Datasets for Multilingual Answer Sentence Selection

Anonymous ACL submission

Abstract

Answer Sentence Selection (AS2) is a critical task for designing effective retrieval-based Question Answering (QA) systems. Most advancements in AS2 focus on English due to the scarcity of annotated datasets for other languages. This lack of resources prevents the training of effective AS2 models in different languages, creating a performance gap between QA systems in English and other locales. In this paper, we introduce new high-quality datasets for AS2 in five European languages (French, German, Italian, Portuguese, and Spanish), obtained through supervised Automatic Machine Translation (AMT) of existing English AS2 datasets such as ASNQ, WikiQA, and TREC-QA using a Large Language Model (LLM). We evaluated our approach and the quality of the translated datasets through multiple experiments with different Transformer architectures. The results indicate that our datasets are pivotal in producing robust and powerful multilingual AS2 models, significantly contributing to closing the performance gap between English and other languages.

1 Introduction

Answer Sentence Selection (AS2) represents a crucial component in many QA systems in both academic and industrial settings. The role of this component is to select the correct answer for a given question among a pool of candidate sentences. While in recent years significant progress has been made in developing models and datasets for AS2 (Wang et al., 2007; Yang et al., 2015; Garg et al., 2020; Di Liello et al., 2022; Gupta et al., 2023), most of these are designed and evaluated in English. By contrast, less attention has been paid to other medium resource languages, such as French, German, Italian, Portuguese, and Spanish, for which researchers struggle to obtain adequate amounts of quality data to train their models. Recently, Machine Translation (MT) has proven to be an effective approach to address the challenges of low-resource language QA systems (Kumar et al., 2021; Ranathunga et al., 2023; Gupta et al., 2023).

For the AS2 task, researchers have released a plethora of AS2 datasets in English, such as ASNQ (Garg et al., 2020), WikiQA (Yang et al., 2015), and TREC-QA (Wang et al., 2007), but there remains a gap for lower-resource languages that still needs to be filled.

In this work, we contribute to this research area by introducing three new large multilingual AS2 corpora[1]. named mASNQ, mWikiQA, and mTREC-QA for the most common European languages, comprising over 100 million question-answer pairs. We prepared these datasets by translating existing datasets (ASNQ, WikiQA, and TREC-QA) into five European languages (French, German, Italian, Portuguese, and Spanish) using a recent state-of-the-art translation model (Team et al., 2022). To validate the effectiveness of our approach, we trained several models using the mASNQ, mWikiQA, and mTREC-QA datasets and evaluated their performance. Our results demonstrate that these new datasets can be reliably used to train robust rankers for lower resource languages, yielding higher performance levels than those other competitors achieve. This contribution helps to reduce the language barrier and provides valuable assets for researchers working in low-resource languages.

2 Related Work

Multilingual Models: The development of multilingual models has seen significant progress due to the necessity of solving multilingual NLP tasks and cross-lingual applications. mBERT (Devlin et al., 2019), an extension of the original BERT model, can handle tasks across multiple languages. XLM-RoBERTa (Conneau et al., 2019), trained on

[1] Resources will be released upon paper acceptance.

100 languages, and mDeBERTa (He et al., 2021b), a variant of DebertaV3, have shown remarkable improvements in cross-lingual tasks. Similarly, mT5 (Xue et al., 2021), a multilingual variant of T5, and BLOOM (Scao et al., 2023), trained on the ROOTS corpus, exemplify advancements in multilingual models. Despite these efforts, multilingual models often underperform compared to their English versions due to the lower availability of training data (Gupta et al., 2023).

Translation Models: State-of-the-art Machine Translation (MT) models have demonstrated remarkable capabilities. OPUS-MT (Tiedemann and Thottingal, 2020), a set of translation tools, supports both bilingual and multilingual translations. The T5 model (Raffel et al., 2020), originally designed for various generative NLP tasks, is widely used for MT. The NLLB model (Team et al., 2022), trained on professionally translated datasets, supports translations between over 200 languages, facilitating broader support for low-resource languages.

Machine-Translated Datasets: MT has been widely used to address the lack of resources for multilingual AS2, showing promising results in the QA domain (Vu and Moschitti, 2021a; Kumar et al., 2021; Ranathunga et al., 2023). The itSQuAD dataset (Croce et al., 2018), the Spanish SQuAD (Carrino et al., 2019), and XQuAD (Dumitrescu et al., 2021) are examples of datasets translated via MT, used to build QA systems in different languages. The MLQA dataset (Lewis et al., 2019), mMARCO (Bonifacio et al., 2021), and Mintaka QA dataset (Sen et al., 2022) further highlight the success of machine-translated datasets in QA. Xtr-WikiQA and TyDi-AS2 (Gupta et al., 2023) are recent additions that extend AS2 datasets to multiple languages.

2.1 Answer Sentence Selection (AS2)

The AS2 task involves selecting the correct sentence from a pool of candidates to answer a given question. Early models like Severyn and Moschitti (2016) used separate embeddings for questions and answers, followed by convolutional layers. Garg et al. (2020) implemented Transformer-based models with an intermediate fine-tuning step, creating the ASNQ corpus from the Natural Questions dataset (Kwiatkowski et al., 2019). Contextual information has been shown to enhance AS2 models (Tan et al., 2018; Lauriola and Moschitti, 2021a; Campese et al., 2023). The translation of English AS2 data into target languages has been explored, demonstrating the potential for reducing the complexity of creating multilingual QA systems (Vu and Moschitti, 2021b). Recently, Cross-Lingual Knowledge Distillation (CLKD) (Gupta et al., 2023) has shown impressive results for low-resource languages, although the quality of machine translations remains a critical factor.

3 AS2 Translated Datasets

For dataset translation, we use the NLLB-200-3.3B model, which has 3.3 billion parameters, to translate three datasets: TREC-QA, WikiQA, and ASNQ. Both questions and answers are translated into French, German, Italian, Portuguese, and Spanish. The translation process involves two steps. First, we translate the datasets using the NLLB model. Second, we evaluate and improve the translation quality.

We assess the quality using a cross-language semantic similarity model by Reimers and Gurevych (2020)[2], which compares the semantic similarity between the original English sentences and their translations. Significant deviations indicate poor translations. We then apply heuristics to correct these errors, improving clarity, removing non-original text, and enhancing the overall quality of the translations.

3.1 Datasets

In this work, we considered and translated three datasets for answer sentence selection (AS2) in 5 different locales:

mTREC-QA, originates from TREC-QA (Wang et al., 2007), which is created from the TREC 8 to TREC 13 QA tracks. TREC 8-12 constitutes the training set, while TREC 13 questions are set aside for development and testing. We used the *Clean* setting, meaning that questions without an answer, or with only correct or incorrect answer-sentence candidates are removed.

mWikiQA is the translated version of WikiQA (Yang et al., 2015). It contains 3047 questions sampled from Bing query logs; candidate answer sentences are extracted from Wikipedia, and then manually labeled to assess whether it is a correct answer. Some sentences do not have a correct answer (*all* -), or have only correct answers (*all* +).

[2] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

Language	Split	Examples
ITA	Original Translated	*ISSN 0362 - 4331* ISSN 0362 - 4331 - Non lo so.
DEU	Original Translated	*Kusumanjali Prakashan .* Ich bin ein guter Mensch.
FRA	Original Translated	*Jump up to : " Safe Haven (2013)* Sauter sur refuge

Table 1: Examples of translation artifacts. The artifacts are highlighted in red. Notice that (i) "Non lo so" is an Italian sentence which translated in English means "I don't know", (ii) "Ich bin ein guter Mensch" means "I am a good person" in German, and (iii) the maening of the "Sauter sur refuge" in French is "Jump on refuge".

We trained using *no all* - mode and tested in the *clean* setting (without both *all* + and *all* -).

mASNQ comes from ASNQ (Garg et al., 2020) which is an AS2 dataset created by adapting the Natural Question (Kwiatkowski et al., 2019) corpus from Machine Reading (MR) to the AS2 task. We replicated this passage using the scripts provided by Lauriola and Moschitti (2021b).

We summarize the statistics of these datasets in Appendix H.

3.2 Removing Translation Artifacts

Despite the good quality of the translation, the dataset still presents some inconsistencies and artifacts. We identified four major classes of translation artifacts: (i) Meaning mismatch between the original and the translated sentences, (ii) The addition of not necessary suffixes and prefixes, (iii) The difficulty in interpreting and translating numerical strings, (iv) Out-of-topic translations of partial contexts. We provide some examples of these translation artifacts in Table 1.

To tackle these issues, we apply some heuristics to improve the dataset quality by designing a simple human-centered pipeline to mitigate these artifacts.

In our approach, we first compute the similarity score between every translated example in the dataset and the corresponding original text. Then we filter out translated examples below a similarity threshold of 0.8 and, on the remaining set, we compute the most common $1k$ n-grams with n ranging from 4 to 9. Second, we manually inspect these extracted n-grams, identifying and removing artifact patterns that could distort the data. Subsequently, we systematically remove occurrences of those problematic artifacts from the translated dataset. To further improve this operation, we also identify the examples where the original sentence and the 75% of the not-blank characters are numbers, and if the similarity score is low (under 0.8) we replace the translated sentences with the original one.

3.3 Semantic Similarities

To assess the quality of the translations and to quantify the benefit given by our heuristics, we evaluate the semantic similarity between the original sentences and their translated versions. For each question-answer pair of each dataset, we compare the original sentences in English with their translated version in the target language. The overall similarity measure between the originals and the translated sentences in each dataset is computed considering the mean of the semantic similarity scores across all the question-answer pairs. This average score indicates how closely the translated sentences align with their original counterparts in terms of semantic meaning. In Appendix H we report the comparison between the similarity scores of the translated dataset and the original one.

4 Experiments

In this section, we aim to verify and prove the effectiveness of our contributions. Specifically, we show that the translated data could be used to train state-of-the-art AS2 models in multiple languages. For each considered language, we finetune existing multilingual transformer models on both the original and our translated datasets.

To verify these hypotheses, we consider an existing multilanguage pre-trained cross-encoder transformer model, which is XLM-RoBERTa base[3], and BERT-multilingual[4]. Following the TANDA (Garg et al., 2020) approach, we perform a two-stage training for each model. Precisely, this technique consists of a two-stage training paradigm, where the first training stage, named *transfer step*, involves training the models on ASNQ to teach them to recognize and solve the AS2 tasks. In the second step, named *adaptation step*, the transferred models are fine-tuned on the final target AS2 datasets. In our setting, we apply this paradigm by first training and doing a separate transfer step on each language of mASNQ. Secondly, we finetune the obtained models on mWikiQA and mTREC-QA, comparing the differences with the baselines (XLM-RoBERTa on the original English splits of WikiQA and TREC-QA) in terms of P@1 and MAP

[3] https://huggingface.co/xlm-roberta-base

[4] https://huggingface.co/bert-base-multilingual-cased

Language	Transfer	Adapt	mWikiQA		mTrecQA		Xtr-WikiQA	
			MAP	P@1	MAP	P@1	MAP	P@1
ENG		✓	0.796 (± 0.011)	0.691 (± 0.015)	0.866 (± 0.015)	0.853 (± 0.031)	0.796 (± 0.011)	0.691 (± 0.015)
	✓	✓	0.874 (± 0.007)	0.813 (± 0.017)	0.892 (± 0.007)	0.871 (± 0.019)	0.874 (± 0.007)	0.813 (± 0.017)
DEU		✓	0.770 (± 0.024)	0.657 (± 0.031)	0.870 (± 0.011)	0.871 (± 0.016)	0.779 (± 0.012)	0.669 (± 0.015)
	✓	✓	0.853 (± 0.005)	0.767 (± 0.006)	0.904 (± 0.006)	0.903 (± 0.017)	0.868 (± 0.005)	0.800 (± 0.011)
FRA		✓	0.769 (± 0.012)	0.662 (± 0.018)	0.872 (± 0.007)	0.859 (± 0.013)	0.760 (± 0.009)	0.646 (± 0.012)
	✓	✓	0.836 (± 0.006)	0.752 (± 0.012)	0.891 (± 0.010)	0.874 (± 0.029)	0.844 (± 0.005)	0.778 (± 0.014)
ITA		✓	0.768 (± 0.019)	0.660 (± 0.026)	0.855 (± 0.024)	0.844 (± 0.049)	0.761 (± 0.021)	0.657 (± 0.027)
	✓	✓	0.828 (± 0.004)	0.749 (± 0.008)	0.870 (± 0.006)	0.871 (± 0.016)	0.820 (± 0.012)	0.742 (± 0.027)
POR		✓	0.798 (± 0.011)	0.704 (± 0.019)	0.855 (± 0.021)	0.704 (± 0.019)	0.780 (± 0.011)	0.684 (± 0.020)
	✓	✓	0.853 (± 0.018)	0.781 (± 0.029)	0.874 (± 0.009)	0.781 (± 0.029)	0.849 (± 0.023)	0.775 (± 0.042)
SPA		✓	0.795 (± 0.009)	0.691 (± 0.016)	0.882 (± 0.013)	0.900 (± 0.016)	0.786 (± 0.015)	0.691 (± 0.024)
	✓	✓	0.847 (± 0.013)	0.768 (± 0.020)	0.898 (± 0.004)	0.929 (± 0.019)	0.859 (± 0.010)	0.790 (± 0.020)

Table 2: Performance comparison of XLM-RoBERTa on mTREC-QA, mWikiQA, and Xtr-WikiQA (zero-shot from the model trained on mWikiQA). The transfer step is done on mASNQ, while the Adaptation is on mTREC-QA and mWikiQA. Results in terms of MAP and P@1, for various language and model configurations. The experiments on the English split represent the models trained and tested on the original, not translated versions of ASNQ, WikiQA and TREC-QA.

on the test sets of mWikiQA, mTREC-QA and Xtr-WikiQA[5]. It is important to note that for the test on Xtr-WikiQA, we considered the model fine-tuned on mWikiQA. Indeed, this allows us to showcase the results on existing test sets and demonstrate that even though both mASNQ and mWikiQA were translated using the same methodology, this does not introduce bias when evaluated on non-automatically translated datasets.

For space reasons, we propose additional experiments using different multilingual models in Appendix C and the training details in Appendix A.

4.1 Results

In this section, we present the experimental results of our approaches on three different AS2 datasets: mWikiQA, mTREC-QA, and the existing Xtr-WikiQA dataset (Gupta et al., 2023). Table 2 provides an overview of the performance achieved by our models. First, we observe that our models achieve performance levels comparable to those of English models. This finding is particularly noticeable when considering the Portuguese language across all datasets. When evaluating the models on Xtr-WikiQA, which can be considered as a zero-shot scenario, as the models are trained on mWikiQA and tested on Xtr-WikiQA, we find that our approaches demonstrate robustness even when dealing with datasets translated using a different translation pipeline. Specifically, Xtr-WikiQA is translated using Amazon Translate. The results obtained on Xtr-WikiQA validate the effectiveness of our procedures in handling such translation variations.

Additional results and comparisons of various models on Xtr-WikiQA in a zero-shot setting can be found in Appendix D. These results showcase the effectiveness and robustness of our approaches on AS2 datasets, highlighting competitive performance and the superiority of certain training configurations and models over others.

5 Conclusion

Our study tackles the language barrier in QA systems by focusing on European languages such as Italian, German, Portuguese, Spanish, and French. We introduced new large multilingual AS2 datasets (mASNQ, mWikiQA, and mTREC-QA) by translating existing English AS2 datasets using a state-of-the-art translation model. This approach provides valuable resources for lower-resource languages. Our extensive experiments demonstrated the effectiveness of these datasets in training robust AS2 rankers across various languages, achieving performance comparable to English datasets. This contributes significantly to reducing the language barrier, making AS2 more accessible and effective across different linguistic contexts. To support further research, we will release the new models and multilingual AS2 datasets to the research community. We hope our work inspires future studies to address language diversity challenges in QA, leading to more inclusive and effective solutions for global users.

[5] https://huggingface.co/datasets/AmazonScience/xtr-wiki_qa

Limitations

This paper focuses on five European languages (Italian, German, Portuguese, Spanish, and French). This could represent a limitation since we limit the applicability of the findings to other languages. Another possible limitation is that the accuracy and quality of machine translation can affect the performance of trained models by introducing errors and inconsistencies, compromising dataset reliability. Moreover, biases present in the original English data might be transferred to the translated datasets, potentially resulting in skewed or unrepresentative training examples for specific languages. Finally, we reserve for future analysis on larger and more powerful pre-trained multilingual language models (e.g., XLM-RoBERTa large, and mDeBERTa).

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Luiz Henrique Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, , Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of ms marco passage ranking dataset. *Preprint*, arXiv:2108.13897.

Stefano Campese, Ivano Lauriola, and Alessandro Moschitti. 2023. QUADRo: Dataset and models for QUestion-answer database retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15573–15587, Singapore. Association for Computational Linguistics.

Casimiro Pio Carrino, Marta R Costa-jussà, and José AR Fonollosa. 2019. Automatic spanish translation of the squad dataset for multilingual question answering. *arXiv preprint arXiv:1912.05200*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020a. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020b. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In *AI*IA 2018 – Advances in Artificial Intelligence*, pages 389–402, Cham. Springer International Publishing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Luca Di Liello, Siddhant Garg, Luca Soldaini, and Alessandro Moschitti. 2022. Paragraph-based transformer pre-training for multi-sentence inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2521–2531, Seattle, United States. Association for Computational Linguistics.

Stefan Daniel Dumitrescu, Petru Rebeja, Beata Lorincz, Mihaela Gaman, Andrei Avram, Mihai Ilie, Andrei Pruteanu, Adriana Stan, Lorena Rosia, Cristina Iacobescu, Luciana Morogan, George Dima, Gabriel Marchidan, Traian Rebedea, Madalina Chitez, Dani Yogatama, Sebastian Ruder, Radu Tudor Ionescu, Razvan Pascanu, and Viorica Patraucean. 2021. Liro: Benchmark and leaderboard for romanian language tasks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig, editors. 2022. *Ethnologue: Languages of the World*, twenty-fifth edition. SIL International, Dallas, Texas.

Matteo Gabburo, Siddhant Garg, Rik Koncel-Kedziorski, and Alessandro Moschitti. 2023. Learning answer generation using supervision from automatic question answering evaluators. *Preprint*, arXiv:2305.15344.

Matteo Gabburo, Rik Koncel-Kedziorski, Siddhant Garg, Luca Soldaini, and Alessandro Moschitti. 2022. Knowledge transfer from answer ranking to answer generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9481–9495, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Siddhant Garg, Thuy Vu, and Alessandro Moschitti. 2020. Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7780–7788.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Shivanshu Gupta, Yoshitomo Matsubara, Ankit Chadha, and Alessandro Moschitti. 2023. Cross-lingual knowledge distillation for answer sentence selection in low-resource languages. *Preprint*, arXiv:2305.16302.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. *arXiv preprint arXiv:2106.06797*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Preprint*, arXiv:1901.07291.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2023. The bigscience roots corpus: A 1.6tb composite multilingual dataset. *Preprint*, arXiv:2303.03915.

Ivano Lauriola and Alessandro Moschitti. 2021a. Answer sentence selection using local and global context in transformer models. In *ECIR 2021*.

Ivano Lauriola and Alessandro Moschitti. 2021b. Answer sentence selection using local and global context in transformer models. ECIR.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. *Preprint*, arXiv:2108.12409.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *Preprint*, arXiv:1806.03822.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *COLING 2022*.

Aliaksei Severyn and Alessandro Moschitti. 2016. Modeling relational information in question-answer pairs with convolutional neural networks. *Preprint*, arXiv:1604.01178.

Chuanqi Tan, Furu Wei, Qingyu Zhou, Nan Yang, Bowen Du, Weifeng Lv, and Ming Zhou. 2018. Context-aware answer sentence selection with hierarchical gated recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3):540–549.

Harish Tayyar Madabushi, Mark Lee, and John Barnden. 2018. Integrating question classification and deep learning for improved answer selection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3283–3294, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Thuy Vu and Alessandro Moschitti. 2021a. Multilingual answer sentence reranking via automatically translated data. *Preprint*, arXiv:2102.10250.

Thuy Vu and Alessandro Moschitti. 2021b. Multilingual answer sentence reranking via automatically translated data. *CoRR*, abs/2102.10250.

Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. 2007. What is the Jeopardy model? a quasi-synchronous grammar for QA. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 22–32, Prague, Czech Republic. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *Preprint*, arXiv:1911.00359.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Zeyu Zhang, Thuy Vu, Sunil Gandhi, Ankit Chadha, and Alessandro Moschitti. 2022. Wdrass: A web-scale dataset for document retrieval and answer sentence selection. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 4707–4711, New York, NY, USA. Association for Computing Machinery.

A Training details

To perform our experiments, we train XLM-RoBERTa on ASNQ and mASNQ datasets with specific parameters: batch size of 1024, Adam optimizer with a learning rate of $5e-6$, precision set to 32, and 10 training epochs. For mWikiQA and mTREC-QA datasets, the batch size was 32, Adam optimizer with a learning rate of $5e-6$, precision set to 16-mixed, and 40 training epochs with early stopping. We select the best model maximizing the mean average precision (MAP) on the development set. The same parameters were used for training the Multilingual BERT architecture. All experiments utilized 8 NVIDIA V100 32 GB GPUs.

B ASNQ additional results

Table 3 presents the performance of the XLM-RoBERTa model trained on the development set of the multilingual ASNQ dataset. The performance of XLM-RoBERTa on the original ASNQ development set is also reported. The results indicate that the English baseline outperforms the models trained in other languages, as expected, while the performance of the models trained in different languages is consistent and relatively close. These results highlight that the use of our translated datasets can improve the performance in terms of MAP, P@1, MRR, and NDCG metrics across multiple languages.

Language	With Translated Data				Without Translated Data			
	MAP	P@1	MRR	NDCG	MAP	P@1	MRR	NDCG
ENG	0.870 (± 0.007)	0.812 (± 0.017)	0.745 (± 0.011)	0.831 (± 0.015)	0.870 (± 0.007)	0.812 (± 0.017)	0.745 (± 0.011)	0.831 (± 0.015)
DEU	0.850 (± 0.005)	0.768 (± 0.006)	0.710 (± 0.008)	0.811 (± 0.009)	0.845 (± 0.005)	0.762 (± 0.010)	0.705 (± 0.010)	0.806 (± 0.012)
FRA	0.835 (± 0.006)	0.750 (± 0.012)	0.692 (± 0.009)	0.798 (± 0.011)	0.830 (± 0.010)	0.743 (± 0.017)	0.689 (± 0.014)	0.793 (± 0.017)
ITA	0.842 (± 0.004)	0.755 (± 0.008)	0.699 (± 0.009)	0.804 (± 0.010)	0.835 (± 0.003)	0.748 (± 0.011)	0.692 (± 0.011)	0.798 (± 0.012)
POR	0.853 (± 0.018)	0.781 (± 0.029)	0.715 (± 0.021)	0.822 (± 0.023)	0.848 (± 0.011)	0.777 (± 0.020)	0.712 (± 0.015)	0.818 (± 0.017)
SPA	0.847 (± 0.013)	0.768 (± 0.020)	0.705 (± 0.016)	0.812 (± 0.018)	0.840 (± 0.012)	0.760 (± 0.024)	0.700 (± 0.018)	0.805 (± 0.019)

Table 3: Performance comparison of XLM-RoBERTa on the multilingual ASNQ dataset with and without translated data. Results are reported in terms of MAP, P@1, MRR, and NDCG metrics.

C Results using better multilingual models

In this section, we present the results of our experiments using the newly created multilingual Answer Sentence Selection (AS2) datasets. The goal is to evaluate the performance of our approach using mDeBERTa across different languages and settings. We consider three main tables that provide a comprehensive overview of the results.

Table 4 presents the performance of mDeBERTa on the mASNQ dataset, covering multiple languages (DEU, FRA, ITA, SPA, POR). The metrics reported include Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG), and Precision at 1 (P@1). These results highlight the effectiveness of our multilingual datasets, showcasing the robustness and consistency of the model across different languages.

Approach	MAP	MRR	NDCG	P1
ASNQ	0.677	0.743	0.707	0.638
mASNQ$_{DEU}$	0.633	0.702	0.662	0.590
mASNQ$_{FRA}$	0.629	0.700	0.657	0.591
mASNQ$_{ITA}$	0.639	0.708	0.670	0.597
mASNQ$_{SPA}$	0.632	0.703	0.663	0.589
mASNQ$_{POR}$	0.635	0.706	0.664	0.595

Table 4: Results of mDeberta on mASNQ

Table 5 compares the performance of mDeBERTa-v3-base when transferred on mASNQ and ASNQ, and subsequently tested on mWikiQA. The results are presented in terms of MAP and P@1 for each language considered (ITA, DEU, SPA, POR, FRA). This table demonstrates the improvements achieved by utilizing the mASNQ dataset, with notable gains in performance across all languages.

Table 6 presents a detailed performance comparison of mDeBERTa on three datasets: mWikiQA, mTREC-QA, and Xtr-WikiQA (zero-shot from the model trained on mWikiQA). The results are reported in terms of MAP and P@1 for various language and model configurations. This table illustrates the benefits of the transfer step on mASNQ and the adaptation step on mTREC-QA and mWikiQA, with the mDeBERTa models consistently achieving high performance across all tasks and languages.

	ASNQ		mASNQ	
	MAP	P@1	MAP	P@1
ITA	0.868	0.801	0.884	0.821
DEU	0.882	0.827	0.885	0.821
SPA	0.875	0.811	0.886	0.829
POR	0.882	0.825	0.883	0.830
FRA	0.851	0.766	0.884	0.819

Table 5: Results of mDeBERTa-v3-base transferred on mASNQ and ASNQ and tested on mWikiQA.

The results in these tables provide comprehensive insights into the effectiveness of our multilingual datasets and the benefits of the proposed transfer and adaptation steps. These findings underline the importance of high-quality multilingual datasets in improving the performance of AS2 models across diverse languages, demonstrating the robustness and generalizability of our approach.

D Results in a zero-shot setting

In this section, we present a comparison of various models on Xtr-WikiQA in a zero-shot setting in Table 7. These models have been trained on mASNQ and are evaluated against existing models trained on well-known and extensive passage reranking datasets. We find that models trained on mASNQ for the Xtr-WikiQA task outperform models trained on other datasets such as mMARCO and MSMARCO. This observation suggests that mASNQ is a more suitable dataset for AS2 compared to mMARCO and MSMARCO. Moreover, when comparing the performance of BERT-multilingual and XLM-RoBERTa, we find that, on average, BERT-

Language	Transfer	Adapt	mWikiQA		mTREC-QA		Xtr-WikiQA	
			MAP	P@1	MAP	P@1	MAP	P@1
ENG		✓	0.872 (± 0.008)	0.801 (± 0.013)	0.905 (± 0.004)	0.909 (± 0.022)	0.872 (± 0.008)	0.801 (± 0.013)
	✓	✓	0.901 (± 0.006)	0.854 (± 0.008)	0.921 (± 0.003)	0.950 (± 0.008)	0.901 (± 0.006)	0.854 (± 0.008)
DEU		✓	0.847 (± 0.006)	0.765 (± 0.006)	0.898 (± 0.011)	0.912 (± 0.023)	0.847 (± 0.006)	0.765 (± 0.006)
	✓	✓	0.888 (± 0.007)	0.837 (± 0.013)	0.925 (± 0.008)	0.944 (± 0.016)	0.888 (± 0.007)	0.837 (± 0.013)
FRA		✓	0.848 (± 0.010)	0.778 (± 0.018)	0.900 (± 0.008)	0.924 (± 0.016)	0.848 (± 0.010)	0.778 (± 0.018)
	✓	✓	0.894 (± 0.003)	0.848 (± 0.003)	0.918 (± 0.004)	0.932 (± 0.008)	0.894 (± 0.003)	0.848 (± 0.003)
ITA		✓	0.851 (± 0.010)	0.774 (± 0.011)	0.890 (± 0.009)	0.900 (± 0.016)	0.851 (± 0.010)	0.774 (± 0.011)
	✓	✓	0.885 (± 0.006)	0.822 (± 0.010)	0.919 (± 0.005)	0.938 (± 0.012)	0.885 (± 0.006)	0.822 (± 0.010)
POR		✓	0.846 (± 0.015)	0.765 (± 0.023)	0.896 (± 0.013)	0.900 (± 0.019)	0.846 (± 0.015)	0.765 (± 0.023)
	✓	✓	0.889 (± 0.005)	0.842 (± 0.007)	0.925 (± 0.004)	0.953 (± 0.012)	0.889 (± 0.005)	0.842 (± 0.007
SPA		✓	0.857 (± 0.010)	0.781 (± 0.016)	0.902 (± 0.012)	0.921 (± 0.022)	0.857 (± 0.010)	0.781 (± 0.016)
	✓	✓	0.879 (± 0.007)	0.819 (± 0.011)	0.915 (± 0.007)	0.924 (± 0.019)	0.879 (± 0.007)	0.819 (± 0.011)

Table 6: Performance comparison of mDeBERTa on mWikiQA, mTREC-QA, and Xtr-WikiQA (zero-shot from the model trained on mWikiQA). The transfer step is done on mASNQ, while the adaptation is on mTREC-QA and mWikiQA. Results in terms of MAP and P@1, for various language and model configurations.

multilingual performs better. This finding is evident when analyzing the results across different languages, including Italian, Portuguese, and Spanish. Overall, our results demonstrate the effectiveness and robustness of our approaches on AS2 datasets, showcasing competitive performance and the superiority of certain training configurations and models over others.

E Ablation: Cross-Lingual

This ablation aims to determine the advantages of using the mASNQ dataset to train state-of-the-art answer ranking models on languages different from English. To achieve this, we compare the performance of cross-lingual models trained on ASNQ and WikiQA with models that were first trained on mASNQ and mWikiQA, across the different languages that compose mWikiQA.

Table 8 compares the performance of models trained only on the original versions of ASNQ and WikiQA with the performance of the same architecture (XLM-RoBERTa base) but trained on our multilingual datasets. To achieve this goal, we measure the performance of each model across all the different test sets of mWikiQA and across their languages. For the evaluation, we considered two proxy measures to understand the quality of the models: Mean Average Precision (MAP) and Precision at 1 (P@1). The results show that the models achieve higher MAP and P@1 scores when trained on mASNQ compared to ASNQ, indicating that training on the mASNQ dataset improves the performance of multilingual models in cross-lingual tasks. Across all languages, the models trained on mASNQ consistently outperform the models trained on ASNQ. This suggests that the mASNQ dataset can guarantee a performance boost for non-English target datasets, confirming our hypotheses.

F Ablation: Ranks Correlation

This study compares the ranking outputs of two sets of models, analyzing the correlation between their rankings. The first set comprehends models trained on mASNQ and mWikiQA and then tested on the mWikiQA test set, while the second set contains models trained on ASNQ and WikiQA and evaluated on the original English WikiQA test set.

We design this experiment in order to compare the rank provided for each question q^i_{Eng} of the original English dataset (WikiQA), with the semantically equivalent question q^i_T and its rank for each language T in mWikiQA. To measure the performance, we compute three correlation metrics to properly evaluate the correlation between the rankings of each pair of questions $\{q^i_{Eng}, q^i_T\}$; in this way, we allow determining the level of agreement between the two models' ranking outputs, providing insights into the potential differences between them. Specifically, we consider XLM-RoBERTa base and compute the Kendall, Spearman, and Pearson correlation metrics on mWikiQA and mTREC-QA.

The results in Table 9 show a strong positive correlation between the performance of models trained in English and tested in English, and the models trained in other languages (using mASNQ, mWikiQA, and mTREC-QA). This correlation is evident

Language	Model	MAP	P@1
ENG	✱ mmarco-mMiniLMv2	0.812	0.722
	✱ bert-multilingual-msmarco	0.798	0.714
	xlm-roberta-base	0.855	0.794
	bert-multilingual	0.814	0.724
DEU	✱ mmarco-mMiniLMv2	0.797	0.700
	✱ bert-multilingual-msmarco	0.759	0.663
	xlm-roberta-base	0.844	0.770
	bert-multilingual	0.834	0.753
FRA	✱ mmarco-mMiniLMv2	0.782	0.675
	✱ bert-multilingual-msmarco	0.734	0.621
	xlm-roberta-base	0.813	0.720
	bert-multilingual	0.863	0.807
ITA	✱ mmarco-mMiniLMv2	0.778	0.671
	✱ bert-multilingual-msmarco	0.735	0.634
	xlm-roberta-base	0.830	0.741
	bert-multilingual	0.846	0.765
POR	✱ mmarco-mMiniLMv2	0.809	0.724
	✱ bert-multilingual-msmarco	0.755	0.646
	xlm-roberta-base	0.840	0.761
	bert-multilingual	0.841	0.761
SPA	✱ mmarco-mMiniLMv2	0.791	0.691
	✱ bert-multilingual-msmarco	0.753	0.650
	xlm-roberta-base	0.832	0.737
	bert-multilingual	0.853	0.774

Table 7: Performance comparison of XLM-RoBERTa base model in a zero-shot setting on the Xtr-WikiQA task. Models trained on mASNQ dataset, denoted by ✱, outperform those trained on other datasets like mMARCO and MSMARCO. Moreover, BERT-multilingual consistently performs better than XLM-RoBERTa in various languages (Italian, Portuguese, Spanish), indicating the robustness and competitiveness of the approach on AS2 datasets.

Language	ASNQ		mASNQ	
	MAP	P@1	MAP	P@1
DEU	0.814	0.705	0.839	0.755
FRA	0.819	0.717	0.793	0.671
ITA	0.819	0.715	0.839	0.726
POR	0.822	0.722	0.842	0.755
SPA	0.830	0.738	0.835	0.751

Table 8: Comparison of XLM-RoBERTa base transferred on mASNQ and ASNQ and tested on mWikiQA in a cross-lingual setting.

mASNQ→mWikiQA

Language	Kendall	Spearman	Pearson
DEU	0.720	0.820	0.908
FRA	0.694	0.802	0.872
ITA	0.713	0.817	0.903
POR	0.710	0.824	0.908
SPA	0.698	0.807	0.902

mASNQ→mTREC-QA

Language	Kendall	Spearman	Pearson
DEU	0.547	0.666	0.713
FRA	0.566	0.663	0.709
ITA	0.513	0.629	0.695
POR	0.567	0.669	0.733
SPA	0.587	0.688	0.728

Table 9: Kendall, Spearman and Pearson correlation computed between the ranks originated from model trained the original ASNQ and mASNQ. The reported values are computed using XLM-RoBERTa base models transferred on ASNQ and mASNQ and then finetuned on mWikiQA and mTREC-QA.

across all evaluation metrics, with Kendall correlations ranging from 0.694 to 0.720, Spearman correlations ranging from 0.802 to 0.824, and Pearson correlations ranging from 0.872 to 0.908 for the mASNQ→mWikiQA task. The high correlation values, ranging from 0.547 to 0.733, across all languages for the mASNQ→mTREC-QA task further support this notion. The Kendall, Spearman, and Pearson correlations show consistently high values, indicating that the translation quality and model performance are consistently strong. The results of the analysis demonstrate (i) the effectiveness of the translation process for mASNQ and (ii) the strong performance of the models.

G Ablation: Passage Ranking

To further evaluate the robustness of our datasets, we also perform several experiments on a different task: Passage Reranking (PR). Passage Reranking is an Information Retrieval (IR) task that consists of reordering a set of retrieved passages for a given query. For this reason, we consider a well-known dataset named mMARCO (Bonifacio et al., 2021), well known in the multilingual IR community. Specifically, we select a random language among the ones considered in the previous experiments, and we train several multi-language models. In detail, we split the original Italian dataset into train, validation, and test splits (Tab. 11).

We compare the results obtained by our approaches with two models: the first is a multilingual BERT trained on the English MSMARCO, while the second model is trained on our train split. In Table 10, we present the results of this comparison. They clearly show that our models trained on the mMARCO dataset outperform the model trained on MSMARCO (e.g., 0.687 vs 0.682 in terms of MAP).

Although the improvement is modest, it becomes significant due to the large size of the mMARCO test set. These findings highlight the advantages our datasets offer for tasks beyond AS2. Even with a marginal improvement, it is evident that adapting a model trained on our multilingual datasets can yield further performance enhancements.

Dataset	Transfer	Adapt	mMARCO	
			MAP	P@1
MSMARCO	✓		0.631	0.502
mMARCO		✓	0.682	0.553
mASNQ→mMARCO	✓	✓	0.687	0.559

Table 10: Comparison of BERT-multilingual performance on mMARCO$_{\text{ITA}}$ test set. We train the two baselines respectively on the English MSMARCO and the mMARCO Italian split. The models trained on mASNQ and adapted to mMARCO consistently improve the two presented baselines, showing that the transfer step on mASNQ is helpful in this domain.

H Datasets

In Table 11, we provide the datasets we described in Section 3.

In addition, in Table 12, we report the semantic similarity between ASNQ and mASNQ to support the translation quality further.

Dataset	Split	#Question	#QA Pairs
mASNQ	Train	57240	20377168
	Validation	2672	930062
mWikiQA	Train	2118	20356
	Validation	296	2731
	Validation (++)	126	1130
	Validation (clean)	122	1126
	Test	633	6160
	Test (++)	243	2350
	Test (clean)	237	2340
mTREC-QA	Train	1227	53282
	Validation	65	1117
	Test	68	1441

Table 11: Dataset statistics for mASNQ, mWikiQA, and mTREC-QA for each language. The datasets have the same statistics in their original version, and considering all the languages, the corpora comprehend more than 100M examples. Notice that for mWikiQA we report also the statistics of the clean and the no-all-negatives (++) splits.

	Language	Similarity
Dev	DEU	$0.869 \pm 0.178 \rightarrow 0.991 \pm 0.003$
	FRA	$0.838 \pm 0.223 \rightarrow 0.990 \pm 0.003$
	ITA	$0.913 \pm 0.115 \rightarrow 0.990 \pm 0.001$
	POR	$0.915 \pm 0.117 \rightarrow 0.992 \pm 0.003$
	SPA	$0.857 \pm 0.211 \rightarrow 0.990 \pm 0.001$
Train	DEU	$0.871 \pm 0.175 \rightarrow 0.923 \pm 0.110$
	FRA	$0.841 \pm 0.218 \rightarrow 0.923 \pm 0.101$
	ITA	$0.915 \pm 0.112 \rightarrow 0.940 \pm 0.081$
	POR	$0.915 \pm 0.115 \rightarrow 0.941 \pm 0.082$
	SPA	$0.860 \pm 0.206 \rightarrow 0.932 \pm 0.090$

Table 12: Similarities between ASNQ and mASNQ. On the left of the arrow (→) the similarity reached after the initial translation is reported; on the right side, there is the similarity score after the application of the heuristics.