Transfer Learning Grammar for Multilingual Surface Realisation

***Abstract*—Deep learning approaches to surface realisation are often held back by lack of good quality datasets. These datasets require significant human effort to design and are rarely available for low-resource languages. We investigate the possibility of cross-lingual transfer learning of grammatical features in a multilingual text-to-text transformer. We train several mT5-small transformer models to generate grammatically correct sentences by reordering and inflecting words, first using monolingual data in one language and then in another language. We show that language comprehension and task-specific performance of the models benefit from pretraining on other languages with similar grammar rules, while languages with dissimilar grammar appear to disorient the model from its previous training. The results indicate that a model trained on multiple languages may familiarize itself with their common features and, thus, require less data and processing time for language-specific training. However, the experimental models are limited by lack of resources and an entirely text-to-text approach. A complete multilingual realisation model will, thus, require a larger transformer variant and longer training on more data.**

***Index Terms*—multilingual, surface realisation, universal grammar, transfer learning, transformer**

I. Introduction

Natural Language Generation (NLG) is the automated production of text in human languages. It encompasses everything from summarization and translation, generally referred to as text-to-text generation, to response and content creation, called data-to-text generation. NLG tasks have traditionally been divided into a number of subtasks, such as content determination, text planning, and realisation, to facilitate parallel and incremental development. Surface realisation is the final step in a traditional NLG pipeline and involves the formation of words into grammatically correct sentences [1].

Realisation has traditionally been considered a subfield of computational linguistics and as such, most classical systems employ hand-crafted or rule-based approaches. However, with the growing popularity of sequential deep learning models in NLG, realisation is increasingly being approached as a statistical problem [1]. Wen et al. (2015) and Tran et al. (2017) stacked multiple recurrent networks to jointly optimize sentence planning and realisation, with particular emphasis on generalizability to multiple domains [2], [3]. More recently, there has also been significant interest in multilingual realisation engines. Fan et al. (2020) studied a seq2seq model for the translation of English Abstract Meaning Representations to text in multiple languages [4]. Mille et al. (2020) organize an annual task aimed at the development of neural models for multilingual surface realisation [5].

Such data-driven approaches are often held back by the lack of good quality datasets. Most realisation datasets are formatted in special structures, like Abstract Meaning Representations and Universal Dependencies, designed by expert linguists to incorporate additional semantic and syntactic information [6], [7]. These datasets require significant human effort to design and are rarely available for low-resource languages.

In this paper, we investigate the possibility of cross-lingual transfer learning of grammar in multilingual surface realisation. The research is inspired by the controversial linguistic theory of Universal Grammar which states that all human languages share grammatical features that are also innate to humans [8]. Notwithstanding its status in the linguistics community, the applicability of this theory to a deep learning-based realisation engine is an interesting question since a model trained on multiple languages may familiarize itself with their common features and, thus, require less data and processing time for language-specific training. Such a multilingual model may overcome the lack of data for low-resource languages by generalizing across high-level and low-level languages.

We have chosen the mT5 transformer model for this study because this and multiple other transformer models have been shown to be capable of cross-lingual generalization [9], [10]. The mT5 model is also well-suited to this study because it is a text-to-text generation model, like the original T5, that has been pretrained on a large multilingual corpus. Therefore, it may be able to learn surface realisation directly from text, instead of hand-crafted data structures.

We test our hypothesis by training several mT5-small transformer models, first using monolingual data in one language and then in another language, on the Track 1 dataset from Multilingual Surface Realisation Shared Task that requires the generation of grammatically correct sentences by reordering and inflection of words. We evaluate the models using Language Modelling loss, BLEU-4, and inverse normalised string-edit distance to observe any potential syntactical generalization across the two languages. We show that language comprehension and task-specific performance of the models benefit from pretraining on other languages with similar grammar rules while languages with dissimilar grammar appear to disorient the model from its previous training.

In the remainder of this paper, we provide background information on the dataset and model used in this study (Section 2), followed by our methodology (Section 3). We then describe our experimental setup and discuss the results (Section 4). We conclude with a summary of the study and future work (Section 5).

II. Related Work

A. Multilingual Surface Realisation Shared Task

The Multilingual Surface Realisation Shared Task (MSRST) is an annual task aimed at the development of neural models for multilingual surface realisation. The task includes two tracks: a shallow track that requires the determination of word order and inflection, and a deep track that further requires the addition of missing function words and other syntactic features [11]. The task inputs include two datasets consisting of Universal Dependency structures derived from treebanks in 11 languages and modified by reordering, lemmatizing and, in the second track, removing most syntactic features [12].

The dataset for the first track is unrealistic as it contains a significant amount of syntactic information which would generally be unavailable to a preceding component in the generation pipeline [12]. The dataset for the second track contains mostly semantic information for only three languages but since UD is primarily a syntactic representation, it does not contain enough semantic information for a completely accurate realisation. However, the former contains sentences in 11 languages from 9 families formatted specifically to ensure cross-lingual syntactical uniformity [11]. By converting these structures into text, the dataset can be used to study cross-lingual syntax generalization in a text-to-text generation model. The results reported by participating systems can also be useful for a comparative evaluation of our results.

B. Transformers

Transformers are a family of sequence-to-sequence models based on attention mechanisms. First proposed by Vaswani et al. in 2017 [13], transformer models have achieved state-of-the-art results on many natural language processing tasks. Transformers have outperformed recurrent networks by harnessing the potential of transfer learning whereby models are pretrained on data-rich tasks and then fine-tuned on smaller datasets for tasks of interest [10]. Transfer learning has also been used in multilingual transformer models to improve the performance on low-resource languages. Several transformer models, such as mBERT [14], mBART [9], XLM-R [15] and mT5 [10], have also been shown to be capable of cross-lingual generalization. Therefore, it is reasonable to expect that a transformer model pretrained on multilingual input can also take advantage of transfer learning for syntactical generalisation across languages.

C. mT5

mT5 is the multilingual variant of the Text-to-Text Transfer Transformer (T5) that has been pretrained on data in 101 languages [10]. T5 is particularly distinctive for its text-to-text generation approach for all tasks, including classification tasks like semantic analysis where the model predicts the words 'positive' and 'negative' instead of class labels [10]. It has been trained on the masked language modelling objective where the model is required to denoise a corrupted input [10]. This objective is fairly similar to our task which involves the reordering and de-lemmatization of inputs.

The mT5 has been made available by Xue et al. (2020) in five model sizes. We have selected the mT5-small version (approx. 300M parameters) for this study as it requires the least amount of resources. Its configuration is similar to the original model proposed by Vaswani et al. (2017) and consists of similar encoder and decoder stacks with 8 layers, each with 6 attention heads. Since we are interested in the possibility of transfer learning of grammatical features across a small subset of the 101 languages, the small version is fairly adequate. However, for a proper multilingual surface realisation model, the larger versions may be more suitable.

III. Method

The purpose of this paper is to investigate the possibility of cross-lingual transfer learning of grammatical features using a text-to-text approach. To that end, we employ the following data processing and model training techniques.

A. Data Processing

The MSRST dataset consists of sentences in Universal Dependency structures. These structures are generally designed by expert linguists and contain additional syntactic information generally unavailable to a preceding component in the generation pipeline [12]. We extract the text of the source and target sentences from these structures and discard the additional information in accordance with our text-to-text generation approach.

Source	dark be . room the grow
Target	The room was growing dark.

Fig. 1: A sample input in English

The dataset has been created from several Universal Dependency tree-banks of varying domains and is divided into files according to the source. Therefore, we also consolidate the data for each language for joint training.

B. Model Training

The mT5 model has been pretrained to denoise copious amounts of multilingual data. We want to fine tune this model on the downstream task of word ordering and inflection while investigating cross-lingual generalization. To that end, we train several models using monolingual data in different languages to establish the baselines for comparison of future models. We then train the baseline models on monolingual data in a different language to observe any potential syntactical generalization across the two languages.

We study models trained on a variety of languages including high-resource (English) and low-resource languages (Arabic), languages from different families (French, Japanese), and languages with different scripts (English, Arabic, Japanese). The models are trained for 5 epochs on monolingual sentences in batches of 12 and training is optimized using Adam with a learning rate of 5e-5. The hyperparamters are left unchanged across all training runs to minimize any variance. We generate

a single output sentence with less than 100 tokens using beam search with a beam size of 5.

Table I describes the models studied in this paper. For example, mT5-en is the model trained on 19,976 sentences in English while mT5-en-fr is a version of this model extended by further training on 17,484 sentences in French.

Models	Dataset	Training Samples	Training Time
mT5-ar	Arabic	6075	25 min
mT5-en	English	19976	45 min
mT5-es	Spanish	28492	60 min
mT5-fr	French	17484	45 min
mT5-ja	Japanese	7133	30 min
mT5-en-fr	French	17484	75 min
mT5-es-ar	Arabic	6075	25 min
mT5-fr-en	English	19976	80 min
mT5-ja-ar	Arabic	6075	25 min
mT5-en-fr-ar	Arabic	6075	25 min

TABLE I: Description of trained models.

IV. Experiments

We evaluate the performance of the models using Language Modelling (LM) loss, smoothed BLEU-4 scores and inverse normalised character-based string-edit distance (DIST). LM loss allows us to observe the language comprehension of the models during training while BLEU and DIST are used to evaluate the generated outputs. Bilingual Evaluation Understudy (BLEU) is a precision metric that compares model outputs to reference sentences and scores them based on matching n-grams. It is generally used to evaluate the performance of text generation models, particularly translation engines. DIST scores the outputs according to the minimum number of edits required to convert them into reference sentences.

We use the implementations of BLEU and DIST, and the test dataset, provided by the organizers of the MSR Shared Task [11] to compare our results with participating systems. Since the test dataset is also divided into multiple files, we generate and evaluate the outputs for each file separately and report average results for each language. We compare the results with the best models submitted to the original MSR shared task (2018) [16] as our results calculated on detokenized outputs and are not directly comparable with the recent task results calculated on tokenized outputs [11]. The lack of tokenization also prevents any evaluation on Japanese as it does not use whitespaces.

A. Results and Discussion

The results of our experiments indicate that the language comprehension and task-specific performance of the models benefit from pretraining on other languages.

The LM loss calculated on validation data for baseline models shows that all monolingual models achieve roughly the same level of language comprehension irrespective of the number of training samples (Figure 2). Arabic is the only outlier because of its unique writing order (right-to-left). However, the additional training samples improve the generalization power of high-resource language models (like English), as apparent from BLEU and DIST evaluations.

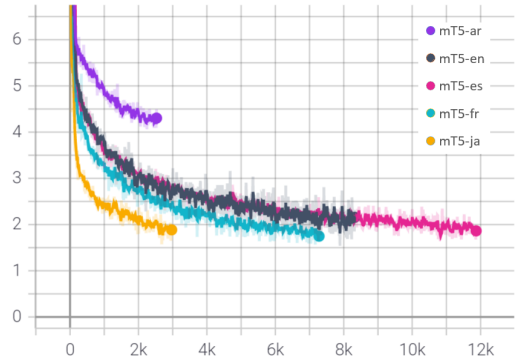

Fig. 2: LM loss for baseline models during training.

When extended by training on a second language, these monolingual models are able to learn the new language earlier and with better comprehension. Figure 3 compares the LM loss of the baseline Arabic model with the bilingual Spanish-Arabic and Japanese-Arabic, and the trilingual English-French-Arabic models. It is apparent that the models pretrained on other languages are able to understand Arabic better despite the differences between the languages. In fact, the only differentiating factor appears to be the total number of training samples as Japanese, with 7,133 sentences, has a higher LM loss than Spanish with 28,492 sentences, which is only slightly worse than English-French with 37,460 sentence.

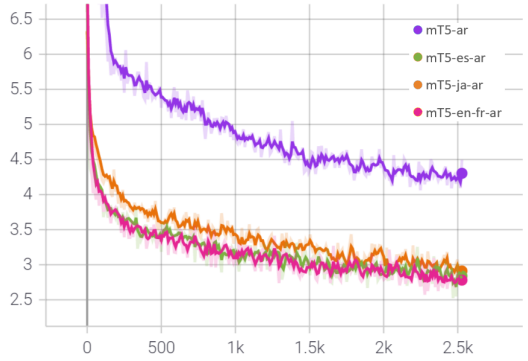

Fig. 3: LM loss for models when trained on Arabic.

The order of training also appears to be irrelevant as the decrease in LM loss when teaching French to an English model is similar to the loss when teaching English to a French model (Figure 4).

Tables II and III present the average BLEU and DIST scores of all models on each language. The results are in no way comparable to the performance of realisation engines submitted for the MSR task. This is primarily because the experimental models are based on the 'small' version of the

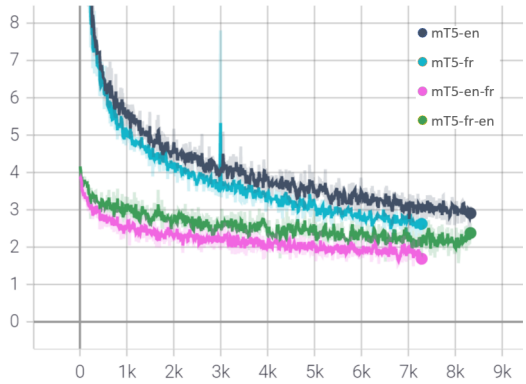

Fig. 4: LM loss for models when trained on English and French.

mT5 model and have also not been trained to completion as apparent from the slope of the validation loss plots. A complete realisation model should ideally be based on the standard or even a larger variant of the mT5 model and trained on all languages in the MSR dataset for tens of epochs. Since such a model would be significantly larger than any of the systems submitted for the MSR task, it involves a trade-off between computing resources and the human effort required for crafting of datasets.

The results show that most of the monolingual models have good zero-shot test scores on other languages. In fact, the French model performs better on English than French. This shows that the monolingual models are mostly dependant on the underlying pretraining of the mT5 transformer, biased towards high-resource languages like English, as their task-based training is inadequate. This also explains the poor performance on low-resource Arabic.

However, the results do show promising signs of cross-lingual generalisation. The multilingual models produce significant gains in performance because of higher combined training time and number of training samples. The results show that training a monolingual model on a new language boosts the performance on both languages. Indeed, teaching French to an English model and vice versa yields better BLEU and DIST scores than the monolingual models. Interestingly, teaching Arabic to a Spanish and an English-French model results in a significant drop in the performance of the models on the original languages. This lends some credibility to the idea that the model is indeed generalizing syntactical features since Arabic has a vastly different grammar than English, French and Spanish. With a completely different language family, script and even order of writing, Arabic appears to disorient the model from its previous training.

Figure 5 shows the sentences generated by the four English models and a French model. Although the English models produce syntactically correct sentences for the most part, the generated outputs contain frequent repetitions. As also apparent from the sentences generated by the French model, although the models seem to have learned to use the input

Models	Arabic	English	Spanish	French
MSRST	**25.65**	**69.14**	**65.31**	**52.03**
mT5-ar	0.02	0.01	0.0	0.01
mT5-en	0.0	7.40	0.99	1.56
mT5-es	0.01	4.44	**7.83**	2.82
mT5-fr	0.01	3.30	1.86	3.22
mT5-ja	0.02	0.47	0.26	0.22
mT5-en-fr	0.01	9.20	2.71	**7.86**
mT5-es-ar	**0.39**	2.26	2.39	0.87
mT5-fr-en	0.01	**13.93**	3.65	4.91
mT5-ja-ar	**0.39**	2.30	2.39	0.87
mT5-en-fr-ar	**0.39**	5.63	0.82	3.05

TABLE II: BLEU-4 scores on trained languages.

Models	Arabic	English	Spanish	French
MSRST	**46.49**	**80.42**	**61.46**	**55.54**
mT5-ar	10.44	3.35	2.14	3.44
mT5-en	5.14	34.40	26.0	28.91
mT5-es	11.66	31.48	**35.0**	33.70
mT5-fr	8.91	27.20	29.71	30.31
mT5-ja	11.37	20.50	22.50	22.42
mT5-en-fr	10.47	38.10	31.94	**37.32**
mT5-es-ar	24.69	31.70	30.90	29.83
mT5-fr-en	11.27	**41.38**	32.84	35.60
mT5-ja-ar	24.69	32.14	30.89	29.83
mT5-en-fr-ar	**25.83**	35.32	28.37	33.50

TABLE III: DIST scores on trained languages.

tokens, they are not yet confident to incorporate all of them. This suggests that the models can be improved by training for more epochs. The second sentence generated by the English-French-Arabic model shows the effects of Arabic training on the model's understanding of English grammar. This further consolidates the idea that the multilingual models do actually generalize the common grammatical features of their respective languages.

V. Conclusion

In this paper, we investigated the possibility of cross-lingual transfer learning of grammatical features in surface realisation using a text-to-text approach. We trained several mT5-small transformer models, first using monolingual data in one language and then in another language, to generate grammatically correct sentences by reordering and inflecting

Input	Google? into if morph GoogleOS what
Reference	What if Google Morphed into GoogleOS?
mT5-en	The GoogleOS morphs into GoogleOS?
mT5-en-fr	Does GoogleOS morph into GoogleOS?
mT5-fr-en	If GoogleOS morphs into GoogleOS?
mT5-en-fr-ar	Does GoogleOS morph into GoogleOS?
mT5-fr	Il morph into GoogleOS?
Input	observation make the of a on he some . few pic' good
Reference	He makes some good observations on a few of the pic's.
mT5-en	He made some pictures on a few pic's on a pic.
mT5-en-fr	They make a good observation on the pic' of a pic'.
mT5-fr-en	He made some good observations on the pic' of a good pic'.
mT5-en-fr-ar	They make a good observation on pic'.
mT5-fr	L' observation of a pic' on a pic'.

Fig. 5: Sample sentences produced by various models.

words. We evaluated the models using Language Modelling loss, BLEU-4, and inverse normalised string-edit distance in order to observe any potential syntactical generalization across the two languages. We found that language comprehension and task-specific performance of the models benefit from pretraining on other languages with similar grammar rules while languages with dissimilar grammar appear to disorient the model from the grammar of the originally trained language.

The results indicate that a model trained on multiple languages may familiarize itself with their common features and, thus, require less data and processing time for language-specific training. However, the experimental models, limited by lack of resources and an entirely text-to-text approach, are not comparable to proper realisation engines. A complete realisation model, based on a larger transformer variant and trained on more languages, is a possible avenue for future research.

Acknowledgments

We thank [REDACTED] for encouragement and guidance throughout this study.

References

[1] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.

[2] T.-H. Wen, M. Gasic, N. Mrksic, P.-H. Su, D. Vandyke, and S. Young, "Semantically conditioned lstm-based natural language generation for spoken dialogue systems," *arXiv preprint arXiv:1508.01745*, 2015.

[3] V.-K. Tran and L.-M. Nguyen, "Semantic refinement gru-based neural language generation for spoken dialogue systems," in *International Conference of the Pacific Association for Computational Linguistics*. Springer, 2017, pp. 63–75.

[4] A. Fan and C. Gardent, "Multilingual amr-to-text generation," *arXiv preprint arXiv:2011.05443*, 2020.

[5] S. Mille, A. Belz, B. Bohnet, T. C. Ferreira, Y. Graham, and L. Wanner, "The third multilingual surface realisation shared task (sr'20): Overview and evaluation results," in *Proceedings of the Third Workshop on Multilingual Surface Realisation*, 2020, pp. 1–20.

[6] J. Nivre, M.-C. De Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira *et al.*, "Universal dependencies v1: A multilingual treebank collection," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 1659–1666.

[7] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider, "Abstract meaning representation for sembanking," in *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, 2013, pp. 178–186.

[8] E. Dabrowska, "What exactly is universal grammar, and has anyone seen it?" *Frontiers in psychology*, vol. 6, p. 852, 2015.

[9] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[10] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020.

[11] S. Mille, A. Belz, B. Bohnet, Y. Graham, and L. Wanner, "The second multilingual surface realisation shared task (sr'19): Overview and evaluation results," in *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, 2019, pp. 1–17.

[12] S. Mille, A. Belz, B. Bohnet, and L. Wanner, "Underspecified universal dependency structures as inputs for multilingual surface realisation," in *Proceedings of the 11th International Conference on Natural Language Generation*, 2018, pp. 199–209.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[14] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" *arXiv preprint arXiv:1906.01502*, 2019.

[15] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.

[16] S. Mille, A. Belz, B. Bohnet, Y. Graham, E. Pitler, and L. Wanner, "The first multilingual surface realisation shared task (sr'18): Overview and evaluation results," in *Proceedings of the First Workshop on Multilingual Surface Realisation*, 2018, pp. 1–12.