Abstract Meaning Representation in Different Languages

Stoica Elias-Valeriu
elias-valeriu.stoica@s.unibuc.ro

Semen Valentin-Ion
valentin-ion.semen@s.unibuc.ro

Abstract

This project documents different attempts of studying how AMR generation behaves from language to language. The tested languages were English, Welsh and Irish. We used a sample (about 1600 sentences) of the massive AMR dataset from which we obtained the Welsh and Irish translation. The new datasets were used in different configurations in order to train different models to observe how they generate AMRs. The fine-tuned models were then used to generate AMR graphs for each sentence in each language and compile them in a CSV.

1 Introduction

Abstract Meaning Representation (AMR) is a way of representing the meaning of a sentence as a structured graph. Instead of just dealing with words, AMR captures **who does what to whom**, the relationships between concepts, and the overall meaning of a sentence in a format that a computer can understand. It's widely used in machine translation, question answering, and text summarization because it provides a deeper understanding of language compared to raw text.

The problem with the basic **AMRlib**, however, is that most models trained for it are focused on English. This limits its usefulness for multilingual tasks and research in other languages. Our project deals with this challenge by training an AMR parser on an English dataset using a multilingual transformer model. To achieve this, we translated the original dataset to Irish and Welsh, using **MarianMT model**. Afterwards, we fine-tuned **Google's t5-small** model to convert sentences into AMR graphs (STOG).

To evaluate the performance of the model, we used the match score, a metric that measures how well the generated AMR graphs match the expected ones.

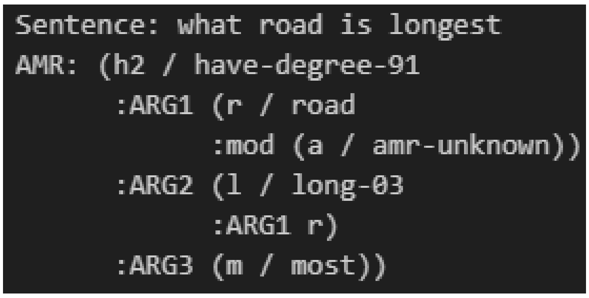
2 Methods

2.1 Dataset

For this project we used a sample of the **Massive-AMR** dataset containing over 1600 English sentences with their related Abstract Meaning Representation graphs.

The dataset covers different types of domains and sentences, such as question answering, factual statements or commands. Each entry contains a **sentence** in the original language (English), an **AMR Graph** that represents the meaning and **additional metadata**.

In order to train the models for multilingual generation, we used the **OPUS-mt** API based on the **MarianMT** model to translate the English sentences from the *"utt"* column into Welsh and Irish.



```
Sentence: what road is longest
AMR: (h2 / have-degree-91
       :ARG1 (r / road
              :mod (a / amr-unknown))
       :ARG2 (l / long-03
              :ARG1 r)
       :ARG3 (m / most))
```

Figure 1: Sample from the Dataset

2.2 Training The T5-small Model on Irish

The first step was to load the newly obtained Irish data by reading the json containing it and filtering out the entries that had missing sentences or missing AMRs, retaining only the sentence and the AMRs. The data was then split into trining, validation and test subsets with 80%, 10%, 10%.

The sentences with an added *"Parse: "* prompt were used as inputs and the AMRs as targets for training the model.

The data was then tokenized and fed into the DataCollator in order to process it in batches.

The model is initialized and trained for 10 epochs, processing the data in batches of 8, with a learning rate of 3e-4 and a weight decay of 0.01 to prevent overfitting.

The next step is to assess the performance of the model by calculating the **Smatch** score using the Irish data. The predictions are obtained using the trained model and are compared to the correct AMRs.

AMRs were checked for valid before the evaluation. The ones that are not valid are skipped and are not used during the **Smatch score** calculation.

The model obtained a **Smatch** score precision of 6.14%, recall of 11.76% and F1 of 11.08%.

Precision	**Recall**	**F1**
6.14 %	11.75 %	11.08 %

Table 1: Smatch scores for T5-small model trained on Irish

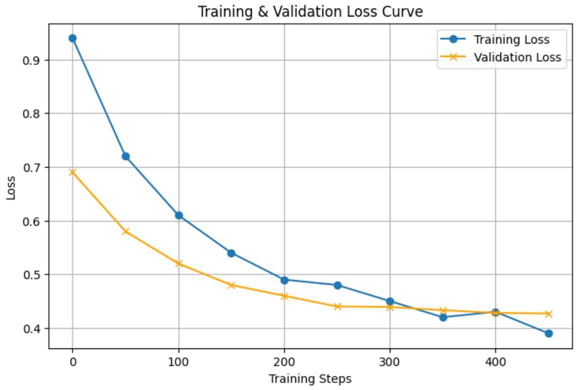


Figure 2: Training & Validation Loss Curve

2.3 Training the T5-small Model on Welsh

In order to train an AMR parsing model for Welsh, we fine-tuned T5-small, a sequence-to-sequence transformer model. We opted for it because it is commonly used for text-to-text tasks, meaning that it fits the task of converting natural language sentences into structured AMR graphs.

As we mentioned before, we used the MASSIVE-AMR dataset, filtering out any incomplete entries. The dataset was split into 80% training data, 10% validation data and 10% test data.

To help the T5 model understand that its task is parsing into AMR graphs, we formatted each input sentence as **"parse: [sentence]"**.

In the same way, we initialized the model with a learning rate of 3e-4, training for 10 epochs with a batch size of 8 per device. A data collator was used for the efficient batch processing. The loss function ignored padding tokens in the labels to prevent unnecessary penalization.

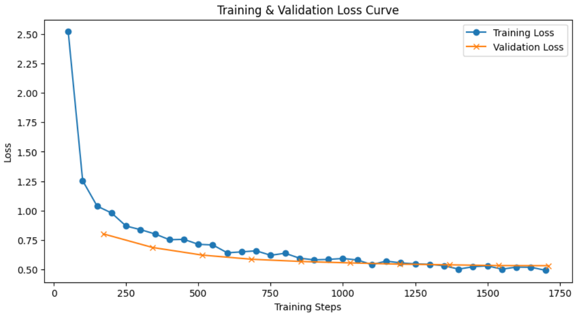


Figure 3: Training & Validation Loss Curve

To track model performance, we recorded training and validation loss. The loss curve shows the training progress. We observe a decline in both scores, indicating that the learning process works well on Welsh text.

The model obtained a **Smatch** score precision of 5.0632%, recall of 11.8316% and F1 of 13.4421%.

Precision	**Recall**	**F1**
5.0632 %	11.8316 %	13.4421 %

Table 2: Smatch scores for T5-small model trained on Welsh

2.4 Training the T5-small Model on Welsh and Irish

In order to make the model multilingual and to better improve its performance, we trained the T5-small Model on a combined dataset of Welsh and Irish sentences and AMR pairs.

The training parameters were similar, with the sentences along the *"Parse: "* prompt as the input and the AMR graphs as the targets.

We maintained the learning rate of 3e-4, the number of training epochs of 10 and a batch size of 8, along with the DataCollator for processing in batches. The weight decay was kept at 0.01 in order to prevent overfitting.

The **Smatch** score for this model was calculated using the combined dataset of both Welsh and Irish sentences, with the AMRs that are not valid being skipped during the evaluation.

The model trained on both languages showed a slight improvement, scoring a **Smatch** score precision of 6.32%, recall of 12.28% and F1 of 11.83%.

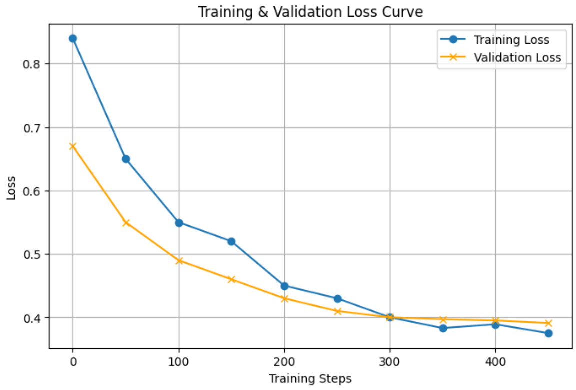


Figure 4: Training & Validation Loss Curve

Precision	Recall	F1
6.32 %	12.28 %	11.83 %

Table 3: Smatch scores for T5-small model trained on Irish and Welsh

2.5 Training the T5-small Model on All Languages

To improve the generalization of AMR parsing across different languages, we fine-tuned T5-small on a merged dataset containing Welsh, English and Irish AMRs. The goal was to train a single model capable of parsing AMRs for multiple languages.

We combined the three datasets and removed incomplete entries, ensuring all examples contained both the sentence and the corresponding AMR graph. We shuffled and split the dataset as before: 80-10-10.

The model was trained following the same sequence-to-sequence paradigm, each input sentence being prefixed with **"parse: "** to specify the task. We maintained the learning rate of 3e-4, training for 10 epoch with a batch size of 8 per device. We used gradient accumulation to balance computational efficiency with model convergence. The optimizer also included **weight decay** to prevent overfitting.

To track model convergence, we monitored training and validation loss at every epoch. As illustrated in the figure, both losses steadily decreased, confirming that the model successfully learned to map multilingual sentences to AMRs. Notably, the validation loss remained close to the training loss, suggesting minimal overfitting and stable generalization across languages.

While multilingual training introduced additional complexity, it proved effective in enhancing the model's robustness for AMR parsing across different linguistic contexts.

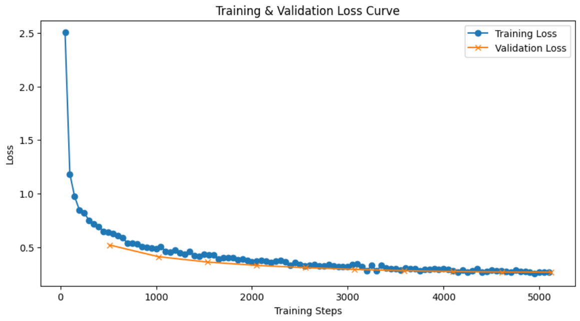


Figure 5: Training & Validation Loss Curve - All Languages

This model which was trained on all languages obtained the best results, scoring a **Smatch** score precision of 7.1939%, recall of 12.7551% and F1 of 11.8776%.

Precision	Recall	F1
7.1939 %	12.7551 %	11.8776 %

Table 4: Smatch scores for T5-small model trained on Irish and Welsh

3 Results

The best Smatch score was output by the small T5 that was trained on the data set made up of the sentences in all three languages, with a **Smatch** score precision of 7.1939%, recall of 12.7551% and F1 of 11.8776%.

All models were used to generate AMR graphs that where then compiled in a csv that contain the following columns: the original sentence, the language,original AMR graph, the AMR predicted by the model that was trained on all languages, AMR predicted by the model trained on Welsh, AMR predicted by the model trained on Welsh and Irish, AMR predicted by the model trained on Irish, AMR generated by the original STOG model from the AMRlib.

The compiled CSV contained 5055 entries.

Model	Precision	Recall	F1
Ir	6.14 %	11.75 %	11.08 %
Wl	5.0632 %	11.8316 %	13.4421 %
Wl + Ir	6.32 %	12.28 %	11.83 %
All	7.1939 %	12.7551 %	11.8776 %

Table 5: Smatch scores for all models

4 Discussion

Overall, the **T5-small** models that were fine-tuned on specific languages (Welsh and Irish) outperform the default **STOG** model in the **AMRlib** when it comes to AMR similarity.

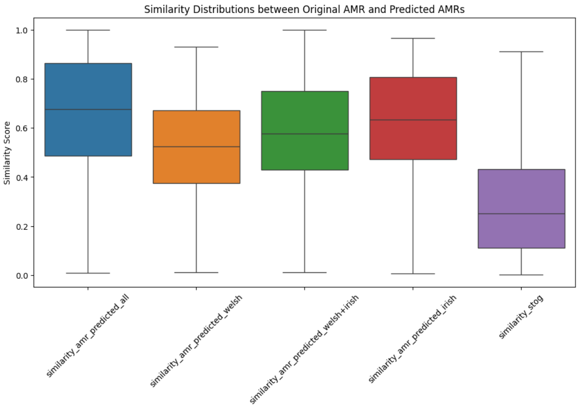


Figure 6: Similarity Distributions between Original AMR and Predicted AMRs

The plot of similarity distributions shows that the AMR graphs predicted by the model trained on Welsh and Irish data produce a higher variance, while the Irish model maintains a more stable prediction structure.

The low scores produced by the STOG model show that the default model from the AMRlib requires fine-tuning in order to improve its performance on non-English languages.

The correlation heatmap provides further insights into the relationships between different AMR representations. The model trained on Welsh and Irish and the model trained only on Irish show a strong correlation of **0.75**, indicating that these models share similar patterns. In contrast, the STOG AMRs show weak correlation with the other models, showing that dialects and non-English sentences require previous training.

We generated scatter plots which show the relationships between different types of similarity scores for American English (AMR) and other dialects.

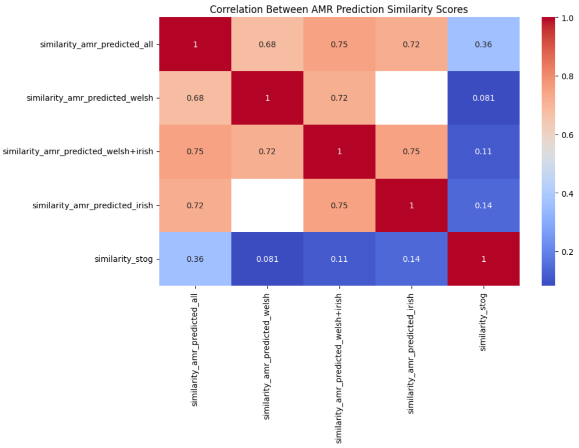


Figure 7: Correlation Heatmap between AMR Prediction

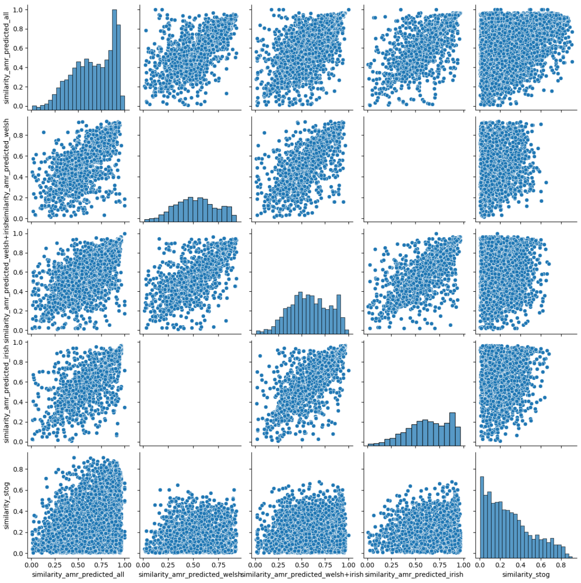


Figure 8: Scatter Plots to Analyze Relationships between Models

If points in a scatter plot tend to cluster together, it indicates a strong correlation between the two of them. For example, the plots showing the realtionship between **"similarity_amr_predicted_irish"** and **"similarity_amr_predicted_welsh+irish"** show a clear clustering pattern. This implies that models trained on Irish and Welsh languages are producing similar scores.

When points are more spread out, it suggests a weaker correlation between the two variables. This implies that standard English models struggle to produce similarity scores consistent with those of the dialectal models.

5 Conclusions

In conclusion, the model for generating AMR graphs require fine-tuning on different languages to produce more accurate results in terms of **similarity** and **Smatch** scores. We managed to generate a dataset of over 5000 AMR graphs in three different languages using different models trained on different combinations of datasets. However, all metrics could be heavily improved.

6 Further Improvements

There are many ways we could improve our results, but we could mostly focus on enhancing the metrics by training the models on larger datasets or utilizing more performant pre-trained models.

Another improvement could involve incorporating another language in the training process, such as Scottish.

During our experimentation, we attempted using T5-base and other models; however, the process was extremely demanding and time-consuming. We encountered memory errors on multiple occasions, even though our system has 16GB of RAM. This highlights the necessity of having high-performance hardware to train such models effectively.

Additionally, we explored training BERT, AMR-lib's STOG, and other related approaches. These experiments provided valuable insights but also reinforced the challenge of computational limitations when working with complex models for AMR parsing.

7 Related Work

7.1 Multilingual Abstract Meaning Representation for Celtic Languages

The authors, Johannes Heinecke and Anastasia Shimorina, present an approach to create a multilingual text-to-AMR model for three Celtic languages, Welsh (P-Celtic) and the closely related Irish and Scottish-Gaelic (Q-Celtic). The main success of this approach are underlying multilingual transformers like mT5. (Heinecke and Shimorina, 2022)

7.2 Multi-dialect Neural Machine Translation and Dialectometry

The authors, Kaori Abe, Yuichiroh Matsubayashi, Naoaki Okazaki, and Kentaro Inui, present a multi-dialect neural machine translation (NMT) model tailored to Japanese. Their experimental results-demonstrate that this model can outperform a baseline dialect translation model. In addition, they show that visualizing the dialect embeddings learned by the model can facilitate geographical and typological analyses of dialects. (Abe et al., 2018)

7.3 Domain and Dialect Adaptation for Machine Translation into Egyptian Arabic

The authors, Serena Jeblee , Weston Feely , Houda Bouamor, Alon Lavie, Nizar Habash and Kemal Oflazer, present a statistical machine translation system for English to Dialectal Arabic (DA), using Modern Standard Arabic (MSA) as a pivot. (Jeblee et al., 2014)

7.4 MASSIVE-AMR

The author, Emilio Monti, gathered a dataset with more than 84,000 text-to-graph Abstract Meaning Representation (AMR) annotations for 1,685 information-seeking utterances mapped to 50+ typologically diverse languages. (Science, 2025)

References

Kaori Abe, Yuichiroh Matsubayashi, Naoaki Okazaki, and Kentaro Inui. 2018. Multi-dialect neural machine translation and dialectometry. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, pages 1–10, Hong Kong. PACLIC.

Johannes Heinecke and Anastasia Shimorina. 2022. Multilingual abstract meaning representation for celtic languages. In *Proceedings of the CLTW 4 @ LREC2022*, pages 14–19, Marseille, France. European Language Resources Association (ELRA). Licensed under CC-BY-NC-4.0.

Serena Jeblee, Weston Feely, Houda Bouamor, Alon Lavie, Nizar Habash, and Kemal Oflazer. 2014. Domain and dialect adaptation for machine translation into egyptian arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 196–206, Doha, Qatar. Association for Computational Linguistics.

Amazon Science. 2025. Massive-amr: A large-scale dataset for amr parsing. Accessed: 2025-02-03.