Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on. pp. 813–820. IEEE (2015)
15. Forner, P., Peñas, A., Agirre, E., Alegria, I., Forăscu, C., Moreau, N., Osenova, P., Prokopidis, P., Rocha, P., Sacaleanu, B., et al.: Overview of the clef 2008 multilingual question answering track. In: Workshop of the Cross-Language Evaluation Forum for European Languages. pp. 262–295. Springer (2008)
16. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. The Journal of Machine Learning Research **17**(1), 2096–2030 (2016)
17. Ghosh, S., Ghosh, S., Das, D.: Complexity metric for code-mixed social media text. arXiv preprint arXiv:1707.01183 (2017)
18. Haas, C., Riezler, S.: Response-based learning for machine translation of open-domain database queries. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1339–1344 (2015)
19. Hadla, L.S., Hailat, T.M., Al-Kabi, M.N.: Evaluating arabic to english machine translation. Editorial Preface **5**(11) (2014)

20. Hakimov, S., Jebbara, S., Cimiano, P.: Amuse: Multilingual semantic parsing for question answering over linked data. In: International Semantic Web Conference. pp. 329–346. Springer (2017)
21. Honnibal, M., Johnson, M.: An improved non-monotonic transition system for dependency parsing. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1373–1378 (2015)
22. Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al.: Google's multilingual neural machine translation system: enabling zero-shot translation. arXiv preprint arXiv:1611.04558 (2016)
23. Joty, S., Nakov, P., Màrquez, L., Jaradat, I.: Cross-language learning with adversarial neural networks: Application to community question answering. arXiv preprint arXiv:1706.06749 (2017)
24. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
26. Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Peñas, A., Peinado, V., Verdejo, F., de Rijke, M.: Creating the disequa corpus: a test set for multilingual question answering. In: Workshop of the Cross-Language Evaluation Forum for European Languages. pp. 487–500. Springer (2003)
27. Magnini, B., Romagnoli, S., Vallin, A., Herrera, J., Penas, A., Peinado, V., Verdejo, F., de Rijke, M.: The multiple language question answering track at clef 2003. In: Workshop of the Cross-Language Evaluation Forum for European Languages. pp. 471–486. Springer (2003)
28. Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., De Rijke, M., Rocha, P., Simov, K., Sutcliffe, R.: Overview of the clef 2004 multilingual question answering track. In: Workshop of the Cross-Language Evaluation Forum for European Languages. pp. 371–391. Springer (2004)
29. Martino, G.D.S., Romeo, S., Barrón-Cedeno, A., Joty, S., Marquez, L., Moschitti, A., Nakov, P.: Cross-language question re-ranking. arXiv preprint arXiv:1710.01487 (2017)
30. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
31. Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., Verspoor, K.: Semeval-2017 task 3: Community question answering. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 27–48 (2017)
32. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
33. Raghavi, K.C., Chinnakotla, M.K., Shrivastava, M.: Answer ka type kya he?: Learning to classify questions in code-mixed language. In: Proceedings of the 24th International Conference on World Wide Web. pp. 853–858. ACM (2015)
34. Riedl, M., Biemann, C.: Unsupervised compound splitting with distributional semantics rivals supervised methods. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 617–622 (2016)
35. Ruder, S.: A survey of cross-lingual embedding models. arXiv preprint arXiv:1706.04902 (2017)

36. Sacaleanu, B., Neumann, G., Spurk, C.: Dfki-lt at qa@ clef 2008. In: Workshop of the Cross-Language Evaluation Forum for European Languages. pp. 429–437. Springer (2008)
37. Sasaki, Y., Lin, C.J., Chen, K.h., Chen, H.H.: Overview of the ntcir-6 cross-lingual question answering task. In: Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, May 15-18. pp. 153–163. Citeseer (2007)
38. Sugiyama, K., Mizukami, M., Neubig, G., Yoshino, K., Sakti, S., Toda, T., Nakamura, S.: An investigation of machine translation evaluation metrics in cross-lingual question answering. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 442–449 (2015)
39. Tan, M., dos Santos, C., Xiang, B., Zhou, B.: Improved representation learning for question answer matching. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 464–473 (2016)
40. Tuggener, D.: Incremental coreference resolution for German. Ph.D. thesis, Universität Zürich (2016)
41. Ture, F., Boschee, E.: Learning to translate for multilingual question answering. arXiv preprint arXiv:1609.08210 (2016)
42. Vallin, A., Magnini, B., Giampiccolo, D., Aunimo, L., Ayache, C., Osenova, P., Peñas, A., De Rijke, M., Sacaleanu, B., Santos, D., et al.: Overview of the clef 2005 multilingual question answering track. In: Workshop of the Cross-Language Evaluation Forum for European Languages. pp. 307–331. Springer (2005)
43. Veyseh, A.P.B.: Cross-lingual question answering using common semantic space. In: Proceedings of TextGraphs-10: the Workshop on Graph-based Methods for Natural Language Processing. pp. 15–19 (2016)