

# Towards End-to-End Multilingual Question Answering

Ekaterina Logina<sup>1</sup><sup>[0000-0002-0911-0213]</sup>, Stalin Varanasi<sup>2</sup>, and Günter Neumann<sup>2</sup>

<sup>1</sup> University of Ghent

{ekaterina.loginova}@ugent.be

<sup>2</sup> DFKI, Saarbrücken, Germany

{stalin.varanasi, neumann}@dfki.de

**Abstract.** Multilingual question answering (MLQA) is a critical part of an accessible natural language interface. However, current solutions demonstrate performance far below that of monolingual systems. We believe that deep learning approaches are likely to improve performance in MLQA drastically. This work aims to discuss the current state-of-the-art and remaining challenges. We outline requirements and suggestions for practical parallel data collection and describe existing methods, benchmarks 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 overview will re-ignite interest in multilingual question answering, especially with regard to neural approaches.

**Keywords:** question answering · multilingual natural language processing · neural natural language processing · deep learning · multilingual question answering · cross-lingual question answering

## 1 Introduction

Natural language processing (NLP) systems are mostly monolingual. For English language, researchers achieved remarkable performance in the area of information retrieval and question answering. Still, few systems efficiently integrate and present knowledge across several languages. As a result, the research community is virtually ignoring large data available in languages other than English [7]. It is especially relevant for opinionated information, such as news, blogs and social media. In the era of fake news and deliberate misinformation, NLP systems turn out to be biased by design, as they mainly take English language data into account. Furthermore, selecting English as the primary development language is poorly motivated regarding the number of users: Mandarin and Spanish have more native speakers than English, and Hindi and Arabic follow closely. Developing multilingual solutions will thus lead to more equal and convenient information access for millions of people.

In this paper, we focus on Question Answering (QA) task. The goal of QA is to find an answer  $a$  to the query  $q$  in the collection of the documents  $\{D\}_i^N$ . In the cross-lingual setting,  $q$  and  $\{D\}_i^N$  are in different languages. For example, the users ask a question in Portuguese, and we search an English document for an answer. The result can be either translated to Portuguese as well or left as is. Previously, devising such systems required expensive manual feature engineering and linguistic resources. Traditional approaches handle multilingual QA by translating either query or documents and converting the problem to a more familiar monolingual setting. After that, the general workflow involves three steps. First, we need to determine the languages of the query  $q$  and the documents  $D$  (a language identification problem). Second, we need to use information retrieval to search for relevant documents containing candidate answers. Finally, depending on the type of QA, we need to either rank the candidates according to how likely they are to be the right answer or extract an exact answer. In the traditional approach, researchers employ text classification methods for the former and information retrieval with rule-based patterns for the latter.

However, this conventional approach has several shortcomings. One of the issues is that there might be multiple languages in one document. Moreover, both query and documents might contain transliterated text from non-Roman alphabets. Besides, speakers of languages using non-Roman based alphabets often transliterate words, which leads to high spelling variations. While language identification for monolingual texts on a document level was widely addressed before, for multi-lingual documents and word-level, it is still an open research area. Another critical problem with the traditional approach is its dependence on machine translation quality. Furthermore, we often require language-dependent tools (such as POS-taggers and NER-recognisers) to perform machine translation and information retrieval. Developing such tools is costly and time-consuming.

With the advent of deep learning approaches, which demonstrate spectacular performance while working in an end-to-end fashion, we strongly feel it is time to reconsider the state of multilingual automated systems. In this work, we provide an overview of the current state of the field for multilingual (MLQA) and cross-lingual (CLQA) subtasks. We also include a preliminary analysis of the performance for a deep learning model in multiple language setting. Our research objective is to compare the performance of the same model on original Arabic (English) texts and their English (German) translations.

This paper is divided into six sections. First, we overview how deep learning has transformed the fields closely related to MLQA. The second 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 third section examines approaches to the problem and reports state-of-the-art results on several shared tasks. In the fourth 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 fifth section. Our conclusions are drawn in the final section.

## 2 Deep learning in related fields

**Monolingual QA.** Deep learning approaches for NLP demonstrate excellent results and do not require manual feature engineering [86]. They have been successfully applied to QA tasks in English, surpassing human performance on SQuAD - a large collection of more than 100,000 questions [62]. The release of this huge dataset ignited an active competition between teams all over the world, resulting in the development of a plethora of the neural architectures. The majority of them were variations of LSTM [31] and PointerNet [79] combinations with advanced neural attention components [3]. Most remarkable ones include BiDAF [68], match-LSTM [81], Multi-Perspective Context Matching [82], Dynamic Coattention Networks [84], DRQA[13], FastQA [83], Ruminating Reader [26], and, finally, the current leader - BERT [19]. Participating teams introduced novel features of neural attention such as multi-stage hierarchical attention, co-attention, gated multi-hop attention [26] and Transformers, which completely forego recurring or convolutional neural structures [19], [77]. In less than three years, the performance increased from the baseline F1 score of 51% to astonishing 93.16%.

**Machine Translation.** The widespread adoption of neural machine translation methods started with the introduction of encoder-decoder architecture and attention concept in 2014 [3]. One of the recent breakthroughs in the area enables neural zero-shot translation, allowing us to translate between pairs of languages which the system has not seen before [37]. This remarkable result can alleviate the issue of requiring multiple corpora for MLQA. Another interesting finding is that neural machine translation implicitly learns the shared semantic space for several languages [65]. In the case of [37], they have visualised this space and concluded that it might contain universal interlingua representations, encoding semantics of the phrases and not just the alignment of translations. Such interlingua space also harbours the potential for immense improvement in the performance of MLQA systems.

**Cross-lingual embeddings.** The shared semantic space can also be learned explicitly by using cross-lingual word embeddings. In general, word embeddings map natural language words to numerical vectors. The enticing aspect of word embeddings is that they can learn and link concepts and organise them in hierarchies in an unsupervised way [52]. Cross-lingual embeddings additionally allow us to transfer knowledge between languages and reason about semantics in multilingual context [65]. The wide range of approaches is discussed in detail in [65]. An empirical comparison of different approaches can also be found in [75]. The study claims that the benefit of cross-lingual embeddings is more for semantic tasks than syntactic. Regarding QA and adjacent fields, cross-lingual embeddings have been applied in information retrieval for Dutch and English [80].

**Text Classification.** One of the earliest studies on deep learning for this task applied convolutional neural networks (CNNs) for sentence classification [41]. Since that, hierarchical attention networks [85] and recurrent convolutional networks [44] have also been successfully employed. In multilingual context,

CNNs were used for sentiment classification in English, German and Arabic [2]. Concerning more recent developments, Universal Language Model Fine-Tuning (ULMFiT) significantly out-performs previous classification models and enables sample-efficient transfer learning for any NLP task [34]. ULMFiT is open-sourced, which eases the adoption of the approach.

**Language Identification.** Despite the claims that language identification is a solved problem, recent studies have shown that without simplifying assumptions, it is still a rather challenging area [46]. Most of the solutions use traditional machine learning approaches, such as Naive Bayes and Support Vector Machines [36]. However, some recent studies have also implemented deep learning based approaches. For example, [35] achieves F1 score of 77.1 using a combination of CNN and LSTM on monolingual documents. The architecture allows word-level prediction on code-switching documents as well. *Code-mixing*, a linguistic phenomenon describing the usage of several languages simultaneously in one text, is one of the main challenges in the area. Language identification in code-mixed texts has been addressed during shared tasks on EMNLP 2014 and FIRE 2014, as well as ACL’s Workshops on Computational Approaches to Code Switching [6]. The first ACL workshop covered the following language pairs: Modern Standard Arabic/Dialectal Arabic, Mandarin-English, Nepali-English and Spanish-English [70]. For token-level language identification, the best average F1 score ranged from 0.799 to 0.959. The result heavily depends on the relatedness of languages. The second workshop covered tweets with Arabic dialects and Spanish/English mixing [54]. Weighted F1 over code-switched and monolingual texts for Spanish/English was 0.913, and for Modern Standard Arabic/Dialectal Arabic it was 0.83. FIRE 2014 Transliterated Search Task [17] investigated word-level labelling in documents written in a mix of English with one of six Indic languages. Depending on the language, the Exact Query Match Fraction ranged from 0.218 to 0.847. Deep learning (more precisely, auto-encoders) has been used to improve the results for Hindi, reaching Mean Average Precision (MAP) of 0.50 and Mean Reciprocal Rank (MRR) of 0.87 [27]. While machine learning methods achieve decent accuracy for language identification in mixed-language documents, they operate under the assumptions that the languages are known a priori [42]. This assumption, which does not reflect real-world scenarios, is a major drawback for the majority of works. However, some recent studies report encouraging results without assuming the language pair at inference time: the average token-level accuracy is 93.4 over English, Spanish, Czech, Basque, Hungarian, Croatian, Slovak, and Hindi [87]. The developed system, based on feed-forward neural networks, is especially appealing as it is fast, compact, and robust to informal style.

### 3 Datasets

#### 3.1 Existing datasets

**Parallel** A parallel corpus is a term from machine translation, meaning a corpus in which the source and the target texts are aligned. The construction of a

reusable, multilingual collection of questions with the related answer-document pairs has been the target of Cross-Language Evaluation Forum (CLEF) [49]. Over the course of several years, CLEF provided the following multilingual datasets: Multisix corpus [48], (200 questions, 6 languages), the DISEQuA corpus [47] (450 questions, 4 languages), the Multieight-04 corpus [49] (700 questions, 7 languages), and the Multi9-05 [76] (900 questions, 9 languages). Some corpora also include question type in their annotation [8]. However, CLEF only covers European languages. Regarding Asian languages, a small parallel corpus for Japanese, Chinese and English was constructed by the NTCIR organisers [67].

However, the existing parallel corpora are inadequately small compared to monolingual benchmarks, such as SQuAD. At the time of publication, to the best of our knowledge, there are no parallel corpus for QA sufficiently big to take advantage of deep learning techniques fully. Thus, creating a large, high-quality parallel corpora for MLQA is a challenging yet neglected area of research.

**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. A sentence can include combinations of transliterated English stems with native affixes or switching alphabets (cross-script). For instance, the word "rasharit" in Russian. It means "to share" in the context of web files, and it has a stem "r" which is a transliteration of English "share". English roots are also used with German morphological rules, especially in computer science terms: "Das Programm wurde upgedated" ("the program has been updated").

Despite extensive use of code-mixing in informal conversations, there are few QA datasets present. Among recent developments in this area, we can note code-mixed and cross-script QA corpus [5]. 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 [11]. This dataset consists of 1959 code-mixed tweets. FIRE 2015 feature cross-script datasets for information retrieval [69].

### 3.2 Dataset collection and analysis

An MLQA corpus dealing with the above problems 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, the 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-sourced 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. Besides, 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. The number of native speakers who do not speak English can also be a criterion to choose the language for system development. 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):

1. closely related languages with similar alphabets (Italian, Spanish),
2. distant languages with similar alphabets (Danish, French),
3. closely related languages with different alphabets (Polish, Russian),
4. distant languages with different alphabets (English, Chinese).

The properties of MLQA datasets have not been dealt with in depth. We argue that the following characteristics need to be considered: diversity and balance of answer and question types, reasoning type for questions along with the difficulty score, whether the reasoning over multiple sentences or documents is required, 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, which 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. Besides, in a code-mixed scenario, an appropriate metric should be chosen to evaluate the complexity of a corpus, such as the one proposed in [25].

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 [20] is promising for doing so in an end-to-end (semi-) automated fashion, which is essential 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. Despite its appeal, machine translation proved to be an over-simplistic way to create multilingual corpora [45]. The authors create the Korean Question Answering Dataset (K-QuAD) by the semi-automated approach, by automatically translating SQuAD. They concluded that it is necessary to build

language-specific resources. Some authors suggest that the multilingual version of knowledge graphs can be used, such as ad DBPedia [32].

## 4 Methods

In general, one can distinguish the following approaches [66]:

- use machine translation directly beforehand or as part of a QA system; and then to work in the monolingual setup of a target language
  - translate the queries into the target language
  - translate the document collection into the source language
  - translate the queries and the documents into an intermediate representation (interlingua)
- map terms in several languages to a multilingual knowledge base or a semantic graph, such as Wikipedia or BabelNet [9]
- use a universal cross-language representation (such as cross-lingual embeddings)

These three broad categories are discussed in more detail below.

An important question for MLQA is what can be seen as a universal baseline? Clearly defining the baseline approach which can be used in any MLQA study will make a comparison of different approaches easier. 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 [37], [50]. A major drawback is an unequal quality of translation for different languages, which should be taken into account during a comparison.

### 4.1 Machine Translation

Previously, translation relied on lexical resources, such as dictionaries and aligned wordnets. The next step introduced machine learning methods. Nowadays, zero-shot translation alleviates the need for extensive lexical resources and eases the transfer of models to other languages. 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. [74] 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 English translations baseline. Their findings also do not support the hypothesis that learning a custom classifier for each language would outperform the single classifier baseline. As a generalisation of this idea, [28] 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.

Despite its extended use, machine translation remains a source of errors in multilingual text processing pipelines. A loss or corruption of named entities has frequently been observed during translation. Code-mixed texts are even more challenging in this regard [37]. Finally, [71] 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. The authors claim that the conversion of entities into logical forms, typical for methods utilising a knowledge base, can be profoundly affected by a translation. Another potential issue is a change in the word order, which might harm the performance of predicate construction and merging. Nevertheless, the authors conclude that the QA system and humans do estimate the translation quality in a very different way. Machine translation errors seem to have more impact on human evaluation, than on the whole system result [53]. In the context of sentiment analysis, sarcasm, metaphors, and word order were not adequately processed from a human’s viewpoint but did not significantly impact classification results.

#### 4.2 Structured Semantic Information

Regarding the latest machine learning models, [30] considers the use of semantic parsing for MLQA over linked data. The authors propose a model that utilises DUDES (Dependency-based Underspecified Discourse Representation Structures) [18] universal dependencies. 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. [78] 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 the question into English with Google Translate. Another group of approaches aims to transform the natural language question into a language-independent semantic formula. An example is QAKIS, which produced the SPARQL query and used it to query multilingual DBpedia [10]. MTransE also worked with multilingual knowledge graphs, embedding them to synchronise knowledge bases across several languages [15]. QALL-ME, a reusable architecture for MLQA powered by ontologies, highlighted the importance of spatial-temporal context [22].

#### 4.3 Shared Semantic Representation

In cross-lingual question retrieval, the system receives a question in the source language and searches for similar questions in the target language. This task is especially relevant for large community question answering platforms such as

StackOverflow, as it provides the users with the opportunity to ask questions in their native language, but address the entire knowledge base. However, existing approaches suffer from word mismatch and ambiguous question formulations. The baseline approach of translating the queries does not preserve the semantics, and the lexical gap proves to be a challenge for question retrieval methods. A potential way to tackle this is to use cross-lingual embeddings. They allow us to use a dual-language vector space to directly represent the semantic similarity between questions in two different languages. The resulting system would be more robust because it would not rely on the exact lexical similarity of the queries. Cross-lingual embeddings have been used in combination with a feed-forward neural network on Arabic and English data of SemEval competitions [50]. Another exciting result on the same dataset has been reported by [38], who developed a new method based on the Domain Adversarial Neural Network model [24]. They have adapted it to a cross-lingual task by coupling a detection network with a question-answering one. Moreover, CNNs were also employed in some novel solutions for this task [14]. The MAP score increased from 0.095 to 0.289, which hints at the potential advantage of deep learning methods over more traditional approaches. However, neural approaches are yet to outperform machine translation on a wide range of datasets. A recent study achieves a MAP of 0.455 on Yahoo! Answers data using non-negative matrix factorisation to integrate the knowledge from translated representations [88].

#### 4.4 Benchmarks

**CLEF.** As mentioned above, one of the most popular MLQA challenges is the CLEF campaigns. 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 (mean reciprocal rank). [23] distinguishes three "eras" in the campaign, different in the task and dataset properties. The first era covered the period from 2003 to 2006 and required participants to answer mostly factoid questions based on monolingual newspapers (ELRA/ELDA). The next era, until 2008, grouped questions by type and added Wikipedia as a source. Finally, the last competition in 2009 featured multilingual parallel documents from the law domain, and the task was to return supporting passages. More details can be found in [23]. While the performance of monolingual QA systems for major European languages have improved over the years, cross-lingual systems remained an unsolved challenge. Once again, it is attributed to the low quality of machine translation, especially regarding named entities. We now provide a brief overview of each CLEF competition.

CLEF-2004 covered six monolingual (Dutch, French, German, Italian, Portuguese and Spanish) and 50 cross-lingual tasks. The document collections consisted of news articles. Each target language had 200 questions with almost no overlap in test sets. CLEF-2005 featured nine target languages and ten source languages, resulting in 8 monolingual and 73 cross-lingual tasks. CLEF-2006 introduced list questions and required the extraction of passages supporting the an-

swer. In CLEF-2007, topics were introduced to cluster QA pairs, and Wikipedia search was made possible for the competing systems. Answer Validation Exercise and Question Answering in Speech Transcription tasks were added. While CLEF-2006 has seen an increase in the performance for most tasks, in CLEF-2007 the results dropped significantly. The outcome motivated the organisers to relax the evaluation conditions, and in CLEF-2008 the monolingual results increase. However, cross-lingual performance decreased substantially. The organisers hypothesise that the decrease in the accuracy could have been due to linked questions. Overall, the topic resolution was demonstrated to be a problem for MLQA systems.

**NTCIR.** Another important shared task is NTCIR-6 [67]. The target languages are English, Chinese and Japanese. The corpus is based on newspaper articles. There can be only one or no answer, and its type is restricted to a named entity. It was not required to translate answers back into the source language. The performance metric is top-1 accuracy and MRR. The comparison of systems’ performance over the years is provided in the table below. X-Y indicates that questions are given in language X and answers are extracted from documents in language Y. A crucial advantage of NTCIR data is that the question sets were made truly parallel. In NTCIR-7 the nugget pyramid evaluation method was added to enable human-in-the-loop evaluation.

Similarly to CLEF, the major issue in NTCIR-5 was machine translation of named entities. Translated questions had expressions different from established idioms used in news articles containing the answer. Keywords were also mistranslated, encumbering information retrieval. However, a curious observation is that sometimes questions could be answered correctly in the cross-lingual but not monolingual setting.

<b>NTCIR-</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>
<b>C-C</b>	0.445	0.547	CS-CS	0.433 0.461
			CT-CT	0.267 0.283
<b>C-EN</b>	0.065			
<b>EN-C</b>	0.165	0.34	EN-CS	0.221 0.289
			EN-CT	0.204
<b>EN-JA</b>	0.155			0.163 0.211
<b>JA-EN</b>	0.31			
<b>JA-JA</b>		0.36		0.22 0.226

**Table 1.** Evolution of NTCIR results (accuracy). CS = Simplified Chinese, CT = Traditional Chinese, EN = English, JA = Japanese

**SemEval.** Community Question Answering (CQA) was covered in Task 3 of SemEval competitions in 2015, 2016 and 2017. SemEval-2015 included two subtasks [57]. For subtask A, teams classified answers as good or bad, and for subtask B they answered yes/no questions. For the first subtask, two datasets were available - in Arabic and English, but they were not parallel. The best F1

score was 57.19 and 63.7 for the English subtasks A and B respectively, and 78.55 for the Arabic subtask A. SemEval-2016 featured three subtasks for English language: QuestionComment Similarity (subtask A), QuestionQuestion Similarity (B), and QuestionExternal Comment Similarity (C) [58]. For Arabic, there was a separate subtask D, where teams needed to rerank the correct answers for a new question. The best solutions achieved a MAP score of 79.19, 76.70, 55.41, and 45.83 in subtasks A, B, C, and D, respectively. In 2017, these subtasks remained, with the addition of subtask E: Multi-domain Question Duplicate Detection [56]. Besides, new data were added. The best MAP scores were 88.43, 47.22, 15.46, and 61.16 in subtasks A, B, C, and D, respectively. In all three competitions, Support Vector Machines (SVM) were the most popular machine learning approach. However, the share of neural network solutions steadily increased, starting from the use of word embeddings as features. CNNs, LSTMs and Feedforward Neural Networks were applied.

While SemEval included datasets in two languages, it did not cover CLQA in the official evaluation. However, SemEval 2016 dataset was used for cross-language question re-ranking [50], [38]. [50] worked with cross-lingual embeddings to answer Arabic questions using English documents, and achieved the MAP is 77.14, which is almost the same as the monolingual system score of 77.41. Translating the questions lowered the MAP slightly to 76.67. [38] approached another subtask - question-question similarity - by using adversarial neural networks. They achieved a MAP of 76.65 with Arabic as a target language and 76.63 with English.

**QALD.** QA over Linked Data (QALD) is a special type of QA which aims to translate the query to a form compatible with the semantic web. The participants transform a natural language question into a SPARQL query which retrieves answers from RDF datasets and other knowledge sources. QALD-3 introduced the multilingual aspect by providing three datasets: MusicBrainz, English and Spanish DBpedia’s. Knowledge sources were large (DBpedia contains hundred million RDF triples), but the training and testing question collections were rather small (less than a hundred). The questions were general, open-domain and factoid, of varying complexity. In QALD-3, six European languages were considered: English, Spanish, German, Italian, French and Dutch. Some of the questions could not be answered based on the given datasets. Most systems achieved F score between 32% and 36%. QALD-4 added Romanian to the list of languages, extended training dataset to 200 questions, and introduced two additional tasks. The first was hybrid question answering, which required the integration of both structured and unstructured information, and the second focused on the biomedical domain. The best MLQA system demonstrated 0.72 F1. QALD-5 further increased the training dataset size to 350 questions, and QALD-6 added Farsi language. QALD-6 also witnessed a significant increase in quality: the best F1 score was 0.89. However, in QALD-7 the test set questions increased in complexity (by demanding, for example, mathematical operations). As a result, the performance dropped: the best F1 score was 0.720 for French, and 0.469 for English. In QALD-8 challenge, the best F1 score was 0.388, and a

modified F1 QALD metric was introduced. A drop in performance is attributed to test set "curve balls": queries were extracted from logs or search engines to make the evaluation more realistic. Finally, the most recent QALD-9 challenge featured the largest training set (408 questions) and covered the most languages (11). The leading team achieved F1 score of 0.298. QALD also measured a modification of F1 score adapted to the task, which for the winning team was 0.430.

**MSIR & FIRE.** Concerning code-mixed and cross-script 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 [4]. In the question classification task, [60] report an accuracy of 45.00%. [12] 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.

**Common issues.** The datasets used in previous competitions have some common disadvantages. First of all, the majority of questions do not have answers in both corpora in the bilingual setting. Furthermore, the question types are not uniformly distributed. This leads to similar results: cross-lingual systems perform significantly worse than monolingual for most language pairs. For Asian languages in NTCIR, the cross-lingual performance is on average three times worse than monolingual, and the situation is similar for many European language pairs. Constructing a dataset that overcomes these issues is an open research problem.

## 5 Experiments

**Table 2.** Statistics of the datasets

Dataset statistics	InsuranceQA			SemEval		
	Train	Validation	Test	Train	Validation	Test
#Questions	12889	2000	2000	1031	250	1400
#Answers	21325	3354	3354	30411	7384	12600

In this section, we will describe our initial experiments with MLQA. The goal was to train and test neural models for target language using the machine translation of source language corpora.

We have performed our experiments on two datasets: InsuranceQA (version 2) [21] and SemEval 2017 (subtask D) [56]. 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 was used. However, to the best of our knowledge, there does not exist a single study demonstrating its performance on the above-mentioned languages. Therefore, we refer to the results from separated studies. For English - German language pair achieves a BLEU score of 24.60 [37]. For Arabic to English translation, the average precision is 0.449 [29].

The deep learning model we use is an attentional Siamese Bidirectional LSTM. The method is essentially the same as that introduced by [72] with some adjustments in hyper-parameters and loss function. 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 [56]). 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. Parameters are randomly initialised, and the initial state is set to zero.

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 [39]. 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 reduces performance variation. For the English language for InsuranceQA, we also tried custom word2vec [52] embeddings pre-trained on Wikipedia<sup>3</sup> with similar results.

The models are implemented in the Keras [16] for SemEval and PyTorch [59] for InsuranceQA. Training on a single GPU (NVIDIA TITAN Xp) takes approximately 30 and 15 minutes per epoch for PyTorch and Keras respectively.

## 5.1 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 2 presents the statistics of the dataset. There are two versions of the dataset available, the main difference between them being the construction of the wrong answers pool: it is either sampled randomly or retrieved with SOLR<sup>4</sup>. 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,

<sup>3</sup> The parameters are as follows: skip-gram, window 5, negative-sampling rate -1/1000.

<sup>4</sup> <http://lucene.apache.org/solr/>

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 on two test datasets respectively. Despite 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 lower-casing words. For German, we additionally apply compound nouns splitting and compare the performance of [73] and [63]. The dataset authors already performed the tokenisation, and we have also tried the SpaCy tokeniser [33]. The correct choice of splitter and tokenisation is crucial, as it reduces the number of out-of-vocabulary words from 40706 to 5304 and from 52596 to 22387 for FastText and Polyglot embeddings, respectively. While there exist several strategies for handling such words, we chose to omit them completely. Studying the influence of alternative strategies is reserved for the future. We also opted to use fixed word vectors, as we empirically found that training the embeddings resulted in reduced performance.

**Performance** For the InsuranceQA model, there is a single BiLSTM with the hidden size 141. Dropout with  $p = 0.5$  is applied on the output, but no learning rate decay. The loss function is margin ranking loss. The optimiser is SGD with a learning rate of 1.1 (following the original implementation [72]).

We compare the performance of the system on original English texts with that on translations to German. The performance metric is top-1 accuracy. On English texts, we obtain the following scores: validation set - 0.6361, test set - 0.6448. On German texts, the scores are significantly lower: respectively 0.5428 and 0.5507. Further research into errors is 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.

**Table 3.** Performance on InsuranceQA v2 (accuracy) and SemEval 2017 (MAP). For InsuranceQA, texts are originally in English and translated into German. For SemEval, texts are originally in Arabic and translated into English.

Performance	InsuranceQA		SemEval	
	English	German (translated)	Arabic	English (translated)
Validation	0.6361	0.5435		
Test	0.6448	0.5654	0.4997	0.4939

## 5.2 SemEval

**Statistics** The SemEval-2017 Task 3 [56] is concerned with community QA. Subtask D focuses on the Arabic language, and the task is to rank new answers for a given question. The dataset is divided into training, a validation and a test set. Their corresponding statistics are reported in Table 2. 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.

Arabic is believed to be one of the most challenging languages [1] 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 a 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. In the current version, we use a random embedding for OOV token.

The number of OOV words is relatively large and might be critical for the performance if the important content words are not present in the vocabulary. As a possible solution to high OOV rate, one can train custom word embeddings on a corpus with texts closer in style. A similar in spirit, but more in-depth approach to expand the query using concept linking has been recently proposed for the same dataset [61].

**Performance** For the SemEval model, the number of units per two layers of shared BiLSTM is 96 and 64 respectively. A simple regular expression based cleaning procedure was applied to texts to remove special characters. The loss is binary cross entropy, the optimiser is Adam [43] with a learning rate of 0.001, and batch normalisation as well as early stopping by F1 score on validation are used.

We compare the performance of the system trained and tested on original Arabic texts with the one using translations to English. The loss function is cross-entropy, and the performance metric is 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 provided by the organisers, and we are now in the process of establishing the exact reasons for that. Contrary to our expectations, for translated texts, the test MAP score is 0.4939, which is remarkably similar. However, it is still

lower than the baseline, which supports the idea that translating the question is not enough in MLQA.

## 6 Discussion

As can be seen from the experiments, naive approaches are not efficient enough in MLQA. More sophisticated yet generalisable solutions are desiderata. 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. Other research groups might adopt a procedure outlined in section 2 of this paper in most widely spoken languages, such as Chinese, Arabic and Hindi [51]. Spanish has also been mentioned as one of the most under-represented languages at ACL conferences [55].
2. Improving machine translation component. This can be done by either incorporating response-based machine translation or adding monolingual data and back-translations [64].
3. Interpretation and comparison of multilingual 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 why.
4. Code-mixed language detection and translation as a part of the QA pipeline. Further investigation is required to assess whether including these components in joint training with QA model is beneficial.
5. Effect of translation on different types of questions and the way they affect the performance of MLQA systems [40].

More broadly, there are many research questions in need of further study. 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 MLQA setup? Are some types of cross-lingual embeddings better suited for particular language pairs?

## 7 Conclusion

In conclusion, multilingualism is acknowledged as one of the main challenges of question answering [32]. Multilingual QA has attracted significant attention in the past. Despite several competitions on the topic, like QALD, CLEF and NT-CIR, current solutions are over-simplistic. In the classical approaches, solutions were mainly limited to machine translation of input texts and manual feature engineering. However, in recent years, deep learning techniques for natural language processing have been developed which allow us to approach the problem in a new way. Nonetheless, MLQA 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 merely translating texts is not a satisfactory solution, as it results in a significant drop in performance in some cases, and does not apply to code-mixing or cross-script scenario. The existing literature supports this finding. This paper has also highlighted existing problems with resources for multi- and cross-lingual applications. The critical issue is the absence of sufficiently large parallel QA corpora for most widely spoken languages. Motivated by discovered challenges, we provide an agenda for collecting parallel QA corpora and gives an account of recent promising developments in the field.

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.

## 8 Acknowledgements

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

## References

1. Almarwani, N., Diab, M.: Gw\_qa at semeval-2017 task 3: Question answer re-ranking on arabic fora. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 344–348 (2017)
2. Attia, M., Samih, Y., Elkahky, A., Kallmeyer, L.: Multilingual multi-class sentiment classification using convolutional neural networks. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). European Language Resource Association (2018), <http://aclweb.org/anthology/L18-1101>
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
4. 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)

5. 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)
6. Barman, U., Das, A., Wagner, J., Foster, J.: Code mixing: A challenge for language identification in the language of social media. In: Proceedings of the first workshop on computational approaches to code switching. pp. 13–23 (2014)
7. Bender, E.M.: Linguistically naïve!= language independent: why nlp needs linguistic typology. In: Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous? pp. 26–32 (2009)
8. 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)
9. 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)
10. Cabrio, E., Cojan, J., Aprosio, A.P., Magnini, B., Lavelli, A., Gandon, F.: Qakis: an open domain qa system based on relational patterns. In: International Semantic Web Conference, ISWC 2012 (2012)
11. 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)
12. 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)
13. Chen, D., Fisch, A., Weston, J., Bordes, A.: Reading wikipedia to answer open-domain questions. arXiv preprint arXiv:1704.00051 (2017)
14. Chen, G., Chen, C., Xing, Z., Xu, B.: Learning a dual-language vector space for domain-specific cross-lingual question retrieval. In: 2016 31st IEEE/ACM International Conference on Automated Software Engineering (ASE). pp. 744–755. IEEE (2016)
15. Chen, M., Zaniolo, C.: Learning multi-faceted knowledge graph embeddings for natural language processing. In: IJCAI. pp. 5169–5170 (2017)
16. Chollet, F., et al.: Keras. <https://github.com/fchollet/keras> (2015)
17. Choudhury, M., Chittaranjan, G., Gupta, P., Das, A.: Overview of fire 2014 track on transliterated search. Proceedings of FIRE pp. 68–89 (2014)
18. 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)
19. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
20. Du, X., Shao, J., Cardie, C.: Learning to ask: Neural question generation for reading comprehension. arXiv preprint arXiv:1705.00106 (2017)
21. 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)
22. Ferrandez, O., Spurk, C., Kouylekov, M., Dornescu, I., Ferrandez, S., Negri, M., Izquierdo, R., Tomas, D., Orasan, C., Neumann, G., et al.: The qall-me framework:

- A specifiable-domain multilingual question answering architecture. *Web semantics: Science, services and agents on the world wide web* **9**(2), 137–145 (2011)
23. 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)
  24. 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)
  25. Ghosh, S., Ghosh, S., Das, D.: Complexity metric for code-mixed social media text. *arXiv preprint arXiv:1707.01183* (2017)
  26. Gong, Y., Bowman, S.R.: Ruminating reader: Reasoning with gated multi-hop attention. *arXiv preprint arXiv:1704.07415* (2017)
  27. Gupta, P., Bali, K., Banchs, R.E., Choudhury, M., Rosso, P.: Query expansion for mixed-script information retrieval. In: *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. pp. 677–686. ACM (2014)
  28. 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)
  29. Hadla, L.S., Hailat, T.M., Al-Kabi, M.N.: Evaluating arabic to english machine translation. *Editorial Preface* **5**(11) (2014)
  30. 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)
  31. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* **9**(8), 1735–1780 (1997)
  32. Höffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J., Ngonga Ngomo, A.C.: Survey on challenges of question answering in the semantic web. *Semantic Web* **8**(6), 895–920 (2017)
  33. 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)
  34. Howard, J., Ruder, S.: Fine-tuned language models for text classification. *CoRR* **abs/1801.06146** (2018), <http://arxiv.org/abs/1801.06146>
  35. Jaech, A., Mulcaire, G., Hathi, S., Ostendorf, M., Smith, N.A.: Hierarchical character-word models for language identification. *arXiv preprint arXiv:1608.03030* (2016)
  36. Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., Lindén, K.: Automatic language identification in texts: A survey. *arXiv preprint arXiv:1804.08186* (2018)
  37. 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)
  38. 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)
  39. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759* (2016)

40. Kalouli, A.L., Kaiser, K., Hautli-Janisz, A., Kaiser, G.A., Butt, M.: A multilingual approach to question classification. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018) (2018)
41. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
42. King, B., Abney, S.: Labeling the languages of words in mixed-language documents using weakly supervised methods. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1110–1119 (2013)
43. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
44. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: Twenty-ninth AAAI conference on artificial intelligence (2015)
45. Lee, K., Yoon, K., Park, S., Hwang, S.w.: Semi-supervised training data generation for multilingual question answering. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018) (2018)
46. Lui, M., Baldwin, T.: Cross-domain feature selection for language identification. In: Proceedings of 5th international joint conference on natural language processing. pp. 553–561 (2011)
47. 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)
48. 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)
49. 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)
50. 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)
51. Mielke, S.: Language diversity in acl 2004 - 2016 (Dec 2016), <https://sjmielke.com/acl-language-diversity.htm>
52. 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)
53. Mohammad, S.M., Salameh, M., Kiritchenko, S.: How translation alters sentiment. *Journal of Artificial Intelligence Research* **55**, 95–130 (2016)
54. Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., Solorio, T.: Overview for the second shared task on language identification in code-switched data. In: Proceedings of the Second Workshop on Computational Approaches to Code Switching. pp. 40–49 (2016)
55. Munro, R.: Languages at acl this year (Jul 2015), <http://www.junglelightspeed.com/languages-at-acl-this-year/>
56. Nakov, P., Hoogeveen, D., Marquez, 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)

57. Nakov, P., Màrquez, L., Magdy, W., Moschitti, A., Glass, J., Randeree, B.: Semeval-2015 task 3: Answer selection in community question answering. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 269–281 (2015)
58. Nakov, P., Mrquez, L., Moschitti, A., Magdy, W., Mubarak, H., abed Alhakim Freihat, Glass, J., Randeree, B.: Semeval-2016 task 3: Community question answering. In: Bethard, S., Cer, D.M., Carpuat, M., Jurgens, D., Nakov, P., Zesch, T. (eds.) SemEval@NAACL-HLT. pp. 525–545. The Association for Computer Linguistics (2016), <http://dblp.uni-trier.de/db/conf/semEval/semEval2016.htmlNakovMMMMFGR16>
59. 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)
60. 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)
61. Rahman, M.M., Hisamoto, S., Duh, K.: Query expansion for cross-language question re-ranking. arXiv preprint arXiv:1904.07982 (2019)
62. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
63. 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)
64. Rücklé, A., Swarnkar, K., Gurevych, I.: Improved cross-lingual question retrieval for community question answering. In: The World Wide Web Conference. pp. 3179–3186. ACM (2019)
65. Ruder, S.: A survey of cross-lingual embedding models. arXiv preprint arXiv:1706.04902 (2017)
66. Sacaleanu, B., Neumann, G.: Cross-cutting aspects of cross-language question answering systems. In: Proceedings of the Workshop on Multilingual Question Answering-MLQA’06 (2006)
67. 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)
68. Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603 (2016)
69. Sequiera, R., Choudhury, M., Gupta, P., Rosso, P., Kumar, S., Banerjee, S., Naskar, S.K., Bandyopadhyay, S., Chittaranjan, G., Das, A., et al.: Overview of fire-2015 shared task on mixed script information retrieval. In: FIRE Workshops. vol. 1587, pp. 19–25 (2015)
70. Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., et al.: Overview for the first shared task on language identification in code-switched data. In: Proceedings of the First Workshop on Computational Approaches to Code Switching. pp. 62–72 (2014)
71. 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)

72. 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)
73. Tuggener, D.: Incremental coreference resolution for German. Ph.D. thesis, Universität Zürich (2016)
74. Ture, F., Boschee, E.: Learning to translate for multilingual question answering. arXiv preprint arXiv:1609.08210 (2016)
75. Upadhyay, S., Faruqui, M., Dyer, C., Roth, D.: Cross-lingual models of word embeddings: An empirical comparison. arXiv preprint arXiv:1604.00425 (2016)
76. 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)
77. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
78. Veysel, 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)
79. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. In: Advances in Neural Information Processing Systems. pp. 2692–2700 (2015)
80. Vulić, I., Moens, M.F.: Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. pp. 363–372. ACM (2015)
81. Wang, S., Jiang, J.: Machine comprehension using match-lstm and answer pointer. arXiv preprint arXiv:1608.07905 (2016)
82. Wang, Z., Mi, H., Hamza, W., Florian, R.: Multi-perspective context matching for machine comprehension. arXiv preprint arXiv:1612.04211 (2016)
83. Weissenborn, D., Wiese, G., Seiffe, L.: Making neural qa as simple as possible but not simpler. arXiv preprint arXiv:1703.04816 (2017)
84. Xiong, C., Zhong, V., Socher, R.: Dynamic coattention networks for question answering. arXiv preprint arXiv:1611.01604 (2016)
85. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1480–1489 (2016)
86. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. CoRR **abs/1708.02709** (2017), <http://arxiv.org/abs/1708.02709>
87. Zhang, Y., Riesa, J., Gillick, D., Bakalov, A., Baldrige, J., Weiss, D.: A fast, compact, accurate model for language identification of codemixed text. arXiv preprint arXiv:1810.04142 (2018)
88. Zhou, G., Xie, Z., He, T., Zhao, J., Hu, X.T.: Learning the multilingual translation representations for question retrieval in community question answering via non-negative matrix factorization. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) **24**(7), 1305–1314 (2016)