Less is Enough: Less-Resourced Multilingual AMR Parsing

Bram Vanroy, Tim Van de Cruys
KU Leuven
Oude Markt 13, Leuven, Belgium
bram.vanroy@kuleuven.be, tim.vandecruys@kuleuven.be

Abstract

This paper investigates the efficacy of multilingual models for the task of text-to-AMR parsing, focusing on English, Spanish, and Dutch. We train and evaluate models under various configurations, including monolingual and multilingual settings, both in full and reduced data scenarios. Our empirical results reveal that while monolingual models exhibit superior performance, multilingual models are competitive across all languages, offering a more resource-efficient alternative for training and deployment. Crucially, our findings demonstrate that AMR parsing benefits from transfer learning across languages even when having access to significantly smaller datasets. As a tangible contribution, we provide text-to-AMR parsing models for the aforementioned languages as well as multilingual variants, and make available the large corpora of translated data for Dutch, Spanish (and Irish) that we used for training them in order to foster AMR research in non-English languages. Additionally, we open-source the training code and offer an interactive interface for parsing AMR graphs from text.

Keywords: AMR parsing, abstract meaning representation, semantics, corpora

1. Introduction

Abstract Meaning Representation (AMR, Section 2; Banarescu et al., 2013) is a meta-language for describing the semantic content of natural language sentences. It is agnostic to surface form (syntactic and lexical) and attempts to capture the meaning of a sentence in its most abstract form. While nodes are technically labelled with a linguistic form (typically a lemma optionally with a sense ID), these may as well be represented as an arbitrary identifier because they refer to a “meaning” rather than a lexical realisation of a meaning. Thanks to its machine-readable data format (as a directed, rooted graph, or as a sequence of triples) AMR has been employed for a variety of natural language processing (NLP) purposes (Section 3). However, the application of AMR to languages other than English has been stymied by the scarcity of large, annotated datasets that are suitable in size for training deep learning systems. AMR corpora exist, notably the English AMR 3.0 corpus (Knight et al., 2020), but manual annotation is costly and time-consuming. This means that AMR data sources are scarce, particularly for non-English languages.

The issue of resource scarcity is not only confined to languages that are commonly considered low-resource. Even languages like Dutch, which enjoys a relatively higher degree of digital presence and is spoken by around 24 million people, face challenges in annotated data for specialised tasks such as AMR. Even for Spanish, the fourth most spoken language in the world, there is a lack of suitable datasets for building deep learning systems for this task. In terms of task-specific resources, such languages are still less-resourced – their mid-to-high resource nature in the traditional sense unfortunately does not transfer to a high availability of annotated data for all NLP tasks. Addressing this scarcity in terms of data availability, models, and research is crucial for the democratisation of NLP technologies and to ensure that the benefits of automating semantic AMR parsing is not confined to English.

In this context, to seek alternative approaches for performant, non-English text-to-AMR systems, multilingual models offer a promising avenue for exploration. Not only are these models computationally more efficient (training one multilingual model is more economical than training multiple monolingual ones); they also offer the advantage of easier deployment, as a single model can handle multiple languages. This efficiency is particularly salient in scenarios where computational resources are limited, a common situation in academic research and in deployments in less-resource environments. Moreover, multilingual models can be less data-hungry when training for each individual language, thereby partially mitigating the issue of data scarcity, which is the main topic of this paper.

This paper aims to investigate the efficacy of multilingual models in the task of text-to-AMR parsing, focusing particularly on English, Dutch, and Spanish. English serves as a well-resourced Germanic language, boasting a large, human-annotated AMR corpus (around 60,000 entries; Knight et al., 2020). In contrast, Dutch (also Germanic) and Spanish (Romance) are “less-resourced languages” in terms of AMR resources.

20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

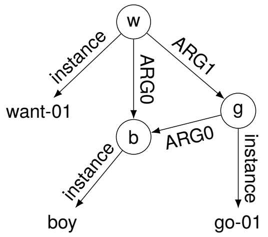
While both languages are widely spoken, they lack annotated and sizeable AMR corpora suitable for machine learning. However, their otherwise higher-resource status does allow for high-quality, automated machine translation (MT). We therefore make use of state-of-the-art MT systems to generate silver datasets for these languages, which we can then use to train deep learning systems (Section 4). We make available these datasets for other researchers as a tangible contribution.

We empirically and statistically evaluate multilingual (English, Spanish, Dutch) models under various configurations and compare them with monolingual counterparts to understand the trade-offs involved in terms of performance on the one hand and computational and data efficiency on the other. Specifically, we gauge how large the performance gap is between monolingual, full-resource models compared to artificially limited-resource, multilingual ones that have been trained on a subset of the data, and other multilingual models that were trained on the combined, full datasets of all languages. Our objective therefore is not to set new state-of-the-art results, although to the best of our knowledge our Dutch models are the best single-model text-to-AMR parsers for Dutch. Instead we offer insights into the advantages and disadvantages of multilingual text-to-AMR parsing and scrutinise the impact of data scarcity.

We provide valuable resources for the broader research community by publishing the models (monolingual models for English, Spanish and Dutch, as well as multilingual ones), the translated datasets for Dutch, Spanish, and Irish Gaelic (the latter not used in this paper but mentioned because it is part of our data release), the training and processing code, and an online interface to generate graphs from text.[1]

2. Abstract meaning representation

AMR describes the meaning of a sentence in terms of “who does what to whom”, in an abstract form that is not bound by lexical or syntactic overt realisations. Therefore different sentences with the same meaning should have the same AMR realisation. AMR can be written as a directed, rooted graph (Figure 1a), e.g. the meaning of a sentence such as “The boy wants to go.” can be denoted with variables that can be used for (co)reference, such as w, b and g. Leaves in the graph are *concepts* so that the variable g refers to the concept `go-01`. These concepts are English words, special entities, or PropBank framesets (Kingsbury and Palmer, 2002), identifiable by their sense identifiers, such as `want-01`, which refers to the first meaning of *want* in the PropBank.[2] Special entities that are specific to AMR include concepts such as `phone-number-entity` and `world-region`. For an exhaustive description of AMR, see the annotation guidelines.[3]



(a) Graph notation

```
(w / want-01
    :ARG0 (b / boy)
    :ARG1 (g / go-01
        :ARG0 b))
```

(b) PENMAN notation

```
( <P1> want-01 :ARG0 ( <P2> boy )
:ARG1 ( <P3> go-01 :ARG0 <P2> ) )
```

(c) Depth-first linearisation following Bevilacqua et al. (2021) (cf. Section 4.1)

Figure 1: AMR notations for the sentence “The boy wants to go.”. Adapted from Banarescu et al. (2013)

The edges in an AMR graph are labelled with the relationships between two nodes, or, rather, the role of the targeted node. Such relationships can be frame arguments that follow PropBank (such as the ARGn roles); general semantic roles such as `:condition` or `:accompanier`; quantities such as `:quant` or `:unit`; date entities like `:day` or `:decade`; and enumerations of different operators in `:op` roles.

An AMR graph can be considered as logical triples of the following types of information: relationships, variables and concepts. Each triple is of the type `role(source, target)` (e.g. `instance(w, want-01)` or `:ARG0(w, b)`.

While the graph notation (and the underlying logical triples) is intended for computational readability, AMR can also be written in PENMAN notation (Matthiessen and Bateman, 1991), which makes it easier to read and write (Figure 1b).

3. Related research

3.1. Datasets

The English-oriented AMR 2.0 and 3.0 corpora (Knight et al., 2017, 2020) have been the cornerstone of much progress in English AMR generation and parsing. These datasets have been made

[1] All resources can be found here: https://github.com/BramVanroy/multilingual-text-to-amr

[2] https://github.com/propbank/propbank-frames/tree/main/frames

[3] https://github.com/amrisi/amr-guidelines/blob/master/amr.md

available through the Linguistic Data Consortium.[4] AMR 2.0 contains 39,260 AMR annotations within the domain of news and weblog data. AMR 3.0 expands on that with 59,255 annotations in total, containing broadcasts and weblogs but also literary translations and Wikipedia articles. For multilingual purposes, the test set of the AMR 2.0 corpus has been partially translated to Spanish, German, Italian and Chinese Mandarin (1371 sentences per language; Damonte and Cohen, 2020), specifically for cross-lingual parsing. In this corpus, descriptively called “AMR 2.0 – Four Translations”, only the English source sentences were translated – the AMR structures remained unchanged. While such resource has been proven useful in multilingual research on AMR, its small size prohibits larger-scale experimentation and applicable.

In this work, we are interested in generating AMR for English but also for Dutch and Spanish. To the best of our knowledge, manually created or verified AMR corpora do not exist for Dutch. For Spanish, in addition to the limited translated AMR 2.0 partition mentioned above, laudable, manual efforts exist to create language-specific corpora. For instance, Migueles-Abraira (2017) annotated 50 sentences from Antoine de Saint-Exupéry’s novella *The Little Prince* translated into Spanish. Wein et al. (2022), on the other hand, defined annotation guidelines for Spanish and applied those guidelines to 486 Spanish sentences from the aforementioned “Four Translations” corpus to create a small but manually annotated gold corpus of Spanish AMR.

To collect multilingual data for AMR-to-text generation, Fan and Gardent (2020) were inspired by the methodology of Damonte and Cohen (2018) to make use of Europarl to create synthetic multilingual data. Europarl is very domain-specific and contains sentence-aligned parliamentary debates for English and many EU languages. The authors first automatically generate AMR from English sentences in the corpus with an existing text-to-AMR system for English. Because the corpus is aligned on the sentence level, this means that the same AMR of an English sentence, is also compatible with the same sentence in the other languages. The resulting, domain-specific, synthetic dataset is not publicly available.

The annotation efforts above are noteworthy and have had a positive impact on the field. However, on the one hand deep learning experiments often require a significantly larger dataset than the manual annotations in Spanish have provided so far, and on the other hand one may prefer general-domain AMR annotations over domain-specific ones for broad applicability. An AMR dataset for Dutch simply does not exist yet.

3.2. AMR parsing

In research on automated text-to-AMR parsing, most work has focused on English – which in part can be attributed to the availability of large corpora, suitable for machine learning, such as the AMR 2.0 and 3.0 corpora described above (Knight et al., 2017, 2020). Performance of automated systems has increased markedly in the last years thanks to innovations such as the Transformer architecture (Vaswani et al., 2017), transfer learning where a pretrained language model is finetuned on the task of AMR parsing, and the use of automatically created, synthetic data for training (also called “silver” data in contrast to manually created “gold” data).

Bevilacqua et al. (2021), for instance, presented SPRING, a text-to-AMR and AMR-to-text model in **English** that was finetuned on a pretrained BART model (Lewis et al., 2020), outperforming previous approaches. They also showed that, in their set up, incorporating silver data did not positively affect the system’s performance. Following up on that, Bai et al. (2022) went a step further by also exploring pretraining a unified model in all directions: text-to-AMR, AMR-to-text, text-to-text, and AMR-to-AMR for English. Similarly, Cheng et al. (2022) proposed to unify AMR-to-text and text-to-AMR tasks but instead of using silver data they employed Bayesian multi-task learning. Also within the Bayesian paradigm, researchers at IBM (Lee et al., 2022) suggested that relying on self-supervised training with silver data in itself is not sufficient to push parsers’ performance higher anymore. In addition, they suggest the use of ensembling multiple system outputs together in combination with distillation for improved performance and efficiency. Noteworthy here is that they also apply their findings on Chinese, German, Italian and Spanish models where they set a new state-of-the-art on the “Four Translation” dataset. Their work relies heavily on earlier findings of Zhou et al. (2021), who explicitly integrated structural information of the AMR graph into pretrained language models. In a similar vein, most recently, Vasylenko et al. (2023) also modify the aforementioned Transformer architecture with adapters that are tailored to contain structural graph information, achieving state-of-the-art results as a non-ensemble system through distillation without the use of additional data.

To the best of our knowledge language-specific models for general-purpose **Dutch**-to-AMR parsing do not exist. Prior work has been done on semantic parsing for Dutch, but as noted in Wang and Bos (2022), no annotated AMR corpora exist for Dutch, so research for Dutch is focused on other

[4] AMR 2.0: `https://catalog.ldc.upenn.edu/LDC2017T10`; AMR 3.0: `https://catalog.ldc.upenn.edu/LDC2020T02`

semantic paradigms, such as Discourse Representation Graphs in multilingual settings (Wang et al., 2023). As mentioned before, some datasets have been created for **Spanish** AMR, but they are relatively small in size for extensive deep learning experimentation or they are not publicly available (Fan and Gardent, 2020), which leads to little research in text-to-AMR for Spanish specifically except for the work that was already referred to above due to their data creation efforts and broader research on multilingual systems.

In the aforementioned work of Lee et al. (2022), **multilinguality** is achieved through the use of machine translation as a data augmentation technique. This is common practice in other research as well in an attempt to automatically create sizable AMR corpora. Mitreska et al. (2022), for instance, establish text-to-AMR and AMR-to-text pipelines for Macedonian, German, Italian, Spanish and Bulgarian. The AMR parsing and generation itself is tailored to English, but they then use machine translation to translate the input or output to the relevant language. Using machine translation to translate English sentences while keeping the same AMR to create synthetic AMR data for other languages has been introduced and proved effective since Damonte and Cohen (2018), who showed a significant boost of performance in their multilingual AMR parsing when using machine-translated data.

While prominent in its descriptive nature for linguistic purposes, AMR's increase of utility should also be mentioned. In the past year, AMR has been applied to NLP tasks ranging from machine translation (Song et al., 2019; Li and Flanigan, 2022) to the realm of multimodal research on the meaning and representation of gestures (Brutti et al., 2022) and images (Abdelsalam et al., 2022). The recent interdisciplinary endeavours underscore the broad exploration of AMR's applicability. However, the impediments of data scarcity across various languages and the absence of automated systems in non-English linguistic domains pose substantial barriers to the advancement of research in this field.

4. Methodology

4.1. Model

In this work, all our models are finetuned from the same base model mBART (Liu et al., 2020), specifically its checkpoint `mbart-large-cc25`, to ensure a fair comparison. Note that despite the base model being multilingual for all our models, in our methodology we often refer to our "monolingual" and "multilingual" models to indicate we finetuned them. This Transformer-based (Vaswani et al., 2017) encoder-decoder model was pretrained on the denoising objective of sequences (recovering an input text that had been scrambled, noised, deleted or otherwise modified) for 25 languages, including English, Spanish and Dutch. The data was resampled so that each language is equally represented in the training data of mBART. If we were to use different base models for each model, e.g. language-specific base models vs. multilingual base models, that would not be a fair comparison and it would not be clear whether the performance difference is caused by the amount of data or the quality of the base model. Therefore, for all of our models, we start from the same base model. Although we created Irish-Gaelic translations for other parts of our research, we did not include it in our model training. The reason is because mBART was not pretrained on Irish-Gaelic so the quality would not be fair compared to the other languages. We began working on this topic in mid-2022, but Heinecke and Shimorina (2022) demonstrated that the mT5 base model (Xue et al., 2021) is a suitable language model that covers Irish-Gaelic. We were not able to redo our work given computational and time constraints but hope that publishing our Irish-Gaelic data alongside the Spanish and Dutch variants enables other researchers to use the insights of Heinecke and Shimorina (2022) together with our data to create Irish-Gaelic AMR parsers.

Due to computational restrictions and because the base models are the same for all models (mBART), we did initial hyperparameter tuning for one model (`en+es+nl-part`) and its hyperparameters were then used to train other models as well. All models were trained for 25 epochs with early stopping.[5] In the remainder of this paper we will make use of our translated AMR 3.0 dataset for training and evaluation.

Because mBART is a sequence-to-sequence model, our input (text) and output (AMR) data has to be formatted as a sequence of tokens. The graphs in the datasets are therefore linearised and delinearised back into a PENMAN representation with a reimplementation of SPRING's (de)linearisation methods (Bevilacqua et al., 2021). They suggest to linearise a graph in a depth-first manner by slightly modifying the PENMAN representation. An example of this process is given in Figure 1c. As much content as possible is retained, such as opening and closing brackets, relations, and concepts. However, instead of variable names they add special tokens to the vocabulary, called pointer tokens. Instance relationships are made implicit by removing the forward slash (/). Concepts and relationships are also added to the vocabulary explicitly instead of rely-

[5] Exact hyperparameters will be given in an appendix in the camera-ready version.

ing on the model's subword tokenizer to ensure that the model learns about those tokens explicitly. When delinearising a model's prediction back into a graph, SPRING uses an iterative graph restoration method to fix potential issues if the predicted tokens could not be readily reconstructed into a graph, which they show works robustly.

4.2. Data

A valuable contribution of our work is the parallel, multilingual dataset that we provide for Spanish, Dutch and Irish Gaelic (Irish not used in this paper). We base our data collection methodology on the premise that "AMR annotations can be successfully shared across languages" (Damonte and Cohen, 2018, p. 1147). Unlike Damonte and Cohen (2020), who translated a relatively small portion of the AMR 2.0 corpus, we employ the more extensive AMR 3.0 corpus (Knight et al., 2020) and automatically translate *all* partitions (train, development, test) to make it usable for deep learning experiments. This corpus comprises 59,255 parallel AMR structures and English sentences, partitioned into canonical training (55,635), development (1,722), and test (1,898) sets. Unlike the domain-specific Europarl corpus used by Fan and Gardent (2020), AMR 3.0 spans a wider array of domains, including discussion forums, Wikipedia, news broadcasts, and literary works. For translation, we opted for Google Translate API v3, which was consulted on September 11th, 2023.[6] All 59,255 sentences were translated into Dutch, Spanish, and Irish Gaelic and manually corrected with regard to formal issues such as unexpected white-spaces or wrongly encoded characters. Similar to previous works mentioned above the AMR side remains unchanged for all languages - only the English source text was automatically translated. This process yields a large, parallel multilingual corpus with aligned AMR annotations that is sufficiently large and diverse in domain for multilingual machine learning experimentation. We make this dataset available with the same license as the original AMR 3.0 corpus on the LDC website under the name "AMR 3.0 - Dutch, Irish, and Spanish Machine Translations".[7]

As described before, the goal of this study is to gauge the performance of multilingual systems compared to their monolingual counter-parts, paying particular attention to the amount of data per-language that the model is trained on. To do so we train monolingual models for English, Spanish and Dutch as the baselines, where each model is trained on their respective full dataset of 55,635 training instances. We also train multilingual models, with English, Spanish and Dutch, and with only Spanish and Dutch. We are mostly interested in the multilingual models that were trained on a subset of the data so that the multilingual model has seen the same number of training samples in total as the monolingual models but distributed across languages. Furthermore, for reference, we also train multilingual models that are trained on the full dataset for each language to see how well multilingual models fare. This data distribution and corresponding models has been illustrated in Table 1. We thus control our training strictly on data size: the baseline models are trained on their full, monolingual dataset (`*-only`), the partial multilingual models (`*-part`) are only trained on a subset per language, and the full multilingual models (`*-full`) are trained on the full dataset of all languages combined. The hypothesis is that the baseline, monolingual models will perform better than the full multilingual models, which in turn will perform better than the partial multilingual models. A small difference would justify the compute efficient (one multilingual model) and data efficient (multilingual model trained on partial datasets) utility of multilingual AMR parsing.

lang. model	en	es	nl
en-only	55 635	0	0
es-only	0	55 635	0
nl-only	0	0	55 635
en+es+nl-part	18 545	18 545	18 545
es+nl-part	0	27 818	27 817
en+es+nl-full	55 635	55 635	55 635
es+nl-full	0	55 635	55 635

Table 1: Contents of the training set for each model. The row sections represent monolingual models (`*-only`), multilingual models that have been trained only on part of the data per-language (`*-part`), and multilingual models that were trained on the full dataset for each language (`*-full`).

4.3. Evaluation

For text-to-AMR parsing, it is common to use Smatch scores (Cai and Knight, 2013), which calculate the precision, recall and Smatch F1 scores on matching the triples of the predicted graph with the reference graph. We report Smatch F1 scores as calculated by `smatchpp` (Opitz, 2023), particularly its ILP solver rather than the hill climber ap-

[6] A sample of Dutch translations was manually verified for quality. For the other languages, we specifically selected Google Translate for its high-quality translations, corroborated by the report, "The State of Machine Translation, 2023" (https://inten.to/machine-translation-report-2023/).

[7] The data submission is accepted by the LDC and its release is planned for the second part of 2024.

proach for the best result. We report pairwise significance levels on the differences between systems based on this F1 score, by sorting systems best to worst and bootstrapping ($n = 1000$).[8] For brevity we show compact tables in the paper that only contain smatch f1 scores and, for the coarse-grained results, their significance compared to the lower performing systems. In addition, we discuss more fine-grained evaluation scores that are common in AMR research, following Damonte et al. (2017). The following categories are reported:

- Unlabeled: Smatch score without considering the edge labels (the relation between two items)
- No WSD: Smatch score without word sense disambiguation (`go` instead of `go-01`)
- Concepts: score of correctly predicting concepts
- Named entities: score of correctly predicting named entities (`:name`)
- Negations: score of correctly predicting negations and polarity (`:polarity`)
- Wiki: score of correctly predicting linked Wikipedia entries (`:wiki`)
- Reentrancy: some nodes can be reentering, for instance due to coreference (so they have more than one parent; like `b` in Fig. 1a).
- SRL: Smatch score for semantic role labelling, i.e., only considering ARGn relations to identify predicate-argument constructions

5. Results

In Tables 2, 4, and 6, we provide for all systems their smatch F1, precision and recall scores. All tables are sorted from worst to best according to the smatch F1 score. For each system the significance compared to only the previous system above it is given for conciseness reasons; other important significant differences as well as overlapping confidence intervals are described in the text. Bold fonts indicate best systems for a given metric. Note that we trained a multilingual model on Spanish and Dutch only to see whether leaving out English as a high-resource language would impact the results for the other languages. Therefore, the English results contain fewer systems than the other two languages.

Detailed scores on specific categories are given in Tables 3, 5, and 7, for English, Spanish and Dutch respectively. Here we report only the F1 scores (multiplied by 100). Digits after the decimal points are not reported for the fine-grained analysis due to the limited decimal precision in the fine-grained evaluation framework.

	smatch f1	smatch p	smatch r
en+es+nl-full	79.07	79.92	78.24
en+es+nl-part	80.14**	81.52	78.81
en-only	**81.30****	**82.34**	**80.29**

Significant differences with the previous row are marked as $^{**}p < 0.01$

Table 2: Smatch F1, precision and recall scores on the English test set

	en+es+nl-full	en+es+nl-part	en-only
unlabeled_f	82	83	**84**
no_wsd_f	79	81	**82**
concepts_f	85	87	**88**
ner_f	84	85	85
negations_f	63	66	**69**
wiki_f	74	74	**75**
reentrancies_f	68	69	**71**
srl_f	78	79	**80**

Table 3: Fine-grained evaluation results for the English test set (F1 score only)

	smatch f1	smatch p	smatch r
en+es+nl-part	73.04	74.59	71.56
es+nl-part	73.36*	74.76	72.79
es+nl-full	73.99	74.97	73.02
en+es+nl-full	74.10	74.99	73.24
es-only	**74.56**	**75.85**	**73.30**

Significant differences with the previous row are marked as $^{*}p < 0.05$

Table 4: Smatch F1, precision and recall scores on the Spanish test set

	en+es+nl	es+nl	es+nl	en+es+nl	es
	part	part	full	full	only
unlabeled_f	77	78	78	78	78
no_wsd_f	73	74	74	74	**75**
concepts_f	76	77	77	77	**78**
ner_f	83	83	83	83	**84**
negations_f	52	58	55	55	**59**
wiki_f	71	72	73	73	73
reentrancies_f	62	62	62	62	**63**
srl_f	70	71	71	71	**72**

Table 5: Fine-grained evaluation results for the Spanish test set (F1 score only)

	smatch f1	smatch p	smatch r
es+nl-part	73.09	74.15	72.07
en+es+nl-part	73.37	74.72	72.07
en+es+nl-full	73.45	74.24	72.66
es+nl-full	74.07	74.92	**73.24**
nl-only	**74.36**	**75.60**	73.15

No significant differences between successive rows

Table 6: Smatch F1, precision and recall scores on the Dutch test set

[8] https://github.com/mdtux89/amr-evaluation/

	es+nl part	en+es+nl part	en+es+nl full	es+nl full	nl only
unlabeled_f	77	77	77	78	78
no_wsd_f	73	73	73	74	74
concepts_f	76	76	76	77	**78**
ner_f	82	84	84	84	84
negations_f	51	54	52	54	**57**
wiki_f	72	72	73	73	73
reentrancies_f	60	61	61	62	62
srl_f	69	70	70	71	71

Table 7: Fine-grained evaluation results for the Dutch test set (F1 score only)

6. Discussion

The central hypothesis of this study posited that monolingual systems `en-only`, `es-only`, and `nl-only` would outperform multilingual systems in terms of F1 Smatch scores. The empirical data affirm this hypothesis, revealing a consistent pattern where monolingual models surpass their multilingual counterparts across all three languages investigated. However, a nuanced interpretation of the statistical significance tests offers some promising insights.

English For coarse-grained results on English (Table 2), the monolingual model was found to be significantly better than both multilingual models. In absolute terms, however, this difference is small: the difference between the multilingual model that was only trained on part of the dataset for each language, and the monolingual model is only 1.2 Smatch F1, and on top of that their confidence intervals overlap. Interestingly, training on the full datasets with all languages combined yields significantly worse performance. This seems to indicate that for English, training on more non-English data deteriorates performance. This is unexpected because the assumption is that training on more data as a whole should yield better results, but given the significant difference between `en+es+nl-full` and `en+es+nl-part` that is not the case for English, i.e., added languages to an English dataset make results significantly worse regardless of the size of the data.

Digging deeper in the English results in Table 3, we find that there is a relatively small increase in scores across categories for each model, with the exception of the "negations" category, where a larger differences can be noted between all models. Negation, or rather the "polarity" of an utterance, has been proven difficult for automatic AMR parsers in earlier work, so much so that it has been suggested to post-process the AMR graph with a heuristic algorithms to re-apply negation based on polarity words in the input (Zhang et al., 2019). Such methods can positively impact performance; however, in this study we are interested in the effect of different data distributions on training results without any other modifications. In terms of negation, we see that mixing in other languages has a strong, negative impact.

Spanish Looking at the main results for Spanish (Table 5), the story changes in some respect. The differences between the monolingual model on the one hand and the full multilingual model `en+es+nl-full` and partial `en+es+nl-part` on the other are significant, with a difference in score of only 0.5 and 1.6 respectively. It is clear that the difference between the Spanish monolingual model and the full multilingual model of 0.5 is small and their confidence intervals overlap greatly. This is in sharp contrast with English, where – even though there also was small overlap between confidence intervals – the difference in Smatch score was larger with 1.2. Unlike English as well we see that the multilingual model trained on partial datasets performs significantly worse than all other models, including the multilingual model trained on all data. So unlike for English, training on full datasets with a lot of data from different languages improves the result, which was expected because that means the model has "seen" more diverse Spanish data as a whole. Scrutinising the bilingual models that were trained on only Spanish and Dutch, we find that the performance between them does not differ significantly. In fact, neither of them differ significantly from the second best performing system, the multilingual model trained on the full datasets `en+es+nl-full`. This sentiment is compounded when looking at the monolingual, best model and the worst bilingual model that was trained only on part of the data. While these models differ significantly, the difference is 1.2 Smatch and their confidence intervals overlap. The bilingual model trained on full datasets does not differ significantly from the monolingual model. So for Spanish, multi/bilingual models trained on the full dataset are viable. Furthermore, while the differences between the partial models and the monolingual one are significant, their differences are relatively small (1.6 and 1.2), and the confidence interval of the Spanish-Dutch model overlap with the one of the monolingual model, which indicates that training on non-English languages together with a Germanic language (Dutch) in limited data availability still yields good results that may be sufficient under data and compute constraints.

In the fine-grained results (Table 5), we see the same tendencies as for English. For all categories there is a slight increase in scores across systems. In many cases scores are even identical across systems, such as for all but the worst system for the unlabeled category, which indicates that the models are all similarly good at predicting the structure of the AMR graph and that a difference in performance is therefore mostly linked to how well they can predict the relations between nodes. Noteworthy, again, is the large difference in how well negations can be predicted. The monolingual model greatly outperforms the multilingual models in this respect but also the bilingual model `es+nl-part` performs well compared to the others, indicating that training on balanced, partial datasets *without* English seems to work well.

Dutch In Dutch we observe similarities with Spanish (Table 6). Multi/bilingual models trained on only a portion of the data perform worse than the monolingual model but absolute differences are small as we hypothesised: the gap between the worst and best model is only 1.3 Smatch. Whereas for Spanish the monolingual model did not differ significantly from the full multi/bilingual models, the monolingual model does differ from `en+es+nl-full` significantly, but only with $p = 0.046$, an absolute difference of 0.9 Smatch F1, and overlapping confidence intervals. Going back to the main interest, the models trained on partial data sets, we find that while the `-part` models differ significantly from the monolingual model (as expected) this is only 1.3 and 1 Smatch F1 respectively and in both cases the confidence intervals overlap. This indicates that for Dutch, training on partial datasets, even combined with a Romance language, yields competitive results compared to a monolingual model.

Dutch fine-grained results are consistent with our earlier findings (Table 7). Performance across categories is similar across all systems with the exception of negations. There, the monolingual model is again greatly outperforming the other systems. However, whereas `es+nl-part` yielded good results in the negation category for Spanish, it performs poorly in this category for Dutch.

7. Conclusion

Our findings suggest that our hypothesis is partially confirmed. For non-English languages, multilingual and even bilingual models achieve good quality. The gap between the worst and best model is 2.2 Smatch F1 for English, but only 1.6 for Spanish and 1.3 for Dutch. If annotated data is scarce for a language, or computational resources are limited to train or deploy multiple language-specific models, it is viable to instead train a single multilingual model with a small trade-off in performance.

Interestingly and unexpectedly, for English, adding too much data of other languages deteriorates model performance. A potential explanation might be that AMR concepts correspond to an English lemmas and training on a mix of plenty of non-English and English data might “confuse” the model.

For all models and languages we confirm the findings of other researchers that polarity prediction is a hard task. We note that this category of errors alone seems to greatly impact performance across all models: in most of the fine-grained categories the performance difference between models is small but for “negations” it is fairly large. Therefore, using techniques such as the post-processing polarity algorithm by Zhang et al. (2019) could close the gap between multilingual and monolingual models even further.

By publishing our detailed findings, our models as baseline references, our multilingual dataset, and our training code, we hope to catalyse additional research in multilingual AMR parsing.

8. Limitations

Our work provides tangible language resources in the form of a multilingual AMR dataset and text-to-AMR models, and also offers insights into advantages and disadvantages of less-resource multilingual models. However, we also acknowledge limitations of our work.

To create our dataset, we make use of Google Translate, one of the best commercial MT systems available. However, we did not post-edit the translations or verified their translation in detail. Secondly, in our study we contrasted full monolingual models with partial and full multilingual models. In this study we did not include additional configurations, such as monolingual models with a subset of the dataset, or other data quantity variations. These were not feasible for us in terms of compute and time but could provide useful insights. Finally, we have based our methodology of training models mostly on the work of Bevilacqua et al. (2021). We have not made use of more recent work, nor used techniques such as multi-task learning, Bayesian learning or distillation. The impact of all those techniques on multilingual AMR parsing with machine-translated data could be promising.

By providing our models as a baseline alongside a multilingual dataset and training code, we aim to engage additional research that addresses these limitations that were out of scope for the current paper but that are noteworthy to investigate further.

Acknowledgements

This research was carried out as part of the Horizon 2020 “SignON” project, Grant Agreement No. 101017255. Model training and data processing was completed with the Flemish Supercomputer Center on grant “Extracting Meaning from Multilingual Text with Deep Neural Networks”.

9. Bibliographical References

Mohamed Ashraf Abdelsalam, Zhan Shi, Federico Fancellu, Kalliopi Basioti, Dhaivat J. Bhatt, vladimir pavlovic, and Afsaneh Fazly. 2022. Visual Semantic Parsing: From Images to Abstract Meaning Representation.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 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.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One SPRING to Rule Them Both: Symmetric AMR Semantic Parsing and Generation without a Complex Pipeline. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12564–12573.

Richard Brutti, Lucia Donatelli, Kenneth Lai, and James Pustejovsky. 2022. Abstract Meaning Representation for Gesture. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1576–1583, Marseille, France. European Language Resources Association.

Shu Cai and Kevin Knight. 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, Sofia, Bulgaria. Association for Computational Linguistics.

Ziming Cheng, Zuchao Li, and Hai Zhao. 2022. BiBL: AMR parsing and generation with bidirectional Bayesian learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5461–5475, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Marco Damonte and Shay B. Cohen. 2018. Cross-lingual Abstract Meaning Representation parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1146–1155, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Damonte, Shay B. Cohen, and Giorgio Satta. 2017. An incremental parser for Abstract Meaning Representation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 536–546, Valencia, Spain. Association for Computational Linguistics.

Angela Fan and Claire Gardent. 2020. Multilingual AMR-to-Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2889–2901, Online. Association for Computational Linguistics.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*, reprinted edition. Number 42 in Studies in Linguistics and Philosophy. Springer-Science+Business Media, B.V, Dordrecht.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Young-Suk Lee, Ramón Astudillo, Hoang Thanh Lam, Tahira Naseem, Radu Florian, and Salim Roukos. 2022. Maximum Bayes Smatch ensemble distillation for AMR parsing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5379–5392, Seattle, United States. Association for Computational Linguistics.

Young-Suk Lee, Ramón Fernandez Astudillo, Radu Florian, Tahira Naseem, and Salim Roukos. 2023. AMR Parsing with Instruction Fine-tuned Pre-trained Language Models. https://arxiv.org/abs/2304.12272v1.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of* *the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Changmao Li and Jeffrey Flanigan. 2022. Improving Neural Machine Translation with the Abstract Meaning Representation by Combining Graph and Sequence Transformers. In *Proceedings of the 2nd Workshop on Deep Learning on Graphs for Natural Language Processing (DLG4NLP 2022)*, pages 12–21, Seattle, Washington. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Christian M. I. M. Matthiessen and John A. Bateman. 1991. *Text Generation and Systemic-Functional Linguistics: Experiences from English and Japanese*. Communication in Artificial Intelligence. Pinter, London.

Maja Mitreska, Tashko Pavlov, Kostadin Mishev, and Monika Simjanoska. 2022. xAMR: Cross-lingual AMR End-to-End Pipeline:. In *Proceedings of the 3rd International Conference on Deep Learning Theory and Applications*, pages 132–139, Lisbon, Portugal. SCITEPRESS - Science and Technology Publications.

Juri Opitz. 2023. SMATCH++: Standardized and extended evaluation of semantic graphs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1595–1607, Dubrovnik, Croatia. Association for Computational Linguistics.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic Neural Machine Translation Using AMR. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS 2017*, pages 1–15, Long Beach, CA, USA.

Pavlo Vasylenko, Pere Lluís Huguet Cabot, Abelardo Carlos Martínez Lorenzo, and Roberto Navigli. 2023. Incorporating graph information in transformer-based AMR parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1995–2011, Toronto, Canada. Association for Computational Linguistics.

Chunliu Wang and Johan Bos. 2022. Comparing Neural Meaning-to-Text Approaches for Dutch. *Computational Linguistics in the Netherlands Journal*, 12:269–286.

Chunliu Wang, Huiyuan Lai, Malvina Nissim, and Johan Bos. 2023. Pre-Trained Language-Meaning Models for Multilingual Parsing and Generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5586–5600, Toronto, Canada. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4003–4012.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019. AMR Parsing as Sequence-to-Graph Transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 80–94, Florence, Italy. Association for Computational Linguistics.

Jiawei Zhou, Tahira Naseem, Ramón Fernandez Astudillo, Young-Suk Lee, Radu Florian, and Salim Roukos. 2021. Structure-aware finetuning of sequence-to-sequence transformers for transition-based AMR parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6290, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

10. Language Resource References

Marco Damonte and Shay Cohen. 2020. Abstract Meaning Representation 2.0 - Four Translations.

Johannes Heinecke and Anastasia Shimorina. 2022. Multilingual Abstract Meaning Representation for Celtic languages. In *Proceedings of* *the 4th Celtic Language Technology Workshop within LREC2022*, pages 1–6, Marseille, France. European Language Resources Association.

Kevin Knight, Bianca Badarau, Laura Banarescu, Claire Bonial, Madalina Bardocz, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, Tim O'Gorman, and Nathan Schneider. 2017. Abstract Meaning Representation (AMR) Annotation Release 2.0.

Kevin Knight, Bianca Badarau, Laura Baranescu, Claire Bonial, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Tim O'Gorman, Martha Palmer, Nathan Schneider, and Madalina Bardocz. 2020. Abstract Meaning Representation (AMR) Annotation Release 3.0.

Noelia Migueles-Abraira. 2017. A study towards Spanish abstract meaning representation. Master's thesis, University of the Basque Country, June.

Shira Wein, Lucia Donatelli, Ethan Ricker, Calvin Engstrom, Alex Nelson, Leonie Harter, and Nathan Schneider. 2022. Spanish Abstract Meaning Representation: Annotation of a general corpus. In *Northern European Journal of Language Technology, Volume 8*, Copenhagen, Denmark. Northern European Association of Language Technology.