# Question-parsing with Abstract Meaning Representation enhanced by adding small datasets

**Johannes Heinecke**[1], **Maria Boritchev**[2], **Frédéric Herledan**[1]
[1]Orange Innovation, 2 avenue Pierre Marzin, 22300 Lannion, France
[2]Paris Télécom, 19 place Marguerite Perey, 91120 Palaiseau, France
{johannes.heinecke,frederic.herledan}@orange.com
maria.boritchev@telecom-paris.fr

## Abstract

Abstract Meaning Representation (AMR) is a graph-based formalism for representing meaning in sentences. As the annotation is quite complex, few annotated corpora exist. The most well-known and widely-used corpora are LDC's AMR 3.0 and the datasets available on the new AMR website. Models trained on the LDC corpora work fine on texts with similar genre and style: sentences extracted from news articles, Wikipedia articles. However, other types of texts, in particular questions, are less well processed by models trained on this data. We analyse how adding few sentence-type specific annotations can steer the model to improve parsing in the case of questions in English.

## 1 Introduction

Abstract Meaning Representation (Banarescu et al., 2013) provides a framework to model the meaning of a sentence, notably actions, events or states and their participants. AMR relies heavily on (verbal) concepts defined in PropBank (Kingsbury and Palmer, 2002; Palmer et al., 2005), e.g. `bear-02` in figure 1, PropBank's sense `-02` for the verb "to bear". Instances are indicated by a following "`/`", e.g., `p` being an instance of the concept `person`. The names of the variables do not have any other semantics than being distinct. Relations are indicated by an initial colon (e.g. `:ARG1`, `:time`). Literals (strings and numbers) lack a preceding instance and "`/`" (c.f. *"Elizabeth"* and *1926* in the example in figure 1). This serialised format, shown in figure 1 left, is called PENMAN (Kasper, 1989).

The largest available corpus used to train models capable of parsing sentences from natural languages into AMR graphs, called AMR 3.0, LDC2020T02[1], is provided by the Linguistic Data Consortium (LDC). This corpus is composed of nearly 59 000 sentences and corresponding AMR graphs. The data contains discussions from forums (partly technical), news reels, translations to English of Chinese news broadcasts, along with a part originating from English Wikipedia pages and Aesop's fables (see LDC2020T02 documentation).

```
(b / bear-02
  :ARG1 (p / person
    :name (n / name
      :op1 "Queen"
      :op2 "Elizabeth"))
  :time (d / date-entity
    :year 1926))
```

Figure 1: AMR graph for "Queen Elizabeth was born in 1926" in PENMAN format.

The problem we address in this article is the following: the gold data currently available for AMR parsing is very homogeneous in form as it is composed of declarative, informative sentences. Training models on such data yields lower-than-expected results for parsing of questions in AMR. We add a small dataset of questions to the training data to bypass this problem. Even if we were intuitively expecting this kind of result, we were able to confirm it and measure improvement.

## 2 Related Work

Domain type adaptation research for AMR has been attempted in several contexts and perspectives, one of the most well-known leading to the development of Bio-AMR[2]. Bio-AMR includes texts from the biomedical domain, extracted from PubMed[3]. Vu et al. (2022) conducted a research

[1] Knight et al. (2020), https://catalog.ldc.upenn.edu/LDC2020T02

[2] Available on the new AMR webpage: https://github.com/flipz357/AMR-World

[3] https://pubmed.ncbi.nlm.nih.gov/

*Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 252–257
March 3-4, 2025 ©2025 University of Tartu Library

on AMR of data outside news article excerpts, focusing on the legal documents domain, using a gold dataset. The parsing results were not very conclusive, and the authors provide a detailed discussion of this result. Among the explanations for the models not-so-good performances, two stand out: first, legal documents contain mostly sentences longer than the ones from LDC datasets; then, the models faced out-of-vocabulary (OOV) issues, as some concepts, specific to the legal domain, were not defined in PropBank. This latter issue comes from the semantic difference between the news and the legal documents domains.

Lee et al. (2022) experimented on sentence-type adaptation through both algorithmic and data-based research. They created and released the QALD-9-AMR corpus, built on top of QALD-9 data (Usbeck et al., 2018). It contains AMR annotations for natural language questions in English, originally provided for executable semantic parsing. Lee et al. (2022) further mention one unavoidable difficulty for domain adaptation which is out-of-vocabulary named entities and their types, that cannot be solved without using domain-specific corpora. The authors compare the usage of silver data with that of human annotation for QALD-9.

## 3 AMR Parsing of Short Questions

In this section, we present our data and our parsing methods, followed by first observations and hypotheses.

**Corpora** The AMR 3.0 corpus mainly contains sentences from newspapers and a small part of Wikipedia. There is almost no real question in this corpus (apart from a few rhetorical ones). Our hypothesis is that a model trained on this data will not perform well on question parsing. Thus, our research question is to see whether it is possible to improve the model's performance on questions by adding a small corpus of short questions to its training data (i.e. AMR 3.0 train).

The data we used in this article is the following:

a. AMR 3.0, about 55 000 sentences for training, 1 722 sentences for validation and 1 898 sentences for test.

b. QALD-9 Lee et al. (2022)[4], contains 400 (train) and 150 (test) questions taken from the QALD-9 project and annotated using AMR. The test set of QALD-9 contains 13 sentences which are also in the train corpus and in the QALD-7 and QALD-8 data which served as input for QUEREO. We deleted them from the QALD-9 test set, and use only the 137 remaining sentences. (c.f. fig. 2).

c. QUEREO: a corpus we created, which contains 406 (training) short, quiz-like questions of the same type as the ones in QALD-9, coming amongst other sources from QALD-7, QALD-8. The 406 sentences are equally divided between questions and the corresponding answers.

About 25% of the questions and all answers in QUEREO were formulated prior to the AMR annotation by human annotators from our team[5] (cf. fig. 3 and 4). Table 1 details the size of the corpora. An answer can often be formulated in various ways: "Edinburgh is the capital of Scotland" and "the Scottish capital is Edinburgh", yielding very similar AMR graphs.

QUEREO was created by two annotators by correcting AMRlib's output annotations and checking PropBank concepts and associated arguments. The computation of pairwise Smatch scores shows a relatively good quality of annotation with an inter-annotator agreement of 87.37%. In case of disagreement, the best annotation was chosen manually by a third annotator.

| corpus | training | dev. | test |
|---|---|---|---|
| AMR 3.0 | 55 635 | 1 722 | 1 898 |
| QALD-9 AMR | 357 | 51 | 137 |
| QUEREO | 358 | 48 | 0 |

Table 1: Number of sentences in the used corpora. QALD-9 only comes with a train and a test set. We split 51 sentences from the training corpus in order to have a development set as well.

**Parser** We use a slightly modified version of the AMRlib[6] parser, which can use as an underlying language model models other than T5. In our case, we adapted AMRlib to use the multilingual version MT5 and FLAN-T5. The base data is the

[4] https://github.com/IBM/AMR-annotations

[5] The annotators used the official AMR annotation guidelines available at https://github.com/kevincrawfordknight/amr-guidelines/blob/master/amr.md. The AMR annotation was undertaken using metAMoRphosED (https://github.com/Orange-OpenSource/metamorphosed/) (Heinecke, 2023)

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

```
(e / erupt-01
    :ARG1 (v / volcano
        :mod (a / amr-unknown)
        :location (c / country
              :name (n / name
                  :op1 "Japan")))
    :time (s / since
          :op1 (d / date-entity
              :year 2000)))
```

“Which volcanos in Japan erupted since 2000?”

Figure 2: Example question from QALD-9 test corpus

```
(g / game
   :name (n / name
      :op1 "Winter"
      :op2 "Olympic"
      :op3 "Games")
   :time (d / date-entity
      :year 2010)
   :location (c / city
      :mod (a / amr-unknown)))
```

“In which city did the 2010 Winter Olympic Games take place?”

Figure 3: Example question from our corpus

AMR 3.0 corpus, which we augment with datasets containing short questions. We trained the models using either T5 (Raffel et al., 2020), FLAN-T5 (Chung et al., 2022) or MT5 (Xue et al., 2021) as underlying language model (base size in all four cases). For evaluation, we use the Smatch package (Cai and Knight, 2013)[7].

**Observations** We have noticed that models relying on the AMR 3.0 corpus perform less well in terms of Smatch F1 when it comes to questions and answer sentences, both in QALD-9 and our own data. Questions in QALD-9 are mostly short sentences, so generally a better performance would be expected. Table 2 shows these initial results on models trained by fine-tuning different language models.

Even though the results for QALD-9 are better than the ones for AMR3.0, we were expecting a larger difference in figures. The sentences in QALD-9 are much shorter compared to the ones from AMR3.0: 43.6 characters/sentence for QALD-9, 112.0 characters/sentence for AMR3.0.

**Hypotheses** This under-performance could be due to two factors: the slightly different syntax of questions with respect to declarative sentences (e.g. “to do” periphrasis in English or the “est-ce que” construction in French), or the missing coverage of vocabulary used in QALD-9 and our questions compared to the AMR 3.0 training corpus (for instance the concepts `abbreviate-01`, `skateboard-01` or `novelist`). Therefore, if the parser encounters an unknown concept, a solution is to use a fake concept appending “`-01`” to the concept’s name. However, we do not encounter this particular problem in our setting yet.

```
(g / game
   :name (n / name
        :op1 "Olympic"
        :op2 "Winter"
        :op3 "Games")
   :time (d / date-entity
        :year 2010)
   :location (c / city
        :name (n2 / name
            :op1 "Vancouver")))
```

“The 2010 Olympic Winter Games took place in Vancouver.”

Figure 4: Example answer from our corpus

| LM | AMR 3.0 | QALD-9 |
|---|---|---|
| T5 | 81.8 | 87.2 |
| FLAN T5 | 82.2 | 86.4 |
| MT5 (en-fr) | 81.6 | 85.7 |

Table 2: Results on the AMR3.0 test corpus and the QALD-9 test corpus. All models were trained on AMR3.0 train corpus only.

In the remainder of this paper we describe our AMR parsing based on our version of AMRlib, the additional data and the obtained results.

## 4 Effect of Adding Questions to the Training Data

After our first observations, we trained models using different combinations of augmented data.

**Experimental setup** In a first step we trained three models using the AMR 3.0 training corpus. This gives us our baseline results (table 2), for the AMR 3.0 test corpus and the QALD-9 test corpus.

We then extended the training data with the QALD-9 training data, with our data, and finally with both. The QALD-9 AMR comes in two files, a training and a test corpus. We took 51 sentences from the training corpus to have a development corpus (see table 1). We used QALD-9 AMR’s test corpus to test our model for the sake of reproducibility of our research results.

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

| LM | lg. | train data | test data | |
|---|---|---|---|---|
| | | | AMR 3.0 | QALD-9 |
| T5 | English | baseline | 81.8 | 87.2 |
| | | + QUEREO | *82.0 (+0.2)* | 86.8 (–0.4) |
| | | + QALD-9 | 81.9 (+0.1) | *90.0 (+2.8)* |
| | | + QR. + Q9 | *82.0 (+0.2)* | 89.5 (+2.3) |
| FLAN-T5 | English | baseline | 82.2 | 86.4 |
| | | + QUEREO | ***82.4 (+0.2)*** | 86.8 (+0.4) |
| | | + QALD-9 | 82.1 (–0.1) | ***89.7 (+3.3)*** |
| | | + QR. + Q9 | 82.1 (–0.1) | 89.6 (+3.2) |
| MT5 | En + Fr | baseline | 81.6 | 85.7 |
| | | + QUEREO | 81.4 (–0.2) | 86.6 (+0.9) |
| | | + QALD-9 | *81.8 (+0.2)* | ***89.8 (+4.1)*** |
| | | + QR. + Q9 | *81.8 (+0.2)* | 89.6 (+3.9) |

Table 3: Test results: Best figures for a test corpus with the same language model (T5, FLAN-T5, MT5) in italics, best overall score in bold. QR stands short for QUEREO, Q9 stands for QALD-9. The baseline is a corpus trained only on the AMR 3.0 training data, the difference with respect to the baseline is shown in small digits. The baseline is taken from table 2.

**Results** The results are shown in table 3. The baseline (already shown in table 2) is given by the models trained only on AMR 3.0 training data provided by LDC. Adding a little additional data to the AMR 3.0 training corpus we were able to improve significantly the parsing results, even for the AMR 3.0 test data. This is independent of the underlying language model.

## 5 Discussion

In this paper we showed that even minor additions to the standard AMR 3.0 training corpus can have big impacts on the performance of an AMR parser for a new sentence type, syntactic in the case of questions. Next, we plan on taking our studies further by annotating a domain specific corpus in the domain of artificial intelligence) or noisy data.

We are aware of the problems of Smatch-based evaluation, and we follow the other algorithms for AMR comparison that have been proposed, in particular semantic Smatch such as S2match (Opitz et al., 2020). In future work, we would like to broaden our exploration of the benefits of adding a small corpus of specialised examples through different dimension of AMR, using different types of evaluation metrics.

The exploration conducted in this paper has focused on one method of parsing, the one provided by our extended version of AMRlib. It would be interesting for us to test whether the data augmentation results presented here are coherent throughout the different parsing methods, in particular in using the most efficient parsing methods for AMR such as MBSE (see Lee et al. (2022)).

We would also like to conduct an exploration of errors similar to the one presented in Boritchev and Heinecke (2023) to be able to quantify and qualify the remaining percentages of mistakes. The goal then would be to use pre- and post-processing methods to accommodate these errors when possible.

We only worked with short questions, quiz-like, since we were not (yet) able to annotate corpora with longer questions or more complex types of questions. The questions in the additional corpora (QALD-9 and QUEREO) are given without a proper context. If we were to parse dialogues, coreference resolution and ellipsis resolution should be considered.

## 6 Further Work: Beyond English

The work presented in the current article only concerns English, since gold AMR data is only available for this language. Another problem is that the AMR3.0 training corpus is translated to other languages using machine translation, so errors in this translation may influence the results.

Even though AMR has explicitly not been developed to be an interlingua for multi-lingual processing, it is in fact used exactly for this. A manual translation of the AMR test corpus sentences into Chinese, German, Italian and Spanish is provided by LDC (LDC2020T07[8]). In order to annotate non-English text in AMR, two variants for multi-lingual AMR can be found in the literature: 1) annotating non-English sentences using a language specific set of concepts, i.e. instead of the (English) PropBank, concepts from language specific thesauri are used (e.g. Chinese AMR, (Li et al., 2016), Spanish (Migueles-Abraira et al., 2018), Turkish (Oral et al., 2022) amongst others) or 2) English AMR graphs represent the meaning of non-English sentences (Damonte and Cohen, 2018; Blloshmi et al., 2020; Uhrig et al., 2021; Cai et al., 2021; Heinecke and Shimorina, 2022). We followed the latter approach by machine-translating the sentences of the AMR 3.0

[8] Damonte and Cohen (2020)

corpus into French and training baseline models using this translation. We used Google Machine Translation (Wu et al., 2016) and No Language Left Behind (NLLB, Costa-jussà et al. (2022)).

For the training of the French corpus we only finetuned MT5. In addition we created a multilingual model (based on MT5) by concatenating and shuffling the English and French training and validation corpora. In this case we have an English and a French sentence for each AMR graph. The first results look very similar to the results described in this paper as shown in table 4 for French on QALD-9. Table 5 shows the results for French (similar to table 3 for English).

Another approach we would like to explore is the transition from AMR to Uniform Meaning Representation (UMR) (Bonn et al., 2024). As UMR is designed to be "cross-linguistically plausible", the multilanguage considerations are inherent to the UMR annotations, making them particularly interesting for our type of investigations.

| LM | AMR 3.0 | QALD-9 |
|---|---|---|
| MT5 (fr) | 74.8 | 81.4 |
| MT5 (en-fr) | 74.6 | 80.8 |

Table 4: French: Results of the AMR3.0 test corpus and the QALD-9 test corpus.

| LM | lg. | train data | test data | |
|---|---|---|---|---|
| | | | AMR 3.0 | QALD-9 |
| MT5 | French | baseline | 74.8 | 81.4 |
| | | + QUEREO | 74.8 (±0.0) | 82.3 (+0.9) |
| | | + QALD-9 | 74.7 (-0.1) | 84.4 (+3.0) |
| | | + QR. + Q9 | *75.0 (+0.2)* | *84.6 (+3.2)* |
| MT5 | En + Fr | baseline | 74.6 | 80.8 |
| | | + QUEREO | 74.5 (–0.1) | 81.6 (+0.8) |
| | | + QALD-9 | **74.9 (+0.3)** | **85.5 (+4.7)** |
| | | + QR. + Q9 | 74.9 (+0.3) | **85.5 (+4.7)** |

Table 5: French test results: Best figures for a test corpus with the same language model (MT5) in italics, best overall score in bold. QR stands short for QUEREO, Q9 stands for QALD-9.

## References

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.

Rexhina Blloshmi, Rocco Tripodi, and Roberto Navigli. 2020. XL-AMR: Enabling Cross-Lingual AMR Parsing with Transfer Learning Techniques. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 2487–2500, Online. Association for Computational Linguistics.

Julia Bonn, Matthew J. Buchholz, Jayeol Chun, Andrew Cowell, William Croft, Lukas Denk, Sijia Ge, Jan Hajič, Kenneth Lai, James H. Martin, Skatje Myers, Alexis Palmer, Martha Palmer, Claire Benet Post, James Pustejovsky, Kristine Stenzel, Haibo Sun, Zdeňka Urešová, Rosa Vallejos, Jens E. L. Van Gysel, Meagan Vigus, Nianwen Xue, and Jin Zhao. 2024. Building a broad infrastructure for uniform meaning representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2537–2547, Torino, Italia. ELRA and ICCL.

Maria Boritchev and Johannes Heinecke. 2023. Error exploration for automatic abstract meaning representation parsing. In *15th International Conference on Computational Semantics*.

Deng Cai, Xin Li, Jackie Chun-Sing Ho, Lidong Bing, and Wai Lam. 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.

Shu Cai and Kevin Knight. 2013. Smatch: an Evaluation Metric for Semantic Feature Structures. In *51st Annual Meeting of the Association for Computational Linguistics*, page 748–752, Sofia, Bulgaria. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Zoph Barret, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehgani, Siddhartha Brahma, Albert Webson, Shane Shixiang Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincen Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. https://arxiv.org/abs/2210.11416.

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

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

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

Johannes Heinecke. 2023. metAMoRphosED: a graphical editor for Abstract Meaning Representation. In *19th Joint ACL – ISO Workshop on Interoperable Semantic Annotation*, Nancy.

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. ELRA.

Robert T. Kasper. 1989. A flexible interface for linking applications to Penman's sentence generator. In *Speech and Natural Language: Proceedings of a Workshop Held at Philadelphia, Pennsylvania, February 21-23, 1989*.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1989–1993, Las Palmas, Canary Islands, Spain. European Language Resources Association.

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

Young-Suk Lee, Ramón Astudillo, Hoang Than Lam, Tahira Naseem, Florian Radu, and Salim Roukos. 2022. Maximum Bayes Smatch Ensemble Distillation for AMR Parsing. In *NAACL*, pages 5379–5392, Seattle, USA. Association for Computational Linguistics.

Bin Li, Yoan Wen, Bu Lijun, Weiguang Qu, and Nianwen Xue. 2016. Annotating the Little Prince with Chinese AMRs. In *Proceedings of LAW X – The 10th Linguistic Annotation Workshop*, pages 7–15, Berlin, Germany. Association for Computational Linguistics.

Noelia Migueles-Abraira, Rodrigo Agerri, and Arantza Diaz de Ilarraza. 2018. Annotating Abstract Meaning Representations for Spanish. In *The eleventh international conference on Language Resources and Evaluation*, pages 3074–3078, Marrakech, Maroc.

Juri Opitz, Anette Frank, and Letitia Parcalabescu. 2020. AMR similarity metrics from principles. *Transactions of the Association for Computational Linguistics*, 8(0):522–538.

Elif Oral, Ali Acar, and Gülşen Eryiğit. 2022. Abstract meaning representation of Turkish. *Natural Language Engineering*, pages 1–30.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Lee Katherine, Sharan Narang, Matena Michael, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21:1–67.

Sarah Uhrig, Yoalli Rezepka García, Juri Opitz, and Anette Frank. 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.

Ricardo Usbeck, Ria Hari Gusmita, Muhamad Saleem, and Axel-Cyrille Ngonga Ngomo. 2018. 9th challenge on question answering over linked data (qald-9). In *Joint proceedings of the 4th Workshop on Semantic Deep Learning (SemDeep-4) and NLIWoD4*.

Sinh Trong Vu, Minh Le Nguyen, and Ken Satoh. 2022. Abstract meaning representation for legal documents: an empirical research on a human-annotated dataset. *Artificial Intelligence and Law*, pages 1–23.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.

Linting Xue, Noa Constant, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *NAACL*, pages 483–498. Association for Computational Linguistics.