Proceedings of the CLTW 4 @ LREC2022 , pages 1–6
Marseille, 20-25 June 2022

Multilingual Abstract Meaning Representation for Celtic Languages

Johannes Heinecke, Anastasia Shimorina
Orange Innovation
22300 Lannion, France
{johannes.heinecke,anastasia.shimorina}@orange.com

Abstract

Deep Semantic Parsing into Abstract Meaning Representation (AMR) graphs has reached a high quality with neural-based seq2seq approaches. However, the training corpus for AMR is only available for English. Several approaches to process other languages exist, but only for high resource languages. We 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. We finally show that machine translated test corpora unfairly improve the AMR evaluation for about 1 or 2 points (depending on the language).

Keywords: AMR, multilingual, low-resource languages, Celtic languages, Welsh

1. Introduction

Abstract Meaning Representation (AMR) is a representation language designed to provide data for natural language understanding, generation, and translation. It implements a simplified, standard neo-Davidsonian semantics (Davidson, 1967; Higginbotham, 1985); its formal origins are in unification systems (Kay, 1979) and other works in the 1980s and 90s. AMR has been formalised by Banarescu et al. (2013), and its motivation is to uniform and organize various semantic annotations like named entities, coreferences, word sense disambiguation, semantic relations, discourse connectives, temporal entities, etc. For verbal predicates, AMR makes extensive use of PropBank framesets as concepts where available (Kingsbury and Palmer, 2002; Palmer et al., 2005). If a concept is not defined in PropBank, English lemmas are used instead. AMR is heavily grounded onto English and is expressively not an interlingua of any kind, even though research work with AMR on languages other than English exists.

AMR graphs are directed, acyclic graphs where nodes are instances or concepts, and edges are relations. An example of an AMR graph is given in Figure 1. Currently AMR does not annotate number, tense or modality, in contrast to UMR (Van Gysel et al., 2021), which proposes to extend AMR in this sense.

Other formalisms to describe the semantics of sentences or texts are, for instance, Discourse Representation Theory (Kamp and Reyle, 1993; Kamp et al., 2011, DRT) and its derivates (Economical DRT, Segmented DRT), Universal Networking Language (Uchida et al., 1996, UNL, http://www.unlweb.net/unlweb/), Universal Conceptual Cognitive Annotation (Abend and Rappoport, 2013, UCCA), or Groningen Meaning Bank (Bos et al., 2017, GMB). However currently AMR seems to be the formalism with the largest interest[1].

```
(h / have-org-role-91                     # instance relation
    :ARG0 (c / city                         # edge relation
       :name (n / name
          :op1 "Cardiff"))                # attribute relation
    :ARG1 (c2 / country
       :name (n2 / name
         :op1 "Wales"))
    :ARG2 (c3 / capital))
```

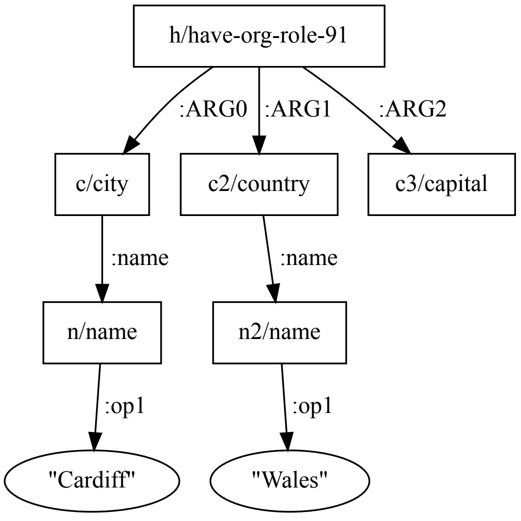

Figure 1: AMR graph (PENMAN format on top, graphical version below) for "Cardiff is the Welsh capital"; the red "/" is an *instance relation* which defines that a variable is an instance of a concept, in blue the *edge relations* which link instances and in green the *attribute relations* which link constants as strings or numbers to an instance. *have-org-role-91* is one out of a short list of special concepts which do not originate in PropBank, but are defined for AMR. Note that in the graphical version instance relations are not explicitely shown with an arrow and a label like $c \xrightarrow{:is\text{-}a} city$ but with a simple "/": *c / city*.

[1] For AMR in comparison to other formalisms see https://github.com/nschneid/amr-tutorial/raw/master/slides/AMR-TUTORIAL-FULL.pdf, pp. 115-121

The main AMR corpora of annotated data are available at Linguistic Data Consortium (LDC) for English[2]:

- LDC2020T02: LDC general release AMR 3.0 (2020), with 59,255 sentences;
- LDC2017T10: LDC general release AMR 2.0 (2017), with 39,260 sentences.

The sentences of the test corpus of AMR 2.0 were translated by human translators into four languages (LDC2020T07: AMR 2.0, four translations of AMR 2.0 test set into Italian, Spanish, German, Chinese, 1371 sentences per language).

However no translations are officially available for any of the Celtic languages. So we prepared translations into Welsh and Irish for the entire corpus (train/dev/test) using the Google Machine Translation (MT) API and had the 1371 sentences of the Welsh test corpus manually corrected and validated by a native speaker of Welsh[3]. Please note that in any case, the AMR graphs in the corpora do not change, "translation of the AMR corpus" means that the only the sentences themselves are translated into another language.

The remainder of this paper describes related work in multilingual parsing into AMR (Section 2) and our experiments (Section 3) on three Celtic languages: Welsh, Irish, and Scottish Gaelic.

2. Related Work

Although AMR had been conceived primarily for English, recently the interest to parse languages other than English into AMR has greatly increased. The approaches vary, and the results come close to the state-of-the-art results obtained for English AMR parsing. However, due to the absence of test data, all multilingual work is concentrated on the four languages for which human translated AMR test corpora exist, Chinese, German, Italian and Spanish; of which three are Indo-European languages, and Italian and Spanish are even more closely related Romance languages.

Currently Spring[4] (Bevilacqua et al., 2021) and X-AMR[5] (Cai et al., 2021), have the best results for English and the latter also for the four languages for which manual translations exist (cf. Table 1). Both Spring and X-AMR modify the AMR structures ("`<n> concept`" notation for variables instead of "`n / concept`" to distinguish variables from constants, since the former do not have semantics), optimize the AMR linearisation and add AMR relations to the underlying mBART tokenizer. Uhrig et al. (2021) chose to simply translate non-English sentences into English before calling an AMR parser (AMRlib[6]). Other approaches have been presented earlier by Damonte and Cohen (2018) (AMREager, using a transition-based parser) and by Blloshmi et al. (2020) (XL-AMR[7], a cross-lingual AMR parser which disposes of word aligners, i.e., word-to-word and word-to-node).

	de	it	es	zh
Damonte and Cohen (2018)	39.0	43.0	42.0	35.0
Blloshmi et al. (2020)	53.0	58.1	58.0	43.1
Uhrig et al. (2021)	67.6	72.3	70.7	59.1
Cai et al. (2021)	**73.1**	**75.4**	**75.9**	**61.9**

Table 1: Smatch scores for multilingual AMR parsing. Best scores in bold. All approaches are based on AMR 2.0

The performance of AMR parsers is evaluated by the *smatch score* which expresses the maximal score over all possible edge alignments (Cai and Knight, 2013)[8]:

$$P = \frac{\#\text{edges}_{\text{correct}}}{\#\text{edges}_{\text{gold}}} \quad R = \frac{\#\text{edges}_{\text{correct}}}{\#\text{edges}_{\text{system}}}$$

$$\text{smatch score (F1)} = \frac{2 \times \#\text{edges}_{\text{correct}}}{\#\text{edges}_{\text{gold}} + \#\text{edges}_{\text{system}}}$$

To calculate the smatch score, the optimal alignment of a gold AMR graph with a predicted AMR graph is to be found, which is a non-trivial task (Cai and Knight, 2013). Different runs of the evaluation can therefore produce slightly different results.

3. Experiments

3.1. General Multilingual Approach

Our approach to multilingual (and Celtic) AMR parsing draws from some of the approaches described in Section 2. As a parser we used a modified version of AMRlib[9], since the code for X-AMR (Cai et al., 2021) was not yet available in late 2021. The baseline was the original AMRlib with its model trained using the AMR 3.0 English corpus and T5 (Raffel et al., 2020) – a large pretrained language model. Expectedly all languages but English have very bad results (cf. first line of Table 2). In order to process other languages than English we first replaced the original T5 language model by the multilingual mT5[10] (Xue et al., 2021), retrained and tested the 4 human translated test corpora (LDC2020T07) on this mT5-based model (Table 2, 2nd line). This replacement shows gains in scores for all languages. In a next step we translated the train and development corpora into Chinese (zh), German (de), Italian (it) and Spanish (es) with MarianMT (Junczys-Dowmunt et al., 2018) and tested again on the 4 human translated test corpora. This time we observed a

[2] Other corpora are available at https://amr.isi.edu/download.html

[3] A great thank you to Delyth Prys, University of Bangor. Diolch yn fawr iawn i'r athro Delyth Prys, Canolfan Bedwyr, Prifysgol Cymru am ei help amhrisiadwy.

[4] https://github.com/SapienzaNLP/spring

[5] https://github.com/jcyk/XAMR

[6] https://github.com/bjascob/amrlib

[7] https://github.com/SapienzaNLP/xl-amr

[8] https://github.com/snowblink14/smatch

[9] https://github.com/bjascob/amrlib

[10] `google/mt5-base` model at HuggingFace.

large increase in Smatch score (Table 2, 3rd line). We then concatenated the English and the translated corpus for each language (both, for training and validation) and tested on the manually translated test sentences. Apart from Chinese we could not observe significant improvements (Table 2, lower four lines). These figures are very close to the SOTA results shown in Table 1. Please note that the evaluations in Table 1 is based on AMR 2.0, while our experiments are based on AMR 3.0. It is reported that AMR 3.0 results are in general slightly lower than AMR 2.0 (Bevilacqua et al., 2021).

trans-	training			test data		
former	data	en	de	es	it	zh
T5	en	81.1	56.5	49.7	45.8	10.7
mT5	en	**81.7**	58.9	62.4	59.7	54.9
mT5	*de/es/it/zh*		**71.1**	**74.4**	73.3	60.2
mT5	en + de	81.2	71.0			
mT5	en + es	81.5		74.3		
mT5	en + it	81.6			**73.9**	
mT5	en + zh	81.5				**61.1**

Table 2: Results (smatch scores) for training with English and translated corpora (MarianMT for train/dev, human translators for test), best scores in bold. *de/es/it/zh* means that the train and development corpora are in the same language as the test corpus. All training corpora are from AMR 3.0.

3.2. Celtic Languages

In this Section we describe our experiments for three Celtic languages: Welsh (cy), Irish (ga) and Scottish Gaelic (gd). Whereas the former is a P-Celtic language, the latter two are closely related Q-Celtic languages. Welsh has about 500,000 native speakers in Wales; Irish, even though the national language of Ireland, and Scottish Gaelic have much less native speakers. Except very young children all native speakers of these three languages are bilingual with English. All Celtic languages are under-resourced languages[11]. For written text, the Welsh Wikipedia, Welsh language press, official language production (Welsh Parliament[12]) provide text corpora of usable size, however linguistically annotated resources are quite limited. It is important to note that Welsh and Irish are amongst the 100 languages used to train mT5, whereas Scottish Gaelic is not included (neither are Breton, Manx and Cornish).

In order to obtain Welsh and Irish training and validation corpora, we used the Google Machine Translation API (the MarianMT models for Welsh[13] did not produce usable results). For Scottish Gaelic we only translated the test corpus. The next steps are identical to the experiments done for the four languages in Section 3.1. Again, we used models trained (on mT5) using the English training corpus, the Welsh/Irish corpus and the concatenated English and Welsh/Irish corpus (cf. Table 3).

trans-	training		test data	
former	data	cy	ga	gd
mT5	en	44.7	44.2	41.7
mT5	cy	73.4	39.9	36.2
mT5	en + cy	**74.3**	40.1	35.3
mT5	ga	39.7	**72.4**	**47.7**
mT5	en + ga	40.0	72.1	47.1

Table 3: Smatch scores for Celtic languages on models trained on English, Welsh, English and Welsh, Irish or English and Irish; best scores in bold.

At least for Welsh, the model trained on the combined data English and Welsh still improves the results, for Irish and Scottish Gaelic no improvement detectable. Using an Irish or Scottish Gaelic test corpus on a model trained on Welsh does not work (as was expected), whereas Scottish Gaelic improves slightly if a model trained on Irish is used (instead of English).

A simple error analysis showed that attribute relations (cf. Figure 1) in contrast to instance and edge relations are less likely to be incorrect. This means that named entities with different labels in other languages are nevertheless correctly rendered using the English label: The sentence *Mae Llundain yn brifddinas Lloegr* ("London is the capital of England") is parsed into the a graph, using the correct English labels "London" and "England" (cf. 2).

```
(h / have-org-role-91
    :ARG0 (c / city
        :name (n / name
            :op1 "London"))
    :ARG1 (c2 / country
        :name (n2 / name
            :op1 "England"))
    :ARG2 (c3 / capital))
```

Figure 2: AMR graph for *Mae Llundain yn brifddinas Lloegr* ("London is the capital of England")

The prediction of edge relations causes the drop in smatch score for all languages, including (the non-translated) English (cf. Table 4[14]).

3.3. The Effect of Machine Translation vs. Human Translation

Until now we have not yet addressed a weak point: for Welsh the entire corpus is machine-translated, includ-

[11] The Universal Dependency project (https://universaldependecies.org) provides treebanks for 5 Celtic languages, however their sizes are comparatively small.

[12] Cf. also the National Corpus of Contemporary Welsh (https://corcencc.org/), which provides a valuable source of written Welsh.

[13] https://huggingface.co/Helsinki-NLP/opus-mt-en-cy

[14] calculated using `smatch.py` (https://github.com/snowblink14/smatch)

	relation type			global
lang.	attribute	instance	edge	smatch score
en	90.5	87.2	73.7	81.7
de	86.5	71.9	68.7	71.0
es	84.1	77.4	71.7	74.3
it	85.3	75.9	71.6	73.9
zh	71.5	63.6	60.3	61.1
cy	83.5	76.7	71.7	74.3
ga	85.7	75.1	68.5	72.1
(gd	69.5	42.1	50.2	47.1)

Table 4: Global smatch scores and smatch scores for different relation types. Test corpora used were the human translations for Chinese, German, Italian and Spanish and machine translations (Google) for Welsh, Irish and Scottish Gaelic. Training was done using mT5 on the concatenated corpus (AMR 3.0) of English and the language concerned (except for Scottish Gaelic, where English and Irish was used instead).

ing the test corpus, whereas for Chinese, German, Italian and Spanish at least the test corpora were translated by human translators. Even though machine translation produces impressive results, it is not always perfect, especially for under-resourced languages like the Celtic languages. Our question is therefore: are the results (for Welsh AMR parsing, Table 3, third line) only as good as they are because the translation is bad and resembles more the source language (English) than proper Welsh? To test our hypothesis, we had the Welsh translation of the test corpus corrected and validated by native Welsh speakers. In parallel, we translated the test corpus from English into the four languages for which human translations exist (de, es, it, zh). For that, we used two MT systems: Google's MT API and MarianMT. We then parsed the translations and evaluated the result.

	de	*es*	*it*	*zh*	*cy*
human tr.	71.0	74.3	73.9	61.1	74.2
MarianMT	**74.8**	**76.2**	75.2	**68.5**	n/a
Google MT	74.0	76.1	**75.6**	68.2	**74.3**
mean diff.	3.40	1.85	1.5	7.25	0.1

Table 5: Comparison of smatch scores with translations (AMR models trained on English + *language*). Welsh was translated with Google MT only because MarianMT did not work well for this language.

Table 5 shows that the machine translated test corpora get a higher smatch score than the human translated ones. This confirms our hypothesis that translations using MT give higher scores due to their possibly greater similarity to English than human translations.

The difference in smatch score between the used MT systems is neglectable, even though for several MT metrics the Google MT API achieves higher values than MarianMT (table 6[15]). Table 6 also shows that there is an inverse correlation between the quality of the translation (with respect to the human translation) and the smatch score of the AMR evaluation: the better the MT evaluation with respect to the human translation, the worse the AMR smatch score. E.g., for German and Spanish all MT metrics show the preference to Google, meaning that its translations are closer to the human references, and Table 5 shows that the parsing of Google translations had a lower smatch score. The AMR parsing of human translations results in a even lower smatch score.

metric	*MT*	*de*	*es*	*it*	*zh*	*cy*
BLEU	M	43.11	59.70	49.82	33.89	n/a
	G	**50.70**	**65.29**	**53.16**	**43.84**	91.89
TER	M	45.12	26.87	36.05	**149.52**	n/a
	G	**38.70**	**22.67**	**32.70**	190.34	5.41
BERTsc.	M	73.77	84.24	78.17	63.20	n/a
	G	**78.31**	**86.31**	**81.10**	**70.51**	98.60
chrF++	M	66.98	78.33	71.55	n/a	n/a
	G	**71.43**	**81.57**	**73.68**	30.10	95.52
BARTsc.	M	-5.53	-5.31	-5.57	-6.92	n/a
	G	**-5.31**	**-5.13**	**-5.44**	**-6.53**	-4.21
	hum.	-3.47	-3.69	-3.65	-3.84	-3.65
=	M	7.5%	11.5%	6.9%	1.8%	n/a
	G	**9.9%**	**13.2%**	**8.1%**	**3.8%**	66.0%
LD (av.)	M	49.07	26.69	35.6	22.62	n/a
	G	**42.74**	**22.99**	**33.08**	**19.36**	12.58
LD (med)	M	39.0	20.0	29.0	18.0	n/a
	G	**35.0**	**18.0**	**27.0**	**15.0**	9.0

Table 6: Comparison of the machine translated test corpora (M: MarianMT, G: Google) with the human translated version, best score for each metric in bold. The BLEU score for the translation of Chinese has been calculated using the *zh*-tokenizer provided by sacreBLEU. Since MarianMT does not output any tokenization for Chinese, the character-based chrF++ metric is not applicable. For BARTscore we added a value for comparing two identical files (human translation: hum.) which is not 0, to have a base value to judge the other BARTscore values better. "=" indicates the percentage of sentences where MT and human translations are identical, "LD" is the average and mean Levenshtein-Damerau distance (Levenshtein, 1966). For TER and Levenshtein 0 is the best score; for BARTscore 0 is the best theoretical value too, but in reality even identical sentences have BARTscores below 0. All other metrics have 100 as best score.

Note that for Welsh, the difference between the human translation and the machine translation is minimal

[15] We use the following tools to calculate the MT metrics: BERTScore (Zhang et al., 2020): https://github.com/Tiiiger/bert_score, BLEU (Papineni et al., 2002; Post, 2018) and TER (Snover et al., 2006): https://github.com/mjpost/sacrebleu, BARTScore (Yuan et al., 2021): https://github.com/neulab/BARTScore and chrF++ (Popović, 2017): https://github.com/m-popovic/chrF

(BLEU 91.89). This may be due to the fact that the Welsh human translated test corpus had been in fact translated from English with MT and then manually corrected and not translated from scratch by a human translator. This is confirmed by the very good values for Welsh in Table 6, and the fact that in 66% of the Welsh sentences, MT and human correction do not differ at all.

4. Conclusion and Perspectives

We showed in this paper that thanks to machine translation and the fact that Welsh and Irish are present in modern multilingual pretrained language models like mT5, it is sufficient to train a model for an AMR parser which produced state-of-the-art results, comparable to AMR parsers for Spanish, Italian, German. A manual correction of the training corpora might improve these figures slightly, however, correcting up to 60,000 machine translated Welsh and Irish sentences would require many resources and is probably not necessary any more. This approach is not restricted to Welsh or Celtic languages. As long as the AMR training corpus can be (machine) translated into any language which in turn is also supported by the underlying language model (mT5), our approach should work for any language.

Even though AMR has been presented in 2013 (Banarescu et al., 2013), due to the lack of tools able to parse (English) sentences into AMR graphs, AMR was not used largely in NLP until recently, with the implementation of Seq2Seq transformer-based tools. The quality obtained with these tools opens the path to many downstream applications based on a more formalized semantics like, multilingual information extraction, question-answering on knowledge bases etc., as the increasing number of papers around AMR shows[16].

5. Acknowledgements

We would like to thank Professor Delyth Prys from Canolfan Bedwyr, University of Bangor for her help in validating and, when necessary, correcting the Welsh translations of AMR's test corpus.

6. Bibliographical References

Abend, O. and Rappoport, A. (2013). Universal Conceptual Cognitive Annotation (UCCA). In *51th Annual Meeting of the Association for Computational Linguistics*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, 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. ACL.

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

Blloshmi, R., Tripodi, R., and Navigli, R. (2020). XL-AMR: Enabling Cross-Lingual AMR Parsing with Transfer Learning Techniques. In *EMNLP*, page 2487–2500, Online. Association for Computational Linguistics.

Bos, J., Basile, V., Evang, K., Venhuizen, N., and Johannes, B. (2017). The Groningen Meaning Bank. In Nancy Ide et al., editors, *Handbook of Linguistic Annotation*, page 463–496. Springer, Berlin.

Cai, S. and Knight, K. (2013). Smatch: an Evaluation Metric for Semantic Feature Structures. In *ACL*, page 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Cai, D., Li, X., Chun-Sing Ho, J., Bing, L., and Lam, W. (2021). Multilingual AMR Parsing with Noisy Knowledge Distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2778–2789, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Damonte, M. and Cohen, S. B. (2018). Cross-lingual Abstract Meaning Representation Parsing. In *NAACL: Human Language Technologies*, pages 1146–1155, New Orleans, Lousiana, USA. Association for Computational Linguistics.

Davidson, D. (1967). The Logical Form of Action Sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press, Pittsburg.

Higginbotham, J. (1985). On semantics. *Linguistic inquiry*, 16(4):547–593.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Fikri Aji, A., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Studies in Linguistics and Philosophy 42. Kluwer, Dordrecht.

Kamp, H., Genabith, J. v., and Reyler, U. (2011). Discourse Representation Theory. In Dov M. Gabbay et al., editors, *Handbook of Philosophical Logic. Vol 15*. Springer, Heidelberg.

Kay, M. (1979). Functional grammar. *Annual Meeting of the Berkeley Linguistics Society*, 5:142–158.

Kingsbury, P. and Palmer, M. (2002). From TreeBank to PropBank. In *LREC*, Las Palmas, Canary Islands, Spain. European Language Resources Association.

Levenshtein, V. I. (1966). Binary codes capable of

[16] https://nert-nlp.github.io/AMR-Bibliography/

correction deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

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, Philadelphia, PA, USA. Association for Computational Linguistics.

Popović, M. (2017). chrF++: Words Helping Character n-grams. In *Proceedings of the Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Raffel, C., Shazeer, N., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and J., L. P. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21:1–67.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Uchida, H., Zhu, M., and Della Senta, T. (1996). UNL: Universal Networking Language. An electronic language for communication, understanding and collaboration. Technical report, Institude of Advanced Studies, United Nations University (IAS/UNU).

Uhrig, S., Rezepka García, Y., Opits, J., and Frank, A. (2021). Translate, then Parse! A strong baseline for Cross-Lingual AMR Parsing. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 58–64, Online. Association for Computational Linguistics.

Van Gysel, J. E. L., Vigus, M., Chun, J., Lai, K., Moeller, S., Yao, J., O'Gorman, T., Cowell, A., Croft, W., Huang, C., Hajič, J., Martin, J. H., Oepen, S., Palmer, M., Pustejovsky, J., Vallejos, R., and Xue, N. (2021). Designing a Uniform Meaning Representation for Natural Language Processing. *Künstliche Intelligenz*, 35:343–360.

Xue, L., Constant, N., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 483–498. Association for Computational Linguistics.

Yuan, W., Neubig, G., and Liu, P. (2021). BARTScore: Evaluating Generated Text as Text Generation. https://arxiv.org/abs/2106.11520.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.