Pre-Trained Language-Meaning Models for Multilingual Parsing and Generation

Wang, Chunliu; Lai, Huiyuan; Nissim, Malvina; Bos, Johan

Published in:
Findings of the Association for Computational Linguistics

DOI:
10.18653/v1/2023.findings-acl.345

Publication date:
2023

Pre-Trained Language-Meaning Models for Multilingual Parsing and Generation

Chunliu Wang*, **Huiyuan Lai***, **Malvina Nissim**, **Johan Bos**
CLCG, University of Groningen / The Netherlands
{chunliu.wang, h.lai, m.nissim, johan.bos}@rug.nl

Abstract

Pre-trained language models (PLMs) have achieved great success in NLP and have recently been used for tasks in computational semantics. However, these tasks do not fully benefit from PLMs since meaning representations are not explicitly included in the pre-training stage. We introduce *multilingual pre-trained language-meaning models* based on Discourse Representation Structures (DRSs), including meaning representations besides natural language texts in the same model, and design a new strategy to reduce the gap between the pre-training and fine-tuning objectives. Since DRSs are language neutral, cross-lingual transfer learning is adopted to further improve the performance of non-English tasks. Automatic evaluation results show that our approach achieves the best performance on both the multilingual DRS parsing and DRS-to-text generation tasks. Correlation analysis between automatic metrics and human judgements on the generation task further validates the effectiveness of our model. Human inspection reveals that out-of-vocabulary tokens are the main cause of erroneous results.

1 Introduction

There are two common tasks in computational semantics: mapping a text to a meaning representation (semantic parsing), and its reverse, producing a text from a meaning representation (semantic generation). These tasks generally rely on corpora that contain texts aligned with meaning representations. While in recent years large pre-trained language models (PLMs), both monolingual as well as multilingual, have brought NLP tasks to a new level, semantic parsing and generation cannot fully benefit from them since the meaning representations are not included in PLMs explicitly.

Our goal in this work is to leverage the principle of pre-trained models and explore the benefit of multilingual semantic parsing and generation of including *in the same model* meaning representations aside from natural language. This would make it possible not only to operate multilingually, thanks to representation neutrality, but also to leverage the bidirectionality of language-meaning alignment. Figure 1 illustrates our idea.

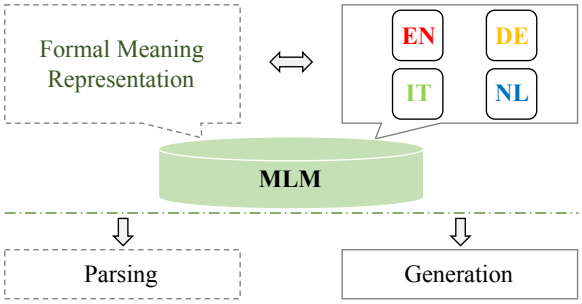


Figure 1: Our **M**ultilingual (English:EN, German:DE, Italian:IT, Dutch:NL) **L**anguage-**M**eaning framework (MLM) for parsing and generation.

Semantic parsing and generation (in different languages) are clearly related, but traditionally they are studied and developed independently of one another, usually focusing on a single language (often English). This results in having to train separate models from scratch for each task and language, and progress has been hampered by data scarcity. This is especially true for languages other than English, where data scarcity is even more severe.

Our proposal to incorporate meaning representations in PLMs and to concurrently embrace a multilingual approach breaks with this tradition yielding a twofold advantage. First, multilingual PLMs enable different languages to be represented in one universal space making it possible to benefit from cross-lingual knowledge transfer in semantic parsing and generation. Second, joining the formal and natural language representations in training makes it possible to leverage one and the same model for parsing and generation. For this approach to work, we need a meaning representation frame-

* Equal contribution.

Findings of the Association for Computational Linguistics: ACL 2023, pages 5586–5600
July 9-14, 2023

work where (i) the formalism is language-neutral, (ii) there is aligned data both in terms of meaning-language(s), but also multilingually across different languages, and (iii) there is enough expressivity to cover for a wide range of language phenomena.

Discourse Representation Structure (DRS), which satisfies our requirements well, is the formal meaning representation proposed in Discourse Representation Theory (DRT, Kamp 1981; Asher 1993; Kamp and Reyle 1993; Kadmon 2001; Kamp et al. 2011; Geurts et al. 2020). It covers a large variety of linguistic phenomena, including anaphors, presuppositions, temporal expressions and multi-sentence discourses and captures the semantics of negation, modals and quantification. Furthermore, DRS provides a language-neutral meaning representation: the same meaning representation associated with text that can be expressed in various languages. While Abstract Meaning Representations (AMR, Banarescu et al. 2013) have been proposed for this task, we believe DRS is more suitable because of its multi-lingual representation capability (all predicates are interpreted), its expressive power (proper treatment of negation and universal quantification), and the comparable annotated data available for multiple languages.

As a first step, we consider DRS as an additional abstract language that will complement the natural languages in our pre-trained model. We take the multilingual PLM mBART (Liu et al., 2020) and further pre-train it with all of our language data, thus both the four natural languages we use as well as the language neutral meaning representations, so that the DRSs and texts are learnt in the same semantic space. As a second step, we introduce a supervised denoising training that exploits more explicitly the relationship between DRS and each corresponding text as well as between the parallel texts in the different languages; we do this combined with denoising training to reduce the gap between the pre-training and fine-tuning objectives. At this point, we have at our disposal a single multilingual language-meaning model which can then be fine-tuned for either parsing (text-to-DRS) or generation (DRS-to-text), in a monolingual or multilingual fashion.

Overall, our main contributions include: (i) A novel task of multilingual DRS-to-text generation, and a framework for a mixed language-meaning modelling in a multilingual setting, serving both parsing and generation. (ii) A pre-training strategy, with self-supervised training followed by supervised training, to reduce the gap between pre-training and fine-tuning; we also employ multilingual transfer techniques to boost performance in languages other than English exploiting language neutrality in DRSs. (iii) Extensive experiments for both parsing and generation across different languages, including both automatic and human evaluation to understand how multilingual models perform.[1]

2 Background and Related Work

This work employs intensive multilingual pre-training techniques for language-meaning modelling for both parsing and generation. In this section, we briefly introduce the concept of DRS, which serves as our meaning representation tool, and relevant background and related work.

Discourse Representation Structures The Parallel Meaning Bank (PMB, Abzianidze et al. 2020) provides a large corpus of sentences annotated with DRSs in different formats for three different degrees of annotation quality: gold (completely checked manually), silver (partially checked manually) and bronze (uncorrected).[2] The box-format of DRS extensively used in Discourse Representation Theory may be convenient for human readability, but it is not suitable for modelling. We thus use the Discourse Representation Graph (DRG) format provided by the PMB and its equivalent variable-free sequential notation (Figure 2). There are three types of nodes in a DRG, a directed acyclic graph: conceptual entities (represented by WordNet (Fellbaum, 1998) synsets), constants (names, quantities, and the discourse deictics `speaker`, `hearer`, and `now`), contexts (defining scope as a box in DRT, represented graphically as a box). Edges between entity nodes denote thematic roles (Agent, Theme, Patient, Experiencer, Stimulus, Time, etc.) and comparison operators (=, $\neq$, $\prec$, $\leq$, $\sim$, and so on); edges between context nodes are discourse relations including negation (Figure 2).

Even though the PMB resorts to the English version of Wordnet (Fellbaum, 1998), we consider a synset as an interlingual way of representing a concept, being a compound of a lemma, part of speech (noun, verb, adjective, adverb) and sense number. This means that DRSs for languages other

[1] Code and models are available at https://github.com/wangchunliu/DRS-pretrained-LMM.

[2] See https://pmb.let.rug.nl/data.php.

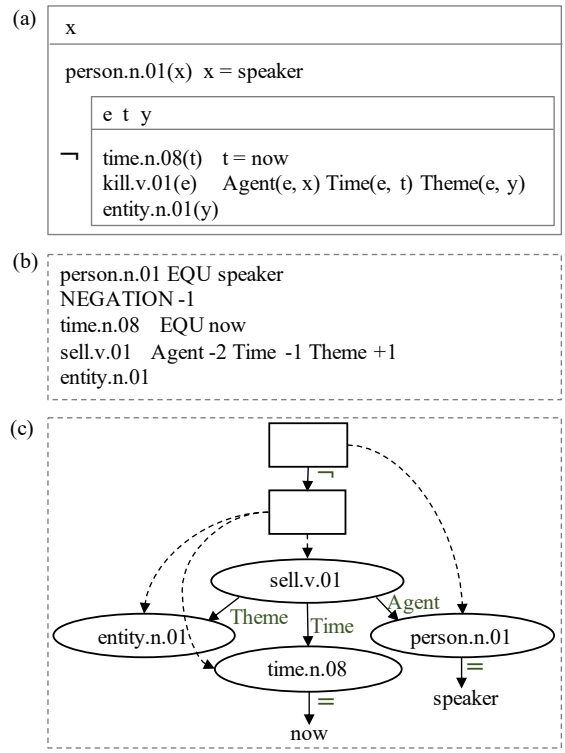


Figure 2: Example of different formats of DRS for the English sentence "I'm not selling anything": the box format (a), the variable-free sequence notation (b), and as directed acyclic graph (c).

than English also employ the synsets of the English WordNet as a sort of interlingua. Only names in a DRS are language-specific — for instance, the city of London would be represented in an Italian DRS as `city.n.01 Name "Londra"`.

The sequence notation for DRGs is based on a variable-free representation of DRS (Bos, 2021). In this notation a DRS is just a sequence of conceptual entities, roles with hooks (indices) or anchors, and discourse relations. Each entity is followed by the roles it introduces. Each thematic role or comparison operator either hooks to another entity via a negative or positive index (-1 relates to the previous entity in the sequence, -2 to the one before that, $+1$ to the next one, and so on). Discourse relations (e.g., NEGATION, NARRATION, ELABORATION) in the sequence notation introduce new contexts (see Figure 2). We make heavily use of this sequential notation because of the many advantages it offers. For example, compared with the box-format DRS, it can be easily converted into a graph structure without the complicated conversion process introduced in previous work (Fancellu et al., 2019; Fu et al., 2020). Compared with the clause-format DRS (van Noord et al., 2018), it omits the use of variables and is therefore simpler. It can also be used directly to train a sequence-to-sequence (seq2seq) neural model.

Text-to-DRS Parsing In the traditional efforts for DRS parsing, it can be roughly divided into two categories, namely rule-based and neural network-based methods. Regarding rule-based methods, Boxer (Bos, 2008) is a classic system based on rules and statistical methods. Recently, Poelman et al. (2022) propose a multilingual DRS parser leveraging existing off-the-shelf Universal Dependency parsers, it can achieve similar or even better performances than BERT-based models. Indeed, neural models have become the most popular methods in this field and usually achieve the best performance (van Noord et al., 2018; Liu et al., 2019b; Evang, 2019; van Noord et al., 2019, 2020a; Wang et al., 2021b). In addition to the seq2seq models above, there are two lines focusing on tree-based approaches (Liu et al., 2018, 2019a) and graph-based approaches (Fancellu et al., 2019; Fu et al., 2020), where Fancellu et al. (2019) is the first attempt at multilingual DRS parsing.

Most of the above works train neural models from scratch, and some make use of PLMs, but the models do not contain meaning representations explicitly during pre-training. Therefore, we aim to leverage the principle of pre-trained models and incorporate both meaning representations and natural language into one model. This, hopefully, can enable different languages to be represented explicitly in one universal space through pre-training, and result in one model for parsing and generation.

DRS-to-Text Generation Compared to DRS parsing, DRS-to-text generation has only recently drawn interest from NLP practitioners (Basile and Bos, 2011; Narayan and Gardent, 2014; Basile, 2015). Similar to DRS parsing, prior work on the generation task can be classified into rule-based methods (Basile and Bos, 2011) and neural network-based methods (Liu et al., 2021; Wang et al., 2021a). All these works focus on English only. Here, we take the first step towards a multilingual generation task and provide a corresponding benchmark, leveraging the representation neutrality in DRS and the bidirectionality of language-meaning alignment in different languages.

Multilingual Pre-Training In recent years, multilingual PLMs have brought NLP to a new era (Liu et al., 2020; Qiu et al., 2020; Xue et al., 2021).

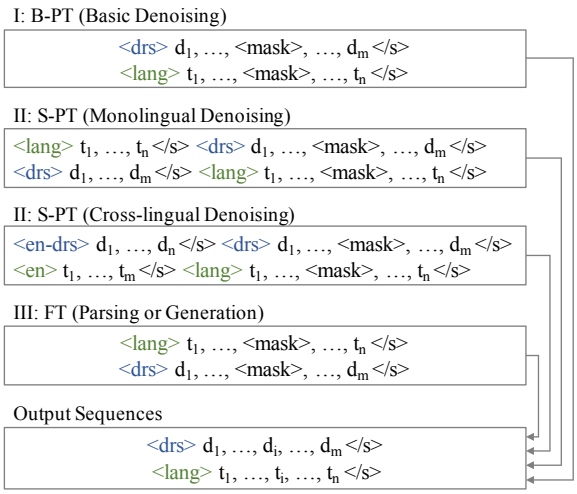


Figure 3: Pre-training and fine-tuning strategies for the language-meaning model. In the B-PT stage, the model is trained with basic denoising. The S-PT stage contains both monolingual and cross-lingual objectives.

They are pre-trained on large-scale unlabeled data in a self-supervised way, which enable different languages to be represented in one semantic space. Therefore, models fine-tuned on high-resource languages can thus transfer knowledge to other lower-resource languages for various tasks, such as Natural Language Inference (Conneau et al., 2018), Question Answering (Clark et al., 2020), Machine Translation (Liu et al., 2020), and formality transfer (Lai et al., 2022b).

Generally, PLMs are pre-trained in a self-supervised manner, which enforces models to reconstruct corrupted text based on denoising objectives (Liu et al., 2020). However, recent work shows that self-supervised pre-training may introduce noisy information that affects the performance of downstream tasks (Feng et al., 2022; Tang et al., 2022). Moreover, it has been shown that supervised pre-training can achieve superior performance compared to the self-supervised approaches (Conneau and Lample, 2019; Tang et al., 2022). In terms of computational semantics, Bai et al. (2022) propose a monolingual framework based on AMR, where the pre-training and fine-tuning share the same data format to facilitate knowledge transfer between them. Inspired by these works, we model meaning representations and natural language jointly leveraging the principle of PLMs in a multilingual fashion, and propose a pre-training strategy to make the pre-training objectives close to target downstream tasks by exploiting the relationship between DRS and its corresponding texts in different languages.

3 Method

We use mBART as our backbone to jointly model natural language and meaning representation in a multilingual manner, thereby enabling the DRS representations and the texts to be learnt in the same semantic space. This one model is then fine-tuned for parsing and generation.

3.1 mBART

mBART is a pre-trained denoising seq2seq model based on the Transformer architecture (Vaswani et al., 2017), derived from the monolingual model BART (Lewis et al., 2020). It is pre-trained to reconstruct the original text from a corrupted version (e.g. token masking). The model then takes the original sequence as input and maps it into the target sequence during fine-tuning and inference on downstream tasks. The novelty of our approach relies on the fact that the sequential DRS format allows for both text-to-DRS parsing and DRS-to-text generation to be performed in a seq2seq way (see Figure 4). For more efficient training, we filter out the unused tokens from mBART's vocabulary after tokenizing the training corpora (including texts and DRSs), which results in a shared vocabulary of 39,981 tokens. Besides, we add a special token `<drs>` as a prefix for DRSs, which is used to distinguish DRSs from natural languages and guide models to produce DRSs as outputs of parsing.

3.2 Multilingual Language-Meaning Models

We introduce a pre-training strategy to model natural language and meaning representation on top of mBART, including (i) basic denoising training and (ii) supervised denoising training.

Basic Denoising Training Since the meaning representations are not included in vanilla mBART, we perform a further pre-training to incorporate DRSs into the model and learn the universal representation. Specifically, we combine all the training data of multiple languages: $\boldsymbol{D} = \{D_1, ..., D_n\}$ where each D_i is a collection of data in a language. Language code `<lang>` and DRS code `<drs>` are used as prefixes for text and DRS sequences, respectively, to differentiate them from each other. As shown in Figure 3 (I: B-PT block), we follow Liu et al. (2020) to conduct a denoising training, which aims to reconstruct the original sequence from a version corrupted with a noise function. Formally, this denoising training can be

formulated as:

$$L_\theta = -\sum \log(T \mid g(T); \theta) \quad (1)$$

where θ are the parameters of mBART and g is the noise function that masks 35% of tokens in each sequence at random.

Supervised Denoising Training Although the basic denoising training makes the model learn the representations for text and DRS in a universal space, during this process the specific relationship between a given DRS and its corresponding texts is not learnt. There is thus a gap between the denoising pre-training and the fine-tuning for the text-to-DRS and DRS-to-text downstream tasks.

To bridge this gap, we perform a supervised denoising training using all parallel language-meaning pairs. This enables our model to learn the transformation connection between text and DRS after the first step of basic denoising training. As shown in Figure 3 (II: S-PT block), we concatenate the text sequences with the corresponding corrupted DRS sequences and conduct denoising training to reconstruct the original DRS in the text-to-DRS direction, and vice versa. Inspired by Wang et al. (2022), who show that retrieving and concatenating training instances relevant to the input can lead to significant gains on language generation tasks, we also perform an English-centric cross-lingual denoising training: English text (or DRS) sequences are concatenated with their corresponding corrupted non-English text (or DRS) sequences and then used for supervised denoising training (and vice versa).

3.3 Parsing and Generation

After denoising pre-training, the single model we have obtained can be fine-tuned with DRS-text pairs for the downstream DRS parsing and DRS-to-text generation tasks. As shown in Figure 3 (III: FT block), given a sequence $\boldsymbol{d} = \{d_1, \cdots, d_n\}$ of DRS and its corresponding text sequence $\boldsymbol{t} = \{t_1, \cdots, t_m\}$, taking DRS-to-text generation as an example, its seq2seq training can be formulated as follows:

$$p_\theta(\boldsymbol{t}|\boldsymbol{d}) = \prod_{i=1}^{m} p_\theta(t_i|t_{1,\ldots,i-1}; \boldsymbol{d}) \quad (2)$$

Similar to previous work (van Noord et al., 2020b; Wang et al., 2021b), we first train the model on gold + non-gold data, and then on gold + silver data.

Data type	Gold			Silver	Bronze
Lang	**Train**	**Dev**	**Test**	**Train**	**Train**
English	8,407	1,147	1,042	119,002	148,164
German	1,730	552	545	5,986	140,654
Italian	682	540	459	3,995	98,382
Dutch	535	435	490	1,363	26,433

Table 1: Documents statistics for PMB release 4.0.0.

Hyper-Parameter	B-PT	S-PT	F-FT	S-FT
Batch size	32	32	32	32
Update Steps	8	8	8	1
Max learning rate	1e-4	1e-5	5e-5	1e-5
Min learning rate	1e-5	1e-5	1e-5	1e-5
Warmup updates	3,000	0	3,000	0
Max decay steps	30,000	0	30,000	0

Table 2: Detailed hyper-parameters in our experiments.

In the first step (F-FT), we use the multilingual DRS-text pairs from dataset $\boldsymbol{D}$ since the same meaning representation can be expressed in various languages. We expect that this process can allow the model to further benefit from knowledge transfer across different languages. After that, the model can be finally fine-tuned on silver and gold data in either a multilingual or monolingual manner (S-FT).

4 Experiments

For all experiments we use PMB release 4.0.0, which contains texts in English, German, Dutch and Italian for three levels of annotation (gold, silver and bronze). Table 1 shows the statistics for the various languages, where each counted instance is a sentence and its corresponding DRS. (A small portion of DRSs that cannot be converted to DRGs were removed from the data set.)

4.1 Training Details

Table 2 reports the detailed hyper-parameters in our experiments. All experiments are implemented atop the Transformers library (Wolf et al., 2020). We use mBART-50 (Tang et al., 2020) as our base model, and train our models with batch size 32, accumulating gradients over 8 update steps in all training except for monolingual fine-tuning which is 1. We use Adam optimiser (Kingma and Ba, 2015) with a polynomial learning rate decay. Additionally, we apply early stopping (patience 5) if validation performance does not improve. Due to the small size of the Dutch dataset, we upsam-

ple them by replication obtaining training sets of 100,000 DRS-text pairs in both pre-training and multilingual fine-tuning.

4.2 Model Settings

To show the effects of each training stage in our framework, we conduct extensive experiments with different settings, yielding five different models. **M1** (FT mBART): fine-tuning vanilla mBART with monolingual data for each task; **M2** (M1 + B-PT): including basic denoising pre-training before monolingual fine-tuning; **M3** (M2 + S-PT): including supervised pre-training before monolingual fine-tuning, after basic pre-training; **M4** (M3 + F-FT): based on M3, and includes first multilingual fine-tuning (F-FT) before monolingual fine-tuning (S-FT); **M5** (monolithic model): based on M4, but using multilingual fine-tuning for S-FT and combining parsing and generation.

For comparison with our models, we also include two parsing systems from Poelman et al. (2022) which use the same DRS data format as we do: (i) UD-Boxer is a rule-based DRS parser based on Universal Dependencies; (ii) Neural Boxer is a seq2seq semantic parser based on Bi-LSTM with mBERT embeddings.

4.3 Automatic Evaluation

For **text-to-DRS parsing**, we follow recent work by Poelman et al. (2022) to convert the linearized DRS into Penman format (Kasper, 1989), as shown in Figure 4. We then adopt Smatch, a standard evaluation tool used in AMR parsing, to compute overlap between system output and gold standard by calculating the F-score of matching triples (Cai and Knight, 2013).

To assess **DRS-to-text generation**, we use three automatic metrics commonly used in text generation: n-gram-based BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007), as well as a neural-based COMET[3] (Rei et al., 2020).

4.4 Automatic Evaluation Results

Table 3 reports the results of DRS parsing in different languages. For English, the performances of the different models are pretty close to each other, with M2 outperforming the others with basic pre-training and monolingual fine-tuning. The models show higher scores for English compared to the other three languages, most likely because the

[3]We use model `wmt-large-da-estimator-1719`.

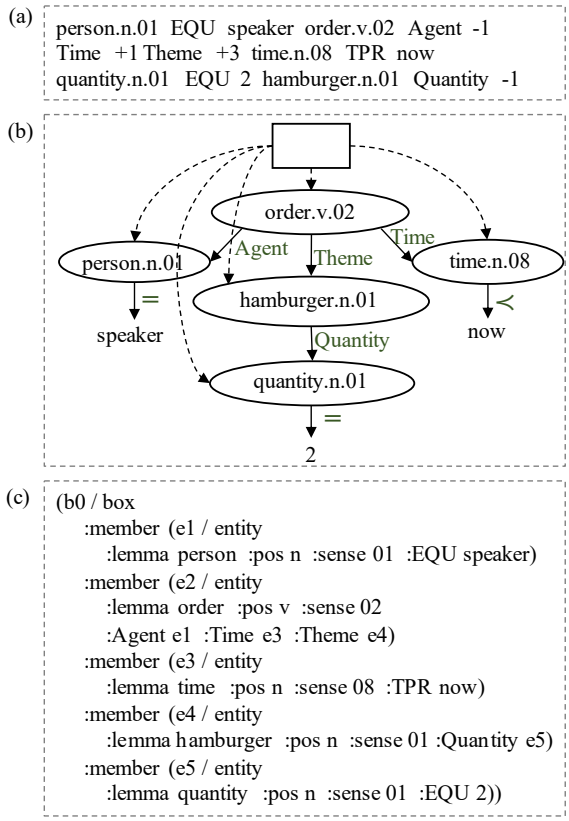


Figure 4: Example of DRS parsing evaluation procedure for sentence *I ordered two hamburgers*: linearized DRS data generated by parser (a), corresponding graphical DRG (b), and Penman format used for evaluation (c).

dataset contains a large amount of gold and silver DRS-text pairs in English, sufficient to fine-tune mBART for parsing without further pre-training.

When looking at the other three languages, we observe performance improvements with the use of different training strategies. Models pre-trained with the basic denoising task produce better results in German, the same F1-score in Italian, and lower results in Dutch, indicating a gap between pre-training and fine-tuning. This gap is bridged by our supervised pre-training strategy, models with the supervised pre-training (M3) yield steady improvements compared to M1 and M2. For M4 fine-tuned with multilingual data, they can further benefit from cross-lingual knowledge transfer and achieve higher scores. It is interesting to see that our monolithic model M5 performs best, thanks to the language-neutral meaning representation.

Compared to existing models UD-Boxer and Neural Boxer, all of our models, especially our main model (M5), achieve higher F1-scores across the board, showing significant improvements in four languages. Our models perform worse than

Model	EN		DE		IT		NL	
	F1	ERR	F1	ERR	F1	ERR	F1	ERR
M1: FT mBART	94.6	0.3	90.3	0.4	90.7	0.9	86.9	1.2
M2: M1 + B-PT	**94.7**	0.3	90.6	0.8	90.7	1.0	85.9	2.4
M3: M2 + S-PT	94.6	0.3	91.3	0.9	90.9	0.7	88.2	1.6
M4: M3 + F-FT	94.5	0.4	**92.0**	0.8	92.8	0.2	92.1	0.2
M5: monolithic model	94.0	0.2	**92.0**	0.4	**93.1**	0.2	**92.6**	0.6
UD-Boxer (Poelman et al., 2022)	81.8	**0.0**	77.5	**0.0**	79.1	**0.0**	75.8	**0.0**
Neural Boxer (Poelman et al., 2022)	92.5	2.3	74.7	0.5	75.4	**0.0**	71.6	1.0

Table 3: Evaluation results for text-to-DRS parsing on the test set of the four languages in the PMB 4.0.0. Notes: (i) ERR is the ill-formed rate (%) of generated DRSs that can not be transformed into a graph structure; (ii) bold numbers indicate best systems for each language.

Model	EN			DE			IT			NL		
	B	M	C	B	M	C	B	M	C	B	M	C
M1: FT mBART	**74.5**	54.7	102.8	45.1	35.1	54.3	44.3	34.4	58.2	34.9	29.5	31.3
M2: M1 + B-PT	73.2	54.0	101.5	45.0	34.8	56.8	44.2	34.2	59.7	38.6	31.8	44.4
M3: M2 + S-PT	74.2	54.6	102.4	52.1	38.4	65.3	49.3	36.6	72.6	47.8	38.6	59.9
M4: M3 + F-FT	**74.5**	54.8	102.4	**56.3**	**40.8**	**76.7**	**58.0**	**41.1**	**85.5**	**60.8**	**43.4**	**79.8**
M5: monolithic model	**74.5**	**55.0**	**102.9**	**56.3**	**40.8**	75.9	56.3	40.1	85.0	59.0	42.6	76.7

Table 4: Automatic evaluation results for DRS-to-text generation on the test sets of the four languages in the PMB 4.0.0 (B = BLEU; M = METEOR; C = COMET).

UD-Boxer in terms of ill-formedness rate, i.e., the proportion of generated DRSs which cannot be converted into a graph structure (and receive an F-score of 0). It is perhaps not surprising that rule-based parsers outperform neural-based parsers in generating well-formed DRS: the UD-Boxer parser is based on Universal Dependency and adds manual transformation rules to finally get the linearized data from the graph structure, and the evaluation process is equivalent to a reverse transformation process. It is worth noting that most of these errors can be corrected by post-processing (see §5.3). We also observe that our models have lower ERR rates than Neural-Boxer, except for Italian. The possible reason for this is that the multilingual training may introduce some noise.

For the generation task, we observe similar trends to parsing, as shown in Table 4. Concretely, Our proposed supervised denoising pre-training and multilingual fine-tuning strategies substantially boost the performances, especially non-English languages. Model M4 has the highest scores in all evaluation metrics across the three languages, the observation that differs slightly from that for the parsing task. We believe the reason is that the output tokens of the generation task are language-particular rather than language-neutral compared to the parsing task. Therefore, for the generation task, the results of fine-tuning with monolingual data are better than those with multilingual data.

Lang	BLEU	METEOR	COMET
EN	-0.098	-0.016	0.775
DE	0.275	0.471	0.687
IT	0.122	0.241	0.768
NL	0.195	0.386	0.686

Table 5: Sentence-level correlations of automatic metrics (against human reference) and human judgments for semantics. Underlined scores indicate $p < 0.01$.

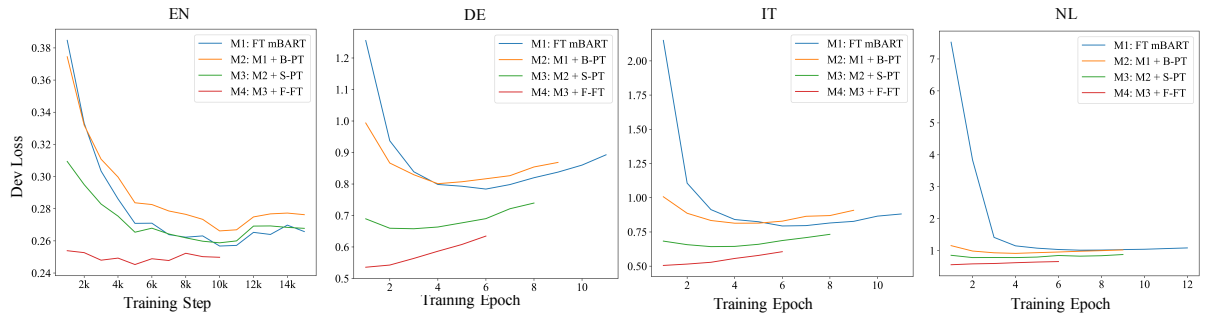
5 Analysis

5.1 Correlation Analysis

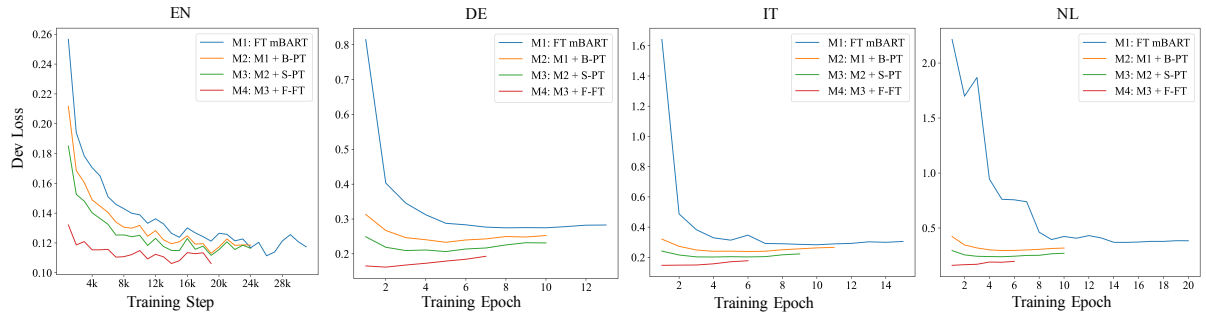
While human evaluation is seen as the most reliable assessment in language generation tasks, due to its costs and availability it can not be easily used during iterative development. We included human evaluation early on in our experiments to check the correlation of human judgement with automatic metrics, so that the latter could be more safely used in the following stages of our experiments.[4]

Table 5 shows the sentence-level biserial corre-

[4]See Appendix 8 for the details on human evaluation.



(a) DRS parsing.



(b) DRS-to-text generation.

Figure 5: Loss curves of monolingual fine-tuning on the development sets.

lations between automatic metrics and expert judgments in meaning preservation.[5] BLEU correlates particularly poorly with human judgments, even showing a negative correlation in English. METEOR also shows a negative correlation with human ratings in English, while it has higher scores than BLEU in non-English languages. Unsurprisingly, we see that COMET has high correlations with human judgements, which is consistent with previous work on other tasks (Rei et al., 2020; Lai et al., 2022a). This observation, therefore, confirms that COMET can be a more reliable metric used for DRS-to-text generation and for comparisons between different models.

5.2 Development Loss

To better understand the training strategies and components in our proposed framework, we examine the loss curves for different monolingual fine-tuned models on the dev sets of different languages (Figure 5).

For DRS parsing, the convergence process of the original mBART (M1) is slow. After adding different training strategies, models have a significantly faster and better convergence process. Specifically, we observe that basic denoising pre-training makes the model learn the representation for DRSs and texts in the same semantic space, but there is still a gap between the basic denoising task and the downstream task. This gap is then eliminated by supervised pre-training as the loss of model M3 is quite flat from the start and is lower than that of M2. Lastly, we see that multilingual fine-tuning consistently helps the model, and it eventually converges fast and well. This suggests that this strategy helps models benefit from the cross-knowledge transfer.

We observe similar trends for the DRS-to-text task, with a large fluctuation in the convergence process without pre-training. Overall, the loss curves for M4 are lower than other models.

5.3 Manual Inspection

In Table 6 we report example DRS outputs from our main model (M5) which differ from the gold standard. We summarize two types of ill-formed DRSs in linear format that cannot be converted to graph structures (and hence are not interpretable). When more tokens are produced than expected, the result is often a sequence of tokens that does not correspond to the graph. For instance, spaces

[5] Since the biserial correlation coefficient is a statistic used to assess the degree of relationship between an artificially created dichotomous nominal scale and an interval scale, it is naturally applicable to our experiments as the generated text is rated by annotators with 0 or 1.

Type	Subtype	Output Meaning	Gold Meaning
Ill-formed	Extra Space	geological_formation.n.01 Name " Himalayas"	geological_formation.n.01 Name "Himalayas"
		driving_ licence.n.01 Owner speaker	driving_licence.n.01 Owner speaker
	Missing Space	person.n.01 Role +1technician.n.01	person.n.01 Role +1 engineer.n.01
Meaning	Wrong Concept	overtreibe.v.01 Patient -1 Time +1	exaggerate.v.01 Agent -1 Time +1
	Wrong Role	blind.a.01 Experiencer -3 Time -2	blind.a.01 Theme -3 Time -2
	Wrong Index	female.n.02 Name "Maria" EQU +1 EQU now	female.n.02 Name "Maria"
	Missing Token	young.a.01 AttributeOf +1 person.n.01	young.a.01 Value + person.n.01 Attribute -1
	Extra Token	more_and_more.a.01 Degree +1 more.r.01	more_and_more.r.01

Table 6: Example outputs produced by our best model (M5) for the parsing task.

Reason	Lang	Generated Text	Gold text
Semantic	IT	Peter sta comprando un gatto male.	Peter sta comprando un gatto maschio.
Grammaticality	NL	Tom foldt zijn kleren.	Tom vouwt zijn kleren op.
Extra Material	EN	My flight arrived exactly at 2:30 p.m.	My flight arrived at 2:30 p.m.
Missing Material	EN	The express arrives at 6:30.	The express arrives at 6:30 p.m.
Word Choice	NL	Charles de Gaulle stierf in 1970.	Charles de Gaulle overleed in 1970.

Table 7: Example outputs generated by our best model (M4) for the generation task.

are included where they shouldn't be, or missing spaces cause subsequent tokens to be erroneously connected to each other.

These syntactic error types occur in a very limited number of models and can be well resolved by post-processing. We focus on the types of errors that affect meaning. We show five typical semantic error types at the bottom of Table 6 that affect the number of matching triples and may lead to a different meaning in gold data. For example, out-of-vocabulary (OOV) words may cause the parser to generate concepts different from gold yielding incorrect meanings. Also, incorrect roles lead to changes in meaning and wrong indices produce different predicate-argument structures. Another problem is when the parser fails to generate a crucial token. In contrast, the parser may hallucinate tokens, which may be added in unexpected places.

In Table 7, we show some examples of DRS-to-text generation which differ from the gold output for various reasons. The model might produce a word which does not convey the intended meaning. For example, in the IT example, the word "male" (EN: "bad") is generated in place of "maschio" (EN: "male"), probably due to the homography of the words across the two languages, without any semantic correspondence. Another example of non-matching is grammatical agreement, which can be due to some underspecified phenomena in DRSs. We also identify three more types: (1) the generated text has redundant information; (2) the generated data lacks some information; (3) the generated words are synonymous with those in the gold references. These types generally degrade automatic evaluation results but may not affect the performance of human evaluation. The generation of these cases is usually random and occurs in all models. Part of it is probably due to the OOV problem, and the rest is mainly related to the training data itself, because the same meaning representation can be paired with multiple expressions.

6 Conclusion and Future Work

Using DRS-based meaning representations as an additional language aside four different natural languages yields a novel multilingual pre-trained language-meaning model that can be fine-tuned for both semantic parsing and generation from formal meaning representations. By doing so, we achieve state-of-the-art performance on both tasks. Exploiting parallel data and DRS language neutrality is key to boost performance in lesser-resourced languages.

We believe our approach can benefit from improvements in its current form, but also opens up to further research in language-meaning models. Regarding future modelling directions, the contribution of graph structures should be further explored in future work. Specifically, it could be possibly to leverage the graph structure to mask tokens in a more meaningful and principled way, designing a denoising training using the rich linguistic phenomena expressed by DRSs.

Limitations

A large part of the dataset that we used in our experiments are semantic annotations for relatively short sentences (as the examples show). So we don't know really know how our multilingual pre-trained language-meaning modelling for DRS parsing and DRS-to-text generation will work on longer sentences.

In our experiments, we converted meaning representation in the sequence notation and modelled them with natural language texts in a seq2seq manner and masked tokens in the DRS sequence randomly. Perhaps a more natural way is to model DRSs as graph structures and let training objectives directly utilize any structural information from DRS. A graph structure would also eliminate the explicit order of concepts that is present in the sequence notation.

Although we say that the DRSs are language-neutral, the concepts in the vocabulary are based on the English WordNet. As a result it might be the case that non-English words do not have a direct correspondence to an appropriate synset, but the number of such cases is likely very small. The only (trivial) language dependence in DRSs are literal occurrences of proper names in cases where they differ across languages (e.g., "London", "Londen", or "Londra"). One way to remedy this is to add alternative spellings to the meaning representation to make it completely interlingual.

Acknowledgments

This work was funded by the NWO-VICI grant "Lost in Translation—Found in Meaning" (288-89-003) and the China Scholarship Council (CSC). We thank the anonymous reviewers of ACL 2023 for their insightful comments. We would also like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

References

Lasha Abzianidze, Rik Van Noord, Chunliu Wang, and Johan Bos. 2020. The parallel meaning bank: A framework for semantically annotating multiple languages. *AMIM*, 25(2):45–60.

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.

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)*, page todo, Online. 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.

Valerio Basile. 2015. *From logic to language: Natural language generation from logical forms*. Ph.D. thesis, University of Groningen.

Valerio Basile and Johan Bos. 2011. Towards generating text from discourse representation structures. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 145–150, Nancy, France. Association for Computational Linguistics.

Johan Bos. 2008. Wide-coverage semantic analysis with Boxer. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 277–286. College Publications.

Johan Bos. 2021. Variable-free discourse representation structures. *Semantics Archive*.

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.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Kilian Evang. 2019. Transition-based DRS parsing using stack-LSTMs. In *Proceedings of the IWCS Shared Task on Semantic Parsing*, Gothenburg, Sweden. Association for Computational Linguistics.

Federico Fancellu, Sorcha Gilroy, Adam Lopez, and Mirella Lapata. 2019. Semantic graph parsing with recurrent neural network DAG grammars. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2769–2778, Hong Kong, China. Association for Computational Linguistics.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. *The MIT Press, Cambridge, Ma., USA*.

Yutong Feng, Jianwen Jiang, Mingqian Tang, Rong Jin, and Yue Gao. 2022. Rethinking supervised pre-training for better downstream transferring. In *International Conference on Learning Representations*.

Qiankun Fu, Yue Zhang, Jiangming Liu, and Meishan Zhang. 2020. DRTS parsing with structure-aware encoding and decoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6818–6828, Online. Association for Computational Linguistics.

Bart Geurts, David I. Beaver, and Emar Maier. 2020. Discourse Representation Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, spring 2020 edition. Metaphysics Research Lab, Stanford University.

Nirit Kadmon. 2001. *Formal Pragmatics*. Blackwell.

H. Kamp. 1981. A theory of truth and semantic representation, 277-322, jag groenendijk, tmv janssen and mbj stokhof, eds. In Jeroen Groenendijk, editor, *Formal Methods in the Study of Language*. U of Amsterdam.

Hans Kamp and U. Reyle. 1993. From discourse to logic: Introduction to model theoretic semantics of natural language, formal logic and discourse representation theory. *Language*, 71(4).

Hans Kamp, Josef van Genabith, and Uwe Reyle. 2011. Discourse Representation Theory. In Dov M. Gabbay and Franz Guenthner, editors, *Handbook of Philosophical Logic*, volume 15, pages 125–394. Elsevier, MIT.

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*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Huiyuan Lai, Jiali Mao, Antonio Toral, and Malvina Nissim. 2022a. Human judgement as a compass to navigate automatic metrics for formality transfer. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 102–115, Dublin, Ireland. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2022b. Multilingual pre-training with language and task adaptation for multilingual text style transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 262–271, Dublin, Ireland. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics.

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.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2018. Discourse representation structure parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 429–439, Melbourne, Australia. Association for Computational Linguistics.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2019a. Discourse representation parsing for sentences and documents. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6248–6262, Florence, Italy. Association for Computational Linguistics.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2019b. Discourse representation structure parsing with recurrent neural networks and the transformer model. In *Proceedings of the IWCS Shared Task on Semantic Parsing*, Gothenburg, Sweden. Association for Computational Linguistics.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2021. Text generation from discourse representation structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 397–415, Online. 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.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 435–445, Baltimore, Maryland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meetings of the ACL*, pages 311–318.

Wessel Poelman, Rik van Noord, and Johan Bos. 2022. Transparent semantic parsing with Universal Dependencies using graph transformations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4186–4192, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, page 1872–1897.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2022. MVP: Multi-task supervised pre-training for natural language generation. *arXiv preprint, arXiv: 2206.12131v1*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint, arXiv: 2008.00401v1*.

Rik van Noord, Lasha Abzianidze, Antonio Toral, and Johan Bos. 2018. Exploring neural methods for parsing discourse representation structures. *Transactions of the Association for Computational Linguistics*, 6:619–633.

Rik van Noord, Antonio Toral, and Johan Bos. 2019. Linguistic information in neural semantic parsing with multiple encoders. In *Proceedings of the 13th International Conference on Computational Semantics - Short Papers*, pages 24–31, Gothenburg, Sweden. Association for Computational Linguistics.

Rik van Noord, Antonio Toral, and Johan Bos. 2020a. Character-level representations improve DRS-based semantic parsing even in the age of BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4587–4603, Online. Association for Computational Linguistics.

Rik van Noord, Antonio Toral, and Johan Bos. 2020b. Character-level representations improve DRS-based semantic parsing even in the age of BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4587–4603, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Chunliu Wang, Rik van Noord, Arianna Bisazza, and Johan Bos. 2021a. Evaluating text generation from discourse representation structures. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 73–83, Online. Association for Computational Linguistics.

Chunliu Wang, Rik van Noord, Arianna Bisazza, and Johan Bos. 2021b. Input representations for parsing discourse representation structures: Comparing English with Chinese. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 767–775, Online. Association for Computational Linguistics.

Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. 2022. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

A Appendix

A.1 Human Evaluation

Human evaluation was performed on a preliminary version of the models to assess correlation with the automatic metrics we planned to use on all larger-scale experiments. We adopt ROSE (Wang et al., 2021a), a human evaluation method that covers three dimensions: *semantics*, *grammaticality* and *phenomenon*, to assess the performance of models' outputs in the generation task. Since we are not investigating a particular linguistic phenomenon, we focus on the first two dimensions only: meaning preservation (whether the generated text has the same meaning as the gold text) and grammaticality (whether the generated text has no grammatical errors). We ask two experts with a linguistic doctorate degree to rate the generated texts with {0: No, 1: Yes} on these two dimensions. To reduce the annotation load, we exclude all outputs that are identical to the corresponding references, and then randomly select 100 samples for each language.

Evaluation Results Table 8 shows that about 26% of the generated sentences in languages other than English correspond to the references, while this rate is about 50% for English it reaches around 50%, due to the larger datasets. While the training data for German and Italian also far exceeds that of Dutch, the evaluation results are very close (including automatic evaluation) for these three languages, suggesting that the models do benefit from cross-lingual knowledge transfer.

Lang	Perfect	Semantics	Grammaticality	Overall
EN	49.3	87.0	90.0	83.0
DE	26.4	54.0	85.0	45.0
IT	27.2	51.0	70.0	38.0
NL	26.3	51.0	74.0	45.0

Table 8: Human evaluation results (%). `Perfect` indicates the ratio of generated sentences which correspond exactly to the human references; (ii) `Overall` indicates the ratio of cases rated 1 for both semantics and grammaticality.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
The section before References.

☒ A2. Did you discuss any potential risks of your work?
There is no potential risk in data, methods and analyses.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1 and section 6.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2 and section 4.

☑ B1. Did you cite the creators of artifacts you used?
Section 2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 2

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2 and section 4.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix A.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.1: just a single run.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.1 and 4.2.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 5.1 and section 5.3.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.