





xAMR: Cross-lingual AMR End-to-End Pipeline

Maja Mitreska¹^a, Tashko Pavlov¹^b, Kostadin Mishev^{1,2}^c and Monika Simjanoska^{1,2}^d

¹*iReason LLC, 3rd Macedonian Brigade 37, Skopje, North Macedonia*

²*Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Rugjer Boshkovic 16, Skopje, North Macedonia*

Keywords: Cross-lingual AMR, AMR Parsing, AMR-to-Text Generation, Multilingual AMR, Cosine Similarity, BLEU, ROUGE, LASER, LaBSE, Distiluse, Low-resource Languages, Europarl.


Abstract: Creating multilingual end-to-end AMR models requires a large amount of cross-lingual data making the parsing and generating tasks exceptionally challenging when dealing with low-resource languages. To avoid this obstacle, this paper presents a cross-lingual AMR (xAMR) pipeline that incorporates the intuitive translation approach to and from the English language as a baseline for further utilization of the AMR parsing and generation models. The proposed pipeline has been evaluated via the cosine similarity of multiple state-of-the-art sentence embeddings used for representing the original and the output sentences generated by our xAMR approach. Also, BLEU and ROUGE scores were used to evaluate the preserved syntax and the word order. xAMR results were compared to multilingual AMR models' performance for the languages experimented within this research. The results showed that our xAMR outperforms the multilingual approach for all the languages discussed in the paper and can be used as an alternative approach for abstract meaning representation of low-resource languages.


1 INTRODUCTION


Abstract Meaning Representation (AMR) (Banarescu et al., 2013) is a language for semantic representation firstly introduced for the English language. The purpose of AMR is presenting a sentence into an AMR graph where nodes portray the entities in the sentence and the edges between them represent the existing semantic relationships. The AMR graphs are rooted, directed, acyclic graphs with labeled edges and leaves. The process of converting sentence to an AMR graph is called AMR parsing as we parse each component of the sentence into its appropriate graph representation. Moreover, the opposite process, the conversion of an AMR graph to sentence is known as AMR-to-text generation. Each of these processes has their own specific characteristics and challenges which encourage researchers to explore the best solutions. Due to the fact that AMR is not intended to be interlingua (Banarescu et al., 2013), there is a huge challenge


when it comes to re-purposing AMR for other languages, i.e., making AMR cross-lingual stable.

There are a lot of different methods of AMR parsing and generating in the English language, and most of them are focusing on either on the parsing, or on the generating part. Each of the research is proposing new and enhanced way of dealing with these tasks. In (Wang et al., 2015) a transition-based framework is presented, where a sentence is first transformed in a dependency tree which is used as input for building an AMR graph. Furthermore, in (Ballesteros and Al-Onaizan, 2017) and (Wang and Xue, 2017) the usage of stacked bidirectional LSTM networks are explored, and even more probabilistic method is introduced in (Lyu and Titov, 2018) where the alignments are treated as latent variables in a joint probabilistic model. In (Zhang et al., 2019) the task of AMR parsing is treated as a sequence-to-graph transduction problem. Many of the papers that are exploring the AMR-to-text generating task build and train graph-to-sequence models, i.e., graph transformer architectures (Song et al., 2018), (Damonte and Cohen, 2019), (Zhu et al., 2019), (Wang et al., 2020). In the recent years, many researchers are succeeding to combine

^a <https://orcid.org/0000-0002-9105-7728>

^b <https://orcid.org/0000-0001-7689-4475>

^c <https://orcid.org/0000-0003-3982-3330>

^d <https://orcid.org/0000-0002-5028-3841>

these tasks, and build end-to-end systems for parsing sentences to AMR graphs and generating sentences from given AMR graphs using advanced neural networks architectures (Konstas et al., 2017), (Blloshmi et al., 2021).

With the progress and improvement of the English-based AMR parsers and generators, the next major step is building multilingual AMR parsers and/or generators. Another important challenge in creating such systems is the small amount of cross-lingual data. The English AMR parsers and generators need training on huge amount of English sentences, but more importantly those systems need annotated AMR graphs so they can train or evaluate their models. Most of the time those AMR graphs are manually annotated. Because of these reasons, the creation of multilingual systems is even more problematic. The need for creating dataset of multilingual AMR graphs requires a lot of both data and human resources.

Encouraged by the mentioned difficulties, we propose new approach for creating end-to-end cross-lingual AMR systems. In this paper, we explain the benefits of the usage of state-of-the-art translators in these kind of problems. Our architecture consists of pretrained translator and state-of-the-art English-based AMR model. Firstly, a given non-English sentence is translated into English and further processed into the AMR parser out of which the AMR graph is obtained. This AMR graph is then run through an AMR generator which results with an English sentence as an output. In the final step, the generated English sentence is translated to the original source language. To measure the preserved semantics of the input and the output sentences, our xAMR pipeline is evaluated by using cosine similarity (Singhal et al., 2001) in the opposite of the foregoing papers where usually the Smatch metric (Cai and Knight, 2013) is used to evaluate the semantic similarity of two AMR graphs. Additionally, the quality of the translated sentences in terms of overlapping words is measured with BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). As a benchmark dataset we are using the Europarl Corpus (Koehn, 2005) which is a parallel corpus deliberately constructed for statistical machine translation and consists of the languages that are spoken in the European Parliament. We are limiting these experiments on German, Italian, Spanish and Bulgarian language because there are available multilingual AMR representations. In addition to these languages, we have manually translated a subset of the English Europarl corpus into Macedonian language. We are introducing the Macedonian language into our experiments as an example of low-resource language for

which creating AMR models would be both time and resources consuming.

In the following sections we explain in details how this research was conducted. In Section 2 we present previous achievements in the field of parsing and generating multilingual AMRs. In Methodology, we are explaining the whole pipeline, the dataset, and the evaluation metrics that are being used to compare the results in Section 4. Finally, we conclude our work and highlight the achievements in the final section 5.

2 RELATED WORK

This section presents a brief review of the recent achievements in the field of multilingual AMR parsing and AMR-to-text generation. While the recent work on both of these tasks is mostly focused on English models, the field of cross-lingual AMR still remains unexplored and a great challenge for research.

The very first research that tackles the issue of multilingual parsing and evaluation (Damonte and Cohen, 2017) uses annotation projection to generate AMR graphs in the target language. They project AMR annotations to a target language using word alignments which means if a source word is word-aligned to a target word, and AMR aligned with a graph node, then the target word is aligned to that node. The newly-generated graphs are used for training AMR parsers. The second problem that they are tackling is the evaluation dataset, that is, the lack of parallel corpora to compare the AMRs graphs. For the validation process a so called silver dataset is used, which is generated in the same way as the training data and relies on the same errors influenced by the errors of the English AMR parser and the projection errors. Therefore, there is a need of a gold dataset. They are solving this problem by inverting the projection process and train another English parser from the target parser. The new parser is then evaluated on a exiting English gold dataset. The comparison between the parsers is done using Smatch score. The pipeline is used on Italian, German and Spanish data retrieved from the Europarl dataset and Chinese data retrieved from TED talks corpus (Cettolo et al., 2012).

Another research in the field of cross-lingual AMR parsing is (Blloshmi et al., 2020). Here the cross-lingual AMR parsing is enabled using Transfer learning techniques. For a given sentence in English and its translations in different languages, first a concept identification is made, and then a relations identification. For the concept identification task, they train sequence-to-sequence model which when a list of words in another languages is given as an input it

generates list of nodes in the original English AMR formalism as output. The new trained network is used to dispose the AMR alignments described in the pioneer paper (Damonte and Cohen, 2017) in this field. The second step is the relations identification. As stated in the paper, this task was inspired by the arc-factored approaches for dependency parsing. Over the identified concepts from the previous task a search is run for the maximum-scoring connected subgraph. Using deep biaffine classifier a prediction is made whether there is an edge between two nodes and what is its label.

In (Cai et al., 2021) a new approach is proposed where they train and fine-tune one multilingual AMR parser for all different languages including English. Similarly to (Biloshmi et al., 2020) they dispose words and nodes alignments using Transformer’s sequence-to-sequence models. Before training, they utilize parameters for initialization of the encoder and the decoder of two pre-trained architectures, one for multilingual denoising autoencoding and the other, for multilingual machine translation. Via knowledge distillation they are trying to transfer the knowledge of an English AMR parser to their multilingual AMR parser and additionally fine-tune the resulting parser on gold AMR graphs. They use gold and silver dataset, and Europarl’s corpus for the knowledge distillation process and evaluate its performances using Smatch.

In contrast to the other papers, (Fan and Gardent, 2020) resolves the issue of multilingual AMR-to-text generation. The model is sequence-to-sequence model, with Transformer encoder and decoder where the input is English AMR which then is used as sample to generate multilingual text from it. Different datasets are used, but for all of them the jamr parser is used. The performance of the model is measured using the BLEU metric, where the multilingual output is compared to its parallel pair.

So far, multiple papers have been discussed for either AMR parsing, or AMR-to-text generation. The research presented in (Xu et al., 2021) describes cross-lingual models that can be used for both AMR parsing and AMR-to-text generation. These models are furthermore fine-tuned on different tasks to explore their performances. The pre-training of the models is done on three tasks, AMR parsing, AMR-to-text generation and machine translation, via multi-task learning. The translation task is included to help the decoder to generate more fluent foreign sentences and to help the encoder to capture beneficial syntax and semantic information. The proposed approach demonstrates higher results when compared to the previous multilingual AMR parsing research, and as

higher results as those in (Fan and Gardent, 2020). For comparison of the performance of the model, for the parsing task the Smatch metric is used and the BLEU metric is used for the generating task.

Another interesting research that has been done in this field is a baseline translation model similar to our proposed approach. Experiments are made with translating multilingual sentences into English sentences and then parsing them with strong English AMR parser. For the translation process, they use Neural Machine Translation system HelsinkiNLP’s Opus-MT models (Tiedemann et al., 2020) and as AMR parser, a model from the Amrlib¹ library is used. The described pipeline is utilized on the benchmark LDC2020T07² for German, Italian, Spanish and Mandarin. As evaluation metric in this parsing task, the Smatch metric is used. In comparison with our method, we are adding the AMR-to-text task to complete the pipeline and create end-to-end multilingual AMR system.

3 METHODOLOGY

This section presents our full end-to-end cross-lingual AMR parsing and text generation pipeline (xAMR)³ in details. The first section is related to the dataset we use for our experiments, and in the following part we describe the approach used to achieve cross-lingual AMR parsing and text generation.

3.1 Datasets

In this research we evaluate xAMR on five different languages. For four of them, Bulgarian, German, Italian and Spanish, we used the parallel corpus from Europarl dataset extracted from the Proceedings of the European Parliament. The corpus is available in 11 different languages and for each language there is a parallel English corpus. Since there is no such paired corpus available for the Macedonian language, a human expert manually translated a subset of 1000 Europarl sentences originally given in English language.

3.2 The xAMR Pipeline

Our technique for multilingual AMR parsing and text generation presented in the following sections is depicted in Figure 1. We provide a new and efficient

¹<https://github.com/bjascob/amrlib>

²<https://catalog.ldc.upenn.edu/LDC2020T07>

³<https://github.com/taskop123/xAMR-Cross-lingual-AMR-End-to-end-Pipeline>

end-to-end multilingual pipeline that solves the problem of huge resources needed for training such models. This unique process includes the usage of state-of-the-art models for translating sentences, and AMR model for parsing and generating English sentences.

3.2.1 Forwards Translation

For the translation of the initial non-English sentence into English, we use two translation pipelines, T_x-tai⁴ and DeepTranslator’s GoogleTranslator⁵ so we can make a comparison of the influence of the chosen translator in our approach.

3.2.2 AMR Parsing

The next step in the xAMR is parsing the translated English sentence into an AMR graph. For this purpose, we use the parser from Amrlib¹. The AMR parser is a pre-trained T5-base model (Roberts et al., 2020) that has been fine-tuned on English sentences and their corresponding AMR graphs using the LDC2020T02⁶ benchmark dataset.

3.2.3 Text Generation

The following phase is a generation of an English sentence from the AMR graph constructed in the previous step. To generate text from AMR graph, we use the Amrlib’s graph to sentence T5-base model.

3.2.4 Backwards Translation

The final step in xAMR is translating the generated English sentence into the original source language. For achieving such translations, we use the aforementioned translation pipelines for Macedonian, German, Italian, Spanish and Bulgarian language.

3.3 Evaluation

The main evaluation metric that we use to evaluate our methodology is cosine similarity on the sentence embeddings to capture the semantic similarity between the sentences. For each of the four steps, translation into English, AMR parsing, AMR-to-text generation and finally translation into the initial, source language, we generate parallel sentences and measure the cosine similarity of their sentence embeddings. We employ multilingual sentence embedding because the sentences in our methodology are not only in English.

⁴<https://neuml.github.io/Txtai/pipeline/text/translation/>

⁵<https://pypi.org/project/deep-translator/#id1>

⁶<https://catalog.ldc.upenn.edu/LDC2020T02>

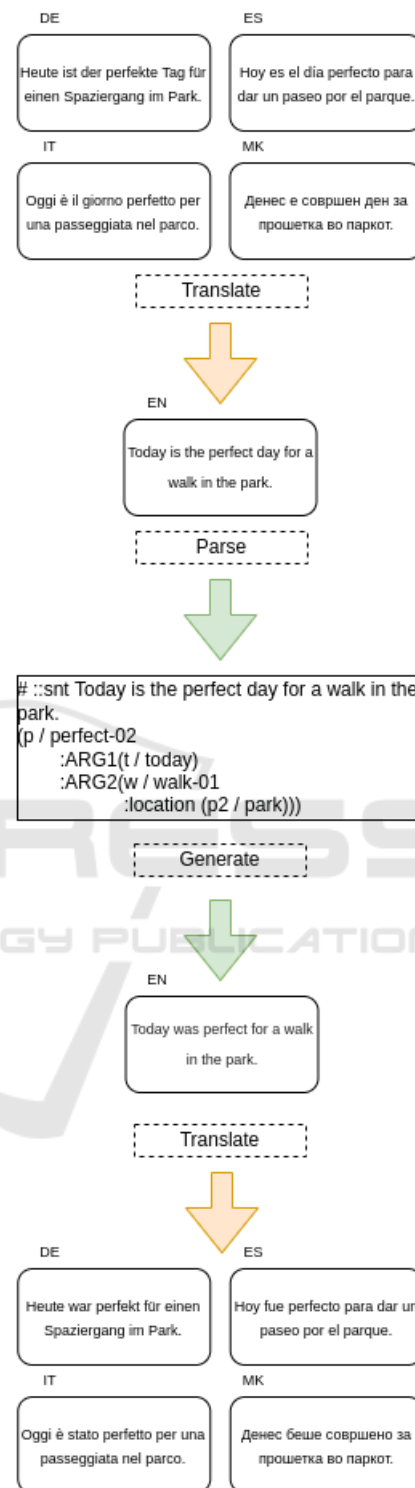


Figure 1: Visual representation of the xAMR pipeline.

The multilingual sentence embeddings that we apply are LASER (Language Agnostic Sentence Rep-

representations)⁷ (Chaudhary et al., 2019) which uses BiLSTM encoder that understands 93 different languages; LaBSE (Language-agnostic BERT Sentence Embedding) (Feng et al., 2020) which is multilingual BERT embedding model that produces cross-lingual sentence embeddings for variety of languages, and Distiluse-Base-Multilingual-Cased-v2 (Reimers and Gurevych, 2019) (Reimers and Gurevych, 2020) which is a modification of the BERT model and is able to derive semantically meaningful sentence embeddings using modified networks.

In addition, we also measure BLEU and ROUGE scores between the input sentences and the output sentences in their initial language. We chose BLEU as a standard metric in translation tasks to measure how much the syntax and the word order alter between the stages of the xAMR pipeline. ROUGE is a similar metric to BLEU. It calculates how many of the overlapping n-grams in the input sentence appeared in the generated output, whereas BLEU focuses more on the overlapping n-grams in the generated sentence that appeared in the input sentence.

4 EXPERIMENTS AND RESULTS

This section presents comparison of our proposed xAMR pipeline to the multilingual AMR-to-text generation model (mAMR) (Fan and Gardent, 2020). As the mAMR has already been evaluated on a subset of the Europarl dataset, we have modified our Europarl datasets to suit the available testing dataset for mAMR. Therefore, 1000 parallel sentences in German, Spanish, Italian, and Bulgarian were selected with their parallel English sentences and the corresponding simplified AMR graph representations. These AMR representations are needed as input for the mAMR model, which generates sentences in the selected language. In this comparison, we exclude the Macedonian language since the mAMR model does not support the Macedonian language. In this manner, we obtain four datasets containing original English sentences, their respective simplified AMR graph, and the parallel sentence in the chosen language that serves as a reference.

The Macedonian language dataset is used only to evaluate the xAMR pipeline, however, the success of a mAMR trained on Macedonian could be assumed from the Bulgarian mAMR as a language most similar to the Macedonian language.

For the translation processes in the pipeline, two different translators had to be considered due to the

inconsistencies in the Bulgarian language translation. The translators that are applied are Txtai on the German, Italian, Spanish and Macedonian dataset and the DeepTranslator’s GoogleTranslator to the Bulgarian dataset.

Considering that the English AMR graphs are acquired from the original English sentences and run through the mAMR model, we skip the first step of our pipeline. Namely, instead of beginning with a non-English sentence, we begin our pipeline from the parsing stage. As input, we have the original English sentence, which is first parsed into an AMR graph. Next, the AMR graph generates another English sentence, which, lastly, is translated to the selected language. This skipping stage is done to obtain a more reliable comparison between the models. Finally, the output of both models is compared to the reference sentence in the selected language to understand the significance of the results.

Table 1 presents the cosine similarity score calculated between the different intermediate states of the input sentence when passed through the pipeline. X denotes the input sentence in the selected language, EN denotes the translated sentence to English, and AMR represents the parsing and the generating task. When the notation is shown in uppercase letters, we compare those two intermediate outputs.

Each column of the tables represents the cosine similarity computed between the original non-English sentence compared with its translated English version, the English sentence after its AMR generation, and the final output of the pipeline, respectively. The cosine similarity is calculated over the different sentence embeddings.

When the sentences are embedded with LASER, the scores are the highest compared to the others embeddings. Moreover, from the $X \rightarrow en \rightarrow amr \rightarrow en \rightarrow X$ column, we see that the original input sentence and the final output of our pipeline have leading scores, hence we can conclude that the pipeline truly keeps the semantic meaning of the sentences. In terms of the languages, the Spanish dataset obtained best scores regardless of the sentence embeddings.

The non-English sentences, the inputs in our pipeline and their respective outputs are further compared using the BLEU and ROUGE metrics. Table 2 presents the results when the sentences are evaluated with BLEU. When calculating the BLEU scores, we take into consideration the overlapping matches in unigrams BLEU-1, bigrams BLEU-2, trigrams BLEU-3 and weighted score BLEU-W. The weighted score calculates the number of matches of unigrams, bigrams, trigrams and four-grams each with 25% weight. In comparison with the cosine similarity,

⁷<https://github.com/facebookresearch/LASER>

Table 1: Cosine similarity comparison between the original non-English input and the intermediate outputs of the pipeline.

SE Model	Language	X→EN→amr→en→x	X→en→amr→EN→x	X→en→amr→en→X
LASER	DE	0.9511	0.8953	0.9633
	ES	0.9670	0.9040	0.9769
	IT	0.9639	0.8983	0.9754
	BG	0.9196	0.6254	0.9593
	MK	0.9357	0.8798	0.9802
LaBSE	DE	0.8879	0.8458	0.9593
	ES	0.9248	0.8755	0.9757
	IT	0.9209	0.8685	0.9696
	BG	0.8980	0.7130	0.9531
	MK	0.8849	0.8393	0.9792
Distiluse	DE	0.9263	0.8893	0.9582
	ES	0.9501	0.9098	0.9731
	IT	0.9394	0.8969	0.9679
	BG	0.8693	0.6842	0.9405
	MK	0.9327	0.8972	0.9754

Table 2: Results for BLEU.

Language	BLEU-1	BLEU-2	BLEU-3	BLEU-W
DE	0.6677	0.4880	0.4209	0.4548
ES	0.7622	0.6213	0.5343	0.5760
IT	0.7010	0.5412	0.4511	0.4952
BG	0.6617	0.5381	0.6511	0.5989
MK	0.8374	0.7376	0.6753	0.7041

Table 3: Results for ROUGE F1 Score.

Language	ROUGE-1	ROUGE-2	ROUGE-L
DE	0.6944	0.4754	0.6703
ES	0.7867	0.6221	0.7703
IT	0.7327	0.5467	0.7152
BG	0.6609	0.5012	0.6591
MK	0.8602	0.7445	0.8546

BLEU obtains lower results because of the paraphrasing that happens during the generation from AMR and the translation processes. Nevertheless, BLEU only provides an overview of the similarity of the syntax and the word order. In the same manner, we compare the sentences with ROUGE calculating matches in unigrams, bigrams and longest common sequence using the F1 measure. The results are shown in Table 3. From the both tables, the conclusion is that the Spanish dataset, again, has the highest BLEU and ROUGE scores, hence the highest scores with the cosine similarity metric, because good amount of the word are overlapping in both input and output sentences.

To evaluate the performance of our method, we compare it with mAMR as previously explained. The English input sentence in both methods is compared with the reference non-English sentence and the generated sentences from the models. The cosine similarity is computed between the reference sentence and the generated outputs. Table 4 presents the obtained results for the cosine similarity with the three sentence embeddings. With OrgX we label the reference sentences, and with GenX we label the corresponding generated sentences. Again, the best results

are acquired when LASER is used for sentence embeddings. It can be noticed that our method produces sentences with higher similarity with the original English sentence at input, unlike mAMR.

The results show that the presented xAMR pipeline generates outputs with less information loss in contrast to the results of mAMR, when compared to the reference non-English sentence. The cosine similarity metric shows almost constant improvement in range 0.03 to 0.04 at xAMR for all the languages, regardless of the applied sentence embedding.

Table 5 and 6 present the results acquired with BLEU and ROUGE when the reference non-English sentence is compared with the generated outputs from our xAMR and the mAMR model. Once more, xAMR surpasses the mAMR model by 0.01 to 0.09 depending on the metric and the counted n-grams, except in two cases when evaluating with BLUE-3 on the German and Italian dataset.

Taking these conclusions into account we can confirm that our method successfully outperforms the previous method for multilingual AMR-to-text generation without the need of pre-training models on large annotated data.

Table 7 displays the results of a comparison between the output of our pipeline and the generated sentences from mAMR. The purpose of comparing these results is to observe how similar sentences both of the strategies produce. The metrics for comparison of such sentences are cosine similarity on the sentence embeddings generated from three different sentence embedding models, BLEU and ROUGE scores. The results obtained from the evaluation of cosine similarity is high, which means that both methods provide sentences with similar meanings, but have distinct sentence organization observed from the BLEU and ROUGE metrics.

Table 4: Cosine similarity comparison on mAMR and xAMR.

SE Model	Language	EN → OrgX		EN → GenX		OrgX → GenX	
		xAMR	mAMR	xAMR	mAMR	xAMR	mAMR
LASER	DE	0.8968	0.8968	0.9543	0.8902	0.9066	0.8645
	ES	0.9070	0.9070	0.9664	0.8974	0.9141	0.8650
	IT	0.8986	0.8986	0.9633	0.8944	0.9088	0.8636
	BG	0.9417	0.9417	0.9643	0.9151	0.9566	0.9139
LaBSE	DE	0.8218	0.8218	0.8835	0.8261	0.8956	0.8617
	ES	0.8592	0.8592	0.9246	0.8633	0.9081	0.8666
	IT	0.8459	0.8459	0.9233	0.8611	0.8902	0.8501
	BG	0.8925	0.8925	0.9215	0.8737	0.9489	0.9113
Distiluse	DE	0.8547	0.8547	0.9389	0.8943	0.8873	0.8583
	ES	0.8696	0.8696	0.9536	0.9096	0.8939	0.8633
	IT	0.8478	0.8478	0.9456	0.8977	0.8749	0.8436
	BG	0.9059	0.9059	0.9447	0.9020	0.9369	0.9013

Table 5: BLEU score comparison on mAMR generation and xAMR.

Language	BLEU-1		BLEU-2		BLEU-3		BLEU-W	
	xAMR	mAMR	xAMR	mAMR	xAMR	mAMR	xAMR	mAMR
DE	0.4971	0.4354	0.3577	0.3194	0.4016	0.4367	0.3643	0.3522
ES	0.5727	0.4982	0.4126	0.3371	0.4003	0.3610	0.3990	0.3386
IT	0.5032	0.4327	0.3641	0.3194	0.3771	0.3826	0.3572	0.3298
BG	0.6585	0.5753	0.5101	0.4256	0.4508	0.4075	0.4770	0.4078

Table 6: ROUGE F1 score comparison on mAMR and xAMR.

Language	ROUGE-1		ROUGE-2		ROUGE-L	
	xAMR	mAMR	xAMR	mAMR	xAMR	mAMR
DE	0.5239	0.4716	0.2800	0.2196	0.4943	0.4252
ES	0.6021	0.5372	0.3751	0.2939	0.5719	0.4785
IT	0.5248	0.4660	0.3022	0.2377	0.4977	0.4211
BG	0.6852	0.6081	0.4835	0.3892	0.6730	0.5735

Table 7: xAMR vs mAMR.

Language	LASER	DistilUse	LaBSE	BLEU-1	BLEU-2	BLEU-3	BLEU-W	ROUGE-1	ROUGE-2	ROUGE-L
DE	0.9033	0.9261	0.9181	0.5957	0.4244	0.4019	0.4091	0.6275	0.3845	0.5740
ES	0.9075	0.9339	0.9222	0.6629	0.4933	0.4264	0.4575	0.6902	0.4810	0.6277
IT	0.9056	0.9233	0.9132	0.5914	0.4292	0.3819	0.4056	0.6265	0.4055	0.5747
BG	0.9277	0.9274	0.9282	0.6419	0.4950	0.4453	0.4647	0.6619	0.4578	0.6269

5 CONCLUSIONS

This paper presents an effort to eliminate the need for large corpora for achieving multilingual AMR representation for a particular language other than English. The main goal of the paper is to propose a novel approach by introducing cross-lingual AMR end-to-end pipeline referred to as xAMR which utilizes state-of-the-art translation and embeddings models along with English-based AMR parsing and generating tasks.

The efficiency of the proposed pipeline has been compared with the existing multilingual AMR pipeline by measuring the sentences' loss of information in terms of cosine similarity, BLEU, and ROUGE scores. To obtain comparable results, Europarl corpus has been used at both pipelines. Additionally, the xAMR pipeline has been evaluated by introducing one low-resource language - the Macedonian lan-

guage.

The results showed that our xAMR significantly surpasses multilingual AMR models for all the languages we experimented with within this paper. Thus, this research revisited the translation as a baseline for developing cross-lingual AMR models.

REFERENCES

- Ballesteros, M. and Al-Onaizan, Y. (2017). Amr parsing using stack-lstms. *arXiv preprint arXiv:1707.07755*.
- Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffith, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

- Blloshmi, R., Bevilacqua, M., Fabiano, E., Caruso, V., and Navigli, R. (2021). SPRING Goes Online: End-to-End AMR Parsing and Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 134–142, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Blloshmi, R., Tripodi, R., and Navigli, R. (2020). Xl-amr: Enabling cross-lingual amr parsing with transfer learning techniques. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2487–2500.
- Cai, D., Li, X., Ho, J. C.-S., Bing, L., and Lam, W. (2021). Multilingual amr parsing with noisy knowledge distillation. *arXiv preprint arXiv:2109.15196*.
- Cai, S. and Knight, K. (2013). Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752.
- Cettolo, M., Girardi, C., and Federico, M. (2012). Wit3: Web inventory of transcribed and translated talks. In *EAMT*.
- Chaudhary, V., Tang, Y., Guzmán, F., Schwenk, H., and Koehn, P. (2019). Low-resource corpus filtering using multilingual sentence embeddings. *arXiv preprint arXiv:1906.08885*.
- Damonte, M. and Cohen, S. B. (2017). Cross-lingual abstract meaning representation parsing. *arXiv preprint arXiv:1704.04539*.
- Damonte, M. and Cohen, S. B. (2019). Structural neural encoders for amr-to-text generation. *arXiv preprint arXiv:1903.11410*.
- Fan, A. and Gardent, C. (2020). Multilingual amr-to-text generation. *arXiv preprint arXiv:2011.05443*.
- Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.
- Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.
- Konstas, I., Iyer, S., Yatskar, M., Choi, Y., and Zettlemoyer, L. (2017). Neural amr: Sequence-to-sequence models for parsing and generation. *arXiv preprint arXiv:1704.08381*.
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Lyu, C. and Titov, I. (2018). Amr parsing as graph prediction with latent alignment. *arXiv preprint arXiv:1805.05286*.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Roberts, A., Raffel, C., and Shazeer, N. (2020). How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.
- Singhal, A. et al. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.
- Song, L., Zhang, Y., Wang, Z., and Gildea, D. (2018). A graph-to-sequence model for amr-to-text generation. *arXiv preprint arXiv:1805.02473*.
- Tiedemann, J., Thottingal, S., et al. (2020). Opus-mt—building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.
- Wang, C. and Xue, N. (2017). Getting the most out of AMR parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1268, Copenhagen, Denmark. Association for Computational Linguistics.
- Wang, C., Xue, N., and Pradhan, S. (2015). A transition-based algorithm for AMR parsing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375, Denver, Colorado. Association for Computational Linguistics.
- Wang, T., Wan, X., and Jin, H. (2020). Amr-to-text generation with graph transformer. *Transactions of the Association for Computational Linguistics*, 8:19–33.
- Xu, D., Li, J., Zhu, M., Zhang, M., and Zhou, G. (2021). XLPT-AMR: Cross-lingual pre-training via multi-task learning for zero-shot AMR parsing and text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 896–907, Online. Association for Computational Linguistics.
- Zhang, S., Ma, X., Duh, K., and Van Durme, B. (2019). Amr parsing as sequence-to-graph transduction. *arXiv preprint arXiv:1905.08704*.
- Zhu, J., Li, J., Zhu, M., Qian, L., Zhang, M., and Zhou, G. (2019). Modeling graph structure in transformer for better amr-to-text generation. *arXiv preprint arXiv:1909.00136*.