Exploring Linguistic Typology Features in Multilingual Machine Translation

Oscar Moreno
Faculty of Science and Engineering
Pontificia Universidad Católica del Perú
omoreno@pucp.edu.pe

Arturo Oncevay
School of Informatics
University of Edinburgh
a.oncevay@ed.ac.uk

Abstract

We explore whether linguistic typology features could impact multilingual machine translation performance (many-to-English) by using initial pseudo-tokens and factored language-level embeddings. With 20 languages from different families or groups, we observed that the introduction of some features such as "Order of Subject (S), Object (O) and Verb (V)", "Position of Negative Word with respect to S-O-V" and "Prefixing vs. Suffixing in Inflectional Morphology" provided slight improvements in low-resource language-pairs, although they not overcome the average performance for all languages.

1 Introduction

Linguistic typology studies language variation and universals (Comrie, 1989), and crafted variables from the World Atlas of Language Structure (WALS; Dryer and Haspelmath, 2013) could support natural language processing (NLP) applications (Ponti et al., 2019). For instance, syntax-based features (e.g. word order) from WALS could benefit multilingual neural machine translation (NMT) approaches such as language clustering and language ranking by taking advantage of a language-level vector space (Oncevay et al., 2020).

In this work, we hypothesise that the direct input of specific typological features in multilingual NMT model could impact the translation performance, as they could behave as labels for language clustering[1]. We considered language embeddings with both factored (concatenation at every input token, (Sennrich and Haddow, 2016)) and initial pseudo-token settings, and features from the WALS' areas of Morphology, Nominal and Verbal Categories, Word Order and Simple Clauses.

[1] Pires et al. (2019) used two Word Order features to group languages for zero-shot transfer in multilingual BERT.

	Morph.	W.O.	Neg.
a	Prefixing vs. Suffixing in Inf. Morph.	Order of Subject-Object-Verb	Negative Morphemes
...			
f	Gender Distinctions in Ind. Personal Pronouns	Order of Adjective and Noun	Position of Negative Word With Respect to S-O-V
...			

Table 1: Examples of features. We use the alphabet sequence to refer to a specific set of features in §3.

2 Experimental setup

Dataset For a many-to-English setting, we chose a subset of 20 languages (one per family) from the TED corpus (Qi et al., 2018). See Appendix A.

WALS variables We selected the most completed features for our 20 languages set. We grouped the features in **Morphology** or *Morph* (14 variables, including Nominal and Verbal Categories), **Word Order** or *WO* (14 feats.) and **Negation** or *Neg* (8 negation-related features from Simple Clauses and Word Order). Appendix B has the full list and some examples are in Table 1.

Model and evaluation We used a small Transformer (Vaswani et al., 2017) with 2 layers for the encoder and decoder in Marian NMT (Junczys-Dowmunt et al., 2018) for both factored and non-factored settings. Moreover, we employed BLEU from sacreBLEU (Post, 2018) for evaluation, and considered a **Pseudo-Token** and **Factored baselines** that only used the language identity tokens (e.g.<es>) at the beginning of the sentence or concatenated in every token, respectively.

3 Results

Factored Typological Features We first studied the factored language-embeddings. We considered each feature as a factor group, and we used a maximum of five factors per experiment (for technical

limitations). This means that every token in the input sequence is concatenated with up to five typological feature-values of the source language. Appendix C shows all factored models tested.

	PT Baseline	Factor Baseline	Neg d|e|f|g|h	Morph a|b	WO a|b|c|d|e
BLEU Avg.	**18.56**	16.55	16.89	17.11	17.36
Stdev.	7.39	7.45	7.11	6.56	7.01
#L improved			5	4	0
Avg. increase			0.16	0.23	0

Table 2: Average BLEU and stdev. for different feature combinations using factors, including number of languages (#L) improved.

In Table 2, we show a combination per feature group[2], and we observe that almost all experiments outperform the Factored baseline but not the Pseudo-Token one. The outcome indicates that factored language-embeddings only encode better information about languages for language similarity tasks (e.g. phylogenetic inference) (Oncevay et al., 2020) but they are redundant and a potential information bottleneck for a translation objective.

There are are some individual improvements, however. Firstly, in *Neg*, the 'd|e|f|g|h' set increased performance in five languages. From experiment in other subgroups (see Appendix C), we considered 'f|g' as more relevant. Secondly, for *Morph*, the 'a|b' set obtained the third best result in average BLEU from all experiments, which confirms their importance. Lastly, for *WO*, the average BLEU in the 'a|b|c|d|e' set is the best one from all the factored systems. As features 'b' (Order of S-V) and 'c' (Order of O-V) are redundant to 'a' (Order of S-V-O), we consider 'a|d|e' (plus Order of Adposition and NP, and Order of Genitive and Noun) as the most important ones.

	Baseline	WO(2) a|d	WO(3) a|d|e	WO(5) a|b|c|d|e	N+M +WO
Multiple TT	18.56	18.40	18.31	17.96	18.36
Unique TT	18.56			18.20	18.41
Factored TF	16.55			17.36	

Table 3: BLEU scores with a multiple and unique typological tokens (TT), and factored typological features.

Multiple Typological Tokens We now focus in adding multiple pseudo-tokens that represent typological features at the beginning of every sentence.

[2]We did not perform an exhaustive search. We divided all features in different groups of five first, and then explore small subgroups of features based on their meaning.

Appendix D has the full results of the experiments and a summary is shown in Table 3, where we notice that the performance decreases by including more features/tokens from *WO*. Moreover, the *WO* 'a|d' set slightly outperform the baseline in four languages (up to +0.5 for Thai) with a comparable average BLEU (-0.16). Besides, we clustered one feature per group based on the previous insights (**N(f)+M(a)+WO(a)** = "Position of Negative Word w.r.t. S-O-V", "Prefixing vs. Suffixing in Inflectional Morph." and "Order of S-O-V"), which achieved a good overall score and the largest individual gain so far (+0.6 for Armenian (hy)).

Unique Typological Token We combine multiple feature-values in a unique variable/token, which is located at the beginning of every sentence. Table 3 shows that N+M+WO could not reach the baseline by 0.15 points, but slightly overcame its analogous multiple token setting (and outperforms the factored baseline with the same features). It also has the largest individual gains (*ka* and *hy* with +0.8 and +0.9, respectively, shown in Appendix D). Besides, all the gains for the shown combinations are in languages with less than 100k samples.

4 Discussion

Adding linguistic features as tokens is a way to tag typologically-based language clusters, and some features allowed slight translation improvement for individual low-resource languages. There is not an overall gain, but there is potential for exploiting multilingual data with a single language objective. However, Mueller et al. (2020) showed that multilingual NMT for low resource languages is highly variable in performance depending on the languages, so more experiments with datasets that have different language samples are needed.

A factor that is not analysed is how the distribution of feature-values seem to impact. Appendix E shows the case for N+M+WO, where there is not high agglomeration of languages in one specific value, except for "Strongly Suffixing". However, the feature-value distribution might be biased for the language sample, as most of the datasets available are usually from related languages (e.g. Indo-European ones). A diversity index (e.g. entropy, Gini) might be useful to make a more informed selection. Finally, another confound to consider is the dataset size: a synthetically reduced-size sample (e.g. up to 50k or 100k sentences) might allow to study better the effect of the typological features?

Acknowledgements

This work could not be possible without the support of REPU Computer Science or REPUcs (Research Experience for Peruvian Undergraduates: https://www.repuprogram.org/), a program that connects Peruvian students with researchers across the world. The first author worked as an intern in the University of Edinburgh as part of the REPUcs' 2021 cohort.

References

B. Comrie. 1989. *Language Universals and Linguistic Typology: Syntax and Morphology*. University of Chicago Press.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Marcin Junczys-Dowmunt, Kenneth Heafield, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. 2018. Marian: Cost-effective high-quality neural machine translation in C++. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 129–135, Melbourne, Australia. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3710–3718, Marseille, France. European Language Resources Association.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2391–2406, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. 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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

A Languages

Table 4 presents the list of 20 languages selected for training the many-to-English multilingual NMT model in all the experiments. We chose one language per family or group, and we prioritise the language with the largest size within a group. We took this decision to push the experiment to the limit: more data usually means a better performance in a massive model overall, and a fewer gap for improvement from extra signals or knowledge.

Id	Language	Family	Size (k)
eu	Basque	Isolate	5
ta	Tamil	Dravidian	6
mn	Mongolian	Mongolic	7
ka	Georgian	Kartvelian	13
hy	Armenian	IE/Armenian	21
sq	Albanian	IE/Albanian	43
id	Indonesian	Austronesian	85
th	Thai	Kra-Dai	96
el	Greek	IE/Hellenic	132
hu	Hungarian	Uralic	145
fa	Persian	IE/Indo-Iranian	148
vi	Vietnamese	Austroasiatic	169
tr	Turkish	Turkic	179
nl	Dutch	IE/Germanic	181
zh	Chinese	Sino-Tibetan	197
it	Italian	IE/Italic	201
ja	Japanese	Japonic	201
ko	Korean	Koreanic	202
ru	Russian	IE/Balto-Slavic	205
ar	Arabic	Afroasiatic	211

Table 4: List of all the languages considered in the study. IE = Indo-European

B WALS typological features

Table 5 shows all the linguistic typology features from WALS considered in the study.

C Experiments with Typological Factors

Table 6 shows all the results from the factor groups manually selected for this study.

D Experiments with Typological Tokens

Table 7 shows all the results using multiple and unique typological tokens.

E Feature-value-language for N+M+WO

- Neg(f): Position of Negative Word With Respect to Subject, Object, and Verb
  - SNegVO (8: vi,it,ru,id,hu,sq,th,zh)
  - MorphNeg (4: tr,fa,ta,ja)
  - More than one position (4: mn,nl,hy,el)
  - SONegV (3: ka,ko,eu)
  - NegVSO (1: ar)
- Morph(a): Prefixing vs. Suffixing in Inflectional Morphology
  - Strongly suffixing (16: it,tr,ru,id,ta,hu, mn,nl,hy,sq,el,zh,ja,ko,ar)
  - Little affixation (2: vi,th)
  - Weakly suffixing (2: fa,ka)
  - Equal prefixing and suffixing (1: eu)
- WO(a): Order of Subject, Object and Verb
  - SOV (8: tr,fa,ta,ka,mn,ja,ko,eu)
  - SVO (7: vi,it,ru,id,sq,th,zh)
  - No dominant order (4: hu,nl,hy,el)
  - VSO (1: ar)

Id	WALS Id	WALS Area	Feature name
Morph a	26A	Morphology	Prefixing vs. Suffixing in Inflectional Morphology
Morph b	27A	Morphology	Reduplication
Morph c	33A	Nominal Categories	Coding of Nominal Plurality
Morph d	36A	Nominal Categories	The Associative Plural
Morph e	37A	Nominal Categories	Definite Articles
Morph f	44A	Nominal Categories	Gender Distinctions in Independent Personal Pronouns
Morph g	46A	Nominal Categories	Indefinite Pronouns
Morph h	48A	Nominal Categories	Person Marking on Adpositions
Morph i	49A	Nominal Categories	Number of Cases
Morph j	50A	Nominal Categories	Asymmetrical Case-Marking
Morph k	51A	Nominal Categories	Position of Case Affixes
Morph l	69A	Verbal Categories	Position of Tense-Aspect Affixes
Morph m	70A	Verbal Categories	The Morphological Imperative
Morph n	71A	Verbal Categories	The Prohibitive
WO a	81A	Word Order	Order of Subject, Object and Verb
WO b	82A	Word Order	Order of Subject and Verb
WO c	83A	Word Order	Order of Object and Verb
WO d	85A	Word Order	Order of Adposition and Noun Phrase
WO e	86A	Word Order	Order of Genitive and Noun
WO f	87A	Word Order	Order of Adjective and Noun
WO g	88A	Word Order	Order of Demonstrative and Noun
WO h	89A	Word Order	Order of Numeral and Noun
WO i	90A	Word Order	Order of Relative Clause and Noun
WO j	92A	Word Order	Position of Polar Question Particles
WO k	94A	Word Order	Order of Adverbial Subordinator and Clause
WO l	95A	Word Order	Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase
WO m	96A	Word Order	Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun
WO n	97A	Word Order	Relationship between the Order of Object and Verb and the Order of Adjective and Noun
Neg a	112A	Simple Clauses	Negative Morphemes
Neg b	113A	Simple Clauses	Symmetric and Asymmetric Standard Negation
Neg c	114A	Simple Clauses	Subtypes of Asymmetric Standard Negation
Neg d	115A	Simple Clauses	Negative Indefinite Pronouns and Predicate Negation
Neg e	143F	Word Order	Postverbal Negative Morphemes
Neg f	144A	Word Order	Position of Negative Word With Respect to Subject, Object, and Verb
Neg g	143E	Word Order	Preverbal Negative Morphemes
Neg h	143A	Word Order	Order of Negative Morpheme and Verb

Table 5: List of all the WALS features considered in the study. “Id” is the identification code used in this paper.

L.	Size(k)	PT Baseline	Factor Baseline	Neg a|b|c|d|e	Neg a|b|d|e|h	Neg d|e|f|g|h	Morph a|b	Morph a|d|f|g|i	Morph c|d|e|f|g	Morph h|i|j|k	Morph l|m|n	WO a|b|c|d|e	WO a|d|f|h|i	WO f|g|h|i|j	WO k|l|m|n
ar	211	22.3	20.5	20.8	21.6	20.8	**22.6**	19.3	19.4	20.6	19.5	20.7	20.8	18.6	19.6
ru	205	18.7	18.2	18	18	18.7	18.1	18.4	18	18.7	17	18.1	18.5	17.5	18.2
ko	202	13.3	11.2	12.7	**13.6**	10.3	**13.6**	12.7	12.5	11.5	12.4	11.8	12.2	12.2	12.8
it	201	29.8	26.1	25.5	25.7	27.2	27.7	26.7	26.9	28.4	23.9	27.2	27.9	23.5	27.9
ja	201	9.6	7.6	9	8.8	8.1	9.1	9.1	8.9	8.7	9.1	8.7	8.8	9.6	9.1
zh	197	14.6	13.6	14.6	14.5	**14.7**	14.6	**14.7**	14.5	14.6	14.5	14.5	**14.7**	14.3	14.6
nl	181	27.6	22.8	24	25.2	23.6	20.5	25.8	25.4	26	22	25.8	26.2	22.1	25.2
tr	179	17.9	11.6	14.4	16.2	13.5	17.7	15.2	14.5	13.8	13.2	14.8	15.3	17.7	14.6
vi	169	21.5	20.6	19.3	19.9	**21.6**	20.4	19.8	19.7	21.4	20.6	20.6	20	19.1	20.1
fa	148	20.2	18.4	15.2	17.1	17.7	15.3	16.2	16.4	18.7	16.1	17.3	17	16	16
hu	145	18.9	15.3	17.7	18.2	**19**	**19**	16.4	16.7	16	18.8	17.2	17.2	16.1	17.2
el	132	30.6	30	29.7	29.8	29.8	29.6	28.1	28.1	29.9	27	29.7	29.7	26.8	29.3
th	96	17.4	16.6	15.7	16.2	**17.7**	16.8	16.7	15.7	**17.5**	16.9	16.8	16.4	15.3	16.6
id	85	25.5	23.1	22.9	23.6	**25.7**	23.6	22.4	22.7	**25.6**	25.5	24.8	23.9	20.3	23.6
sq	43	28.8	**29.2**	18	19.7	22.2	23.1	19.4	21.4	22.9	21.2	27.6	26.5	21.5	26
hy	21	16	15	13.6	15	14.7	**16.2**	13	13.4	13.6	14.6	15.9	**16.1**	13.9	16
ka	13	14.1	12.6	12.2	12.8	12.2	12.7	11.3	11.1	11.6	11.9	13.5	12.5	12.7	11
mn	7	6.7	5.3	5.1	6.4	5.3	6.4	5.5	5.4	5.2	5.4	6.3	5.6	6.3	5.5
ta	6	5.3	3.7	3.3	4.1	4.6	5.1	3	3.4	3.8	4.8	5	4.6	5.2	3
eu	5	12.3	9.5	8.1	11.4	10.3	10	9.5	8.8	9.6	10.7	10.9	9.9	10	9.3
Avg (20 lang)		18.56	16.55	15.99	16.89	16.89	17.11	16.16	16.15	16.91	16.26	17.36	17.19	15.94	16.78

Table 6: BLEU for different feature combinations using factors. Scores better than the PT baseline are in bold.

			Multiple Typological Tokens						Unique Typological Token			
L.	Size (k)	PT Baseline	WO a|d	WO a|d|e	M:a|b + W:a	N+M +W0	WO a|b|c|d|e	N:f|h + M:a|b + W:a|d	N+M +W0	WO a|b|c|d|e	N:f|h M:a|b + W:a|d	M:a|b + W:a
ar	211	22.3	22.2	22	21.9	22	21.6	21.5	21.9	21.7	21.5	21.9
ru	205	18.7	18.3	18.4	18.4	18.4	18.3	18.2	18.4	18.3	18.3	18.4
ko	202	13.3	13.1	12.8	12.6	13	12.7	12.4	13.2	12.9	13.2	13.3
it	201	29.8	29.6	29.5	29.6	**29.9**	28.9	29.5	29.8	28.6	29.1	29.6
ja	201	9.6	9.4	9.3	9.3	9.6	9.2	9.1	9.5	9.4	9.5	9.6
zh	197	14.6	**14.9**	**14.7**	14.6	14.5	14.5	14.3	14.6	14.6	14.6	14.5
nl	181	27.6	27.6	27.3	25.9	26.5	25.6	25.6	27	27	26.6	27.3
tr	179	17.9	16.8	16.6	16.2	16.8	15.7	15.9	16.9	17.1	17	17.1
vi	169	21.5	21.1	21.3	21	**21.7**	21	20.9	21.4	21	21	21.4
fa	148	20.2	20.2	20.2	19.4	19.9	19.7	19.3	19.7	19.8	19.8	19.8
hu	145	18.9	18.1	18.2	18.3	18.8	18.1	18.4	17.7	17.9	17.3	18
el	132	30.6	30.4	30.4	30.4	30.3	30.3	29.9	30.2	30	29.8	30.3
th	96	17.4	**17.9**	**17.6**	**17.5**	**17.7**	17.4	17.2	**17.7**	**17.5**	17.3	**17.5**
id	85	25.5	25.2	25.3	25.4	25.5	25.2	25.1	**25.7**	25.1	24.7	25
sq	43	28.8	28.6	28.5	28.7	**29**	28.3	28.3	**29.3**	28.7	27.2	27.2
hy	21	16	**16.3**	**16.4**	**16.6**	**16.6**	**16.2**	**16.4**	**16.9**	16	**16.3**	**16.1**
ka	13	14.1	**14.4**	**14.3**	13.4	14.1	14	13.9	**14.9**	**14.4**	14.1	**14.2**
mn	7	6.7	**7**	6.4	**6.8**	6.4	6.1	6	6.4	6.2	**7.3**	**6.8**
ta	6	5.3	5.3	**5.4**	**5.4**	**5.5**	**5.5**	4.8	5.1	5.3	5.2	**5.6**
eu	5	12.3	11.6	11.7	11.2	11.1	11	11.6	11.8	**12.4**	**12.4**	11.9
Avg (20 lang)		18.56	18.40	18.32	18.13	18.37	17.97	17.92	18.41	18.20	18.11	18.28

Table 7: BLEU scores using multiple and unique typological tokens.Scores better than the PT baseline are in bold.