Universal Reordering via Linguistic Typology

Joachim Daiber **Miloš Stanojević** **Khalil Sima'an**
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
{initial.last}@uva.nl

Abstract

In this paper we explore the novel idea of building a single *universal* reordering model from English to a large number of target languages. To build this model we exploit typological features of word order for a large number of target languages together with source (English) syntactic features and we train this model on a *single combined* parallel corpus representing all (22) involved language pairs. We contribute experimental evidence for the usefulness of linguistically defined typological features for building such a model. When the universal reordering model is used for preordering followed by monotone translation (no reordering inside the decoder), our experiments show that this pipeline gives comparable or improved translation performance with a phrase-based baseline for a large number of language pairs (12 out of 22) from diverse language families.

1 Introduction

Various linguistic theories and typological studies suggest that languages often share a number of properties and that their differences fall into a small set of parameter settings (Chomsky, 1965; Greenberg, 1966; Comrie, 1981). While this intuition has influenced work on multilingual parsing (Zeman and Resnik, 2008; McDonald et al., 2011), it has found less practical use in other areas of natural language processing, such as the task of machine translation. In machine translation, significant word order differences between languages often constitute a challenge to translation systems. Word order differences are frequently given special treatment, such as in the case of preordering (Xia and McCord, 2004; Neubig et al., 2012; Stanojević and Sima'an, 2015, inter alia), which is a technique heavily used in practice as a means to improve both translation quality and efficiency. In preordering, word order is predicted based on manually created rules or based on statistical models estimated on word-aligned training data exploiting only source language features. This approach works well for some language pairs, however it usually demands a separate, dedicated preordering model for every source-target language pair, trained on a word-aligned corpus specific for the particular language pair.

But if the similarities and differences between languages can indeed be captured with a small set of features, as linguistic theory suggests, then it seems more expedient to try to benefit from the similarities between target languages in the training data, which is not possible when training a separate preordering model for every new target language. Ideally, the word-aligned data obtained for various target languages should be combined to train a single, *universal* reordering model with a single set of features. The questions addressed in this paper are (1) could a linguistically inspired universal reordering model show any promising experimental results and (2) how can such a universal reordering model be built?

For building an effective universal reordering model, we need access to a resource that describes the similarities and differences between (target) languages in a small set of properties. The World Atlas of Language Structures (Dryer and Haspelmath, 2013), WALS, is a major resource which currently specifies the abstract linguistic properties of 2,679 languages.[1] In this paper we explore the use of the

[1] http://wals.info/languoid

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3167–3176, Osaka, Japan, December 11-17 2016.

linguistically defined WALS features for a broad range of target languages and show that these features have merit for building a single, universal reordering model. The universal reordering model is based on a feed-forward neural network which is trained to predict the target word order given source syntactic structure and all available WALS parameter settings for each of the 22 target languages involved. By training the feed-forward neural network on the WALS-enriched data from a broad range of target languages, we enable the universal reordering model to both learn how much to trust the WALS parameters and to exploit possible interactions between them for different target languages. When the universal reordering model is followed by *monotone translation* (no reordering inside the decoder), our experiments show that this pipeline gives comparable or improved translation performance to a Moses baseline with standard distortion settings, for a large number of language pairs. This suggests that typological target language features could play a key role in building better, more general preordering models, which have, heretofore, been trained solely on source sentences and word alignments, but had no access to other target-side information.

We believe that the experiments presented in this paper have both theoretical and practical implications. Firstly, they show the utility and provide empirical support for the value of linguistic typology. Secondly, they enable building more compact preordering models that should generalize to a broad set of target languages and which potentially apply for the low resource setting where no or little parallel data is available for a specific target language.

2 Related Work

The most basic usage of linguistic knowledge in preordering is in restricting the search space of possible reorderings by using syntactic parse trees. Earlier work was done mostly on constituency trees (Khalilov and Sima'an, 2012; Xia and McCord, 2004) while more recent versions of preordering models mostly use dependency trees (Lerner and Petrov, 2013; Jehl et al., 2014). Preordering in syntax-based models (whether dependency or constituency) is done on the local level where for each constituent (or head word) the classifier decides how the children (or dependent words) should be reordered.

Employing classifiers to make local decisions on each tree node is one machine learning approach to solving this problem. An alternative to employing machine learning techniques is the use of linguistic knowledge that can in some cases give clear rules for the reordering of children in the tree. An early example of rule-based preordering is by Collins et al. (2005), who develop linguistically justified rules for preordering German into English word order. Similar in spirit but much simpler is the approach of Isozaki et al. (2010), who exploit the fact that Japanese word order is in large part the mirror image of English word order—the heads of constituents in English are in final position while in Japanese they are in initial position. Preordering English sentences into Japanese word order thus only involves two simple steps: (1) Finding the parse tree of the English sentence (the authors used HPSG derivations) and (2) moving the head of each constituent to the initial position. However, this approach does not seem to scale up easily because manually encoding reordering rules for all the world's language pairs would be a rather difficult and very slow process.

In contrast to manually encoding rules for language pairs, we could use similarities and differences between target languages encoded in existing *typological databases* of structural properties of the world's languages, e.g., the World Atlas of Language Structures, WALS (Dryer and Haspelmath, 2013). Therefore, the challenge taken up in the present work is how to exploit typological databases such as WALS to guide the learning algorithm into making the right decisions about word order. So if, for instance, a feature indicates that the target language follows VSO (verb-subject-object) word order, then the preordering algorithm should learn to transform the English parse tree from SVO into a VSO tree. Using typological features like these in a machine learning system for preordering constitutes a compromise between knowledge-based (rules) and data-driven (learning) approaches to preordering.

Researchers in linguistic typology have produced various initiatives to collect typological data in a centralized and structured format out of which WALS is the most comprehensive one. We briefly discuss WALS in Section 3. WALS has been used before in computational linguistics, e.g., by Östling (2015) who performed a typological study of word order based on a corpus of New Testament translations in 986

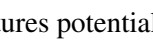
Feature	Name	Distribution of Values	Feature	Name	Distribution of Values
37A	Definite Articles		82A	Order of Subj. and Verb	
46A	Indefinite Pronouns		83A	Order of Object and Verb	
48A	Person Marking on Adpositions		84A	Order of Object, Oblique, and Verb	
52A	Comitatives and Instrumentals		85A	Order of Adposition and NP	
53A	Ordinal Numerals		86A	Order of Genitive and Noun	
54A	Distributive Numerals		87A	Order of Adjective and Noun	
55A	Numeral Classifiers		88A	Order of Demonstrative and Noun	
56A	Conj. and Universal Quantifiers		89A	Order of Numeral and Noun	
57A	Pos. of Pron. Poss. Affixes		90A	Order of Relative Clause and Noun	
61A	Adjectives without Nouns		91A	Order of Degree Word and Adj.	
66A	The Past Tense		92A	Position of Polar Quest. Particles	
67A	The Future Tense		93A	Position of Interr. Phrases	
68A	The Perfect		94A	Order of Adv. Subord. + Clause	
69A	Pos. of Tense-Aspect Affixes		95A	Rel. of Obj. + Verb and Adp. + NP	
81A	Order of Subj., Obj. and Verb		96A	Rel. of Obj. + Verb and Rel. Clause + Noun	
81B	Two Dominant SVO Orders		97A	Rel. of Obj. + Verb and Adj. + Noun	

Table 1: WALS features potentially relevant to determining word order.

languages. This study found that the word order typology created from such data and the information in WALS show a high level of agreement. Finally, Bisazza and Federico (2016) survey word reordering in machine translation and categorize languages based on their WALS features. In the present work, we make novel use of linguistic typological features from WALS for building a universal reordering model.

3 Linguistic Typology

The field of linguistic typology studies the similarities and distinguishing features between languages and aims to classify them accordingly. Among other areas, the World Atlas of Language Structures describes general properties of each language's word order. Overall, WALS contains 192 features, but not all features are relevant to determining word order. Many WALS features deal with phonology, morphology or lexical choice: Feature 129A, for example, describes whether the language's words for "hand" and "arm" are the same. Hence, for simplicity's sake we pre-select the subset of WALS features potentially relevant to determining word order and describe this subset in the following. Table 1 provides an overview of these features, along with an indication of the relative frequency distribution of each of their values over all languages in WALS.

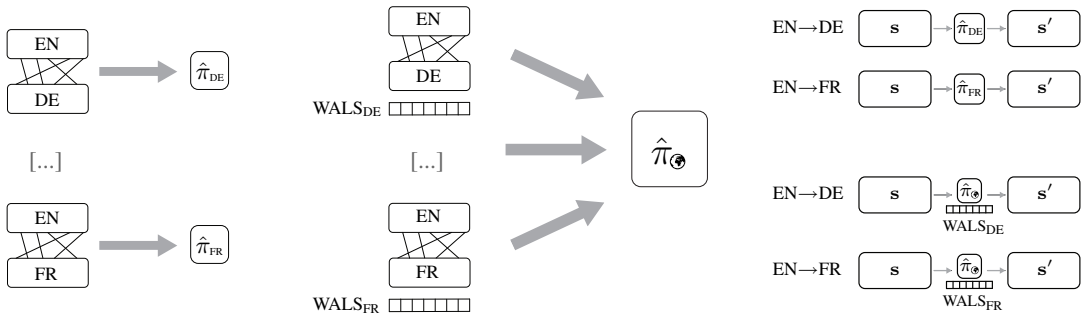
One of the most common ways to classify languages is according to the order of the subject, the object and the verb in a transitive clause. Accordingly, a number of WALS features describe the order of these elements. WALS Feature 81A classifies languages into 6 dominant clause-level word orders. For languages such as German or Dutch, which do not exhibit a single dominant clause-level order, Feature 81B describes 5 combinations of two acceptable word orders. Additionally, two features describe whether the verb precedes the subject (82A) and whether the verb precedes the object (83A). The position of adjuncts in relation to the object and the verb are described in Feature 84A and the internal structure of adpositional phrases is described in Feature 85A, which specifies whether the language uses pre-, post- or inpositions. Finally, the following properties describe the order of words in relation to nouns: Feature 86A specifies the position of genitives (e.g. *the girl's cat*), Feature 87A the position of adjectives (e.g. *yellow house*), Feature 89A the position of numerals (e.g. *10 houses*) and Feature 90A the position of relative clauses (e.g. *the book that I am reading*) in relation to the noun.

4 Universal Reordering Model

Our universal reordering model uses a preordering architecture similar to the (non-universal) preordering model of De Gispert et al. (2015), which in turn is based on the authors' earlier work on logistic regression and graph search for preordering (Jehl et al., 2014).

4.1 Basic Preordering Model

In this neural preordering model, a feed-forward neural network is trained to estimate the swap probabilities of nodes in the source-side dependency tree. The learning task is defined as follows: How likely is it that two nodes *a* and *b* are in the linear order (a, b) or (b, a) in the target language? Preordering then consists of finding the best sequence of swaps according to this model. While De Gispert et al. (2015)



(a) Basic preordering training. (b) Universal reordering model training. (c) Applying the models.

Figure 1: Training and application of basic preordering models and the universal reordering model.

use a depth-first branch-and-bound algorithm to find the best permutation, we use the k-best version of this algorithm and minimize the resulting preordering finite-state automaton to produce a lattice of word order choices (Daiber et al., 2016).

Model estimation Training examples are extracted from all possible pairs of children of the source dependency tree node, including the head itself. The crossing score of two nodes a and b (a precedes b in linear order) and their aligned target indexes A_a and A_b is defined as follows:

$$\mathrm{cs}(a, b) = |\{(i, j) \in A_a \times A_b : i > j\}|$$

A pair (a, b) is swapped if $\mathrm{cs}(b, a) < \mathrm{cs}(a, b)$, i.e. if swapping reduces the number of crossing alignment links. Training instances generated in this manner are then used to estimate the order probability $p(i, j)$ for two indexes i and j. The best possible permutation of each node's children (including the head) is determined via graph search. The score of a permutation π of length k consists of the order probabilities of all possible pairs:

$$\mathrm{score}(\pi) = \prod_{1 \le i < j \le k | \pi[i] > \pi[j]} p(i, j) \cdot \prod_{1 \le i < j \le k | \pi[i] < \pi[j]} 1 - p(i, j)$$

De Gispert et al. (2015) use a feed-forward neural network (Bengio et al., 2003) to predict the orientation of a and b based on 20 source features, such as the words, POS tags, dependency labels, etc.[2]

Permutation lattices To find the sequence of swaps leading to the best overall permutation according to the model, the score of a permutation is obtained by extending a partial permutation π' of length k' by one index i (Jehl et al., 2014). This score can be efficiently computed as:

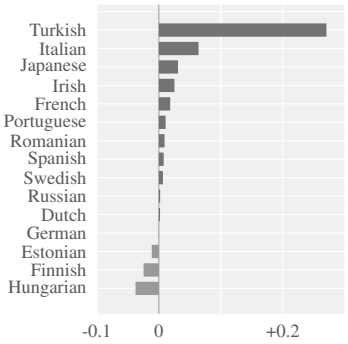
$$\mathrm{score}(\pi' \cdot \langle i \rangle) = \mathrm{score}(\pi') \cdot \prod_{j \in V | i > j} p(i, j) \cdot \prod_{j \in V | i < j} 1 - p(i, j)$$

Instead of extracting the single-best permutation, we use the *k*-best extension of branch-and-bound search (van der Poort et al., 1999). The resulting k-best permutations are then compressed into a minimal deterministic acceptor and unweighted determinization and minimization are performed using OpenFST (Allauzen et al., 2007).

4.2 Estimating a Universal Reordering Model

The universal reordering model differs from the basic neural preordering model in terms of features and training data collection. Differences in training data collection and application are illustrated in Figure 1.

[2]Full set of features: words, word classes, dependency labels, POS tags, coarse POS tags, word and class of the left-most and right-most child token, and the tokens' distance to their parent.



(a) Absolute 1-best Kendall τ improvements.

	Manual word alignments			Automatic word alignments		
Language	τ@10	τ@100	τ@1000	τ@10	τ@100	τ@1000
French	+01.95	+03.05	+03.05	+01.28	+02.12	+02.13
German	+04.03	+05.61	+05.61	+06.85	+08.27	+08.27
Italian	+06.39	+06.75	+06.75	+03.07	+03.32	+03.32
Portuguese	+05.87	+07.89	+07.89	+03.24	+03.55	+03.55
Spanish	+05.97	+06.57	+06.57	+04.28	+05.11	+05.11
Romanian	+02.49	+03.37	+03.37	+01.23	+02.09	+02.10
Swedish	+00.13	+00.42	+00.42	+00.71	+01.18	+01.18

(b) N-best permutation quality on manually and autom. aligned data.

Figure 2: Intrinsic quality of word order predictions (improvement over source word order).

In addition to the source features used in the standard neural preordering model (cf. Section 4.1), we add a feature indicating the source word order of the two tokens, as well as the type of end-of-sentence punctuation. We then add WALS features 37, 46, 48, 52–57, 61, 66–69 and 81–97. WALS features are represented by their ID and the value for the current target language (e.g. "WALS_87A=Adj-Noun" or "WALS_87A=Noun-Adj"). For the most basic word order features (81, 82 and 85–91), we additionally add a feature indicating if the order of the node pair agrees with the order specified by the WALS feature.[3]

While the training data for a standard preordering model consists of source sentences and their target-language order retrieved via word alignments, the training data for the universal reordering model is comprised of training examples from a large number of language pairs. Because of the diversity of this data, special care has to be taken to ensure a balanced dataset. We use an equal number of sentences from each language-specific training subcorpus. Additionally, we reduce class imbalance by further randomly shuffling the source tokens when creating training instances. This ensures a balanced distribution of classes in the training data. The distribution of the two classes is 84.5%/15.5% in the original and 50.1%/49.9% in the randomized dataset.

4.3 Intrinsic Evaluation

We use NPLM (Vaswani et al., 2013) to train a feed-forward neural network to predict the orientation of two nodes a and b based on the features described in Section 4.2. The network consists of 50 nodes on the input layer, 2 on the output layer, and 50 and 100 on the two hidden layers. We use a learning rate of 0.01, batch sizes of 1000/64 and perform 60 training epochs, ensuring convergence of the log-likelihood on a validation set.

Preordering data The training data for the universal reordering model consists of a combined corpus of 30k sentence pairs each from the Tatoeba corpus (Tiedemann, 2012) for French, German, Japanese, Portuguese, Russian, and Spanish as well as 100k sentence pairs each from the OpenSubtitles 2012 corpus (Tiedemann, 2012) for Spanish, Portuguese, Italian, Danish, Romanian, Swedish, French, Greek, Russian, Polish, Arabic, Hebrew, Hungarian, Czech, Finnish, Icelandic, Dutch, Slovak, Chinese, German and Turkish. Word alignments for all corpora were produced using MGIZA (Och and Ney, 2003) using *grow-diag-final-and* symmetrization and performing 6, 6, 3 and 3 iterations of IBM M1, HMM, IBM M3 and IBM M4 respectively. To evaluate the model, we also use sets of manually word-aligned sentence for the following language pairs: En–Ja (Neubig, 2011), En–De (Padó and Lapata, 2006), En–It (Farajian et al., 2014), En–Fr (Och and Ney, 2003), En–Es and En–Pt (Graça et al., 2008).

Quality of word order predictions Figure 2a shows the intrinsically measured quality of the predictions by the universal reordering model. We use Kendall τ (Kendall, 1938) to measure the correlation

[3]Example for WALS feature 87A=Adj-Noun: $f(a,b) = \begin{cases} \text{"W87A:ab"} & \text{if } a = \text{adj} \wedge b = \text{noun} \\ \text{"W87A:ba"} & \text{if } a = \text{noun} \wedge b = \text{adj} \end{cases}$

Language	# sent.	# tok.	BiHDE ↑	Language	# sent.	# tok.	BiHDE ↑	Language	# sent.	# tok.	BiHDE ↑
Spanish	800k	14.29	0.57	Greek	800k	14.36	0.65	Finnish	800k	14.36	0.69
Portuguese	800k	14.29	0.58	Russian	800k	15.08	0.65	Icelandic	800k	14.10	0.69
Italian	800k	14.68	0.61	Polish	800k	14.22	0.67	Dutch	800k	14.37	0.70
Danish	800k	14.50	0.62	Arabic	800k	14.84	0.68	Slovak	638k	15.08	0.70
Romanian	800k	14.24	0.64	Hebrew	800k	14.61	0.68	Chinese	636k	10.39	0.71
Swedish	800k	14.49	0.64	Hungarian	800k	14.33	0.68	German	800k	14.62	0.72
French	800k	14.25	0.65	Czech	800k	14.19	0.69	Turkish	800k	14.25	0.72

Table 2: Properties of training data from the 2012 OpenSubtitles corpus.

between the predicted word order and the *oracle* word order determined via the word alignments. Figure 2a plots absolute Kendall τ improvement over the original, i.e. unreordered, source sentence for the single best permutation for a number of language pairs. The three worst-performing target languages in Figure 2a, Estonian, Finnish and Hungarian, are all morphologically rich, indicating that additional considerations may be required to improve word order for languages of this type. Figure 2b shows the quality of n-best permutations of the universal reordering model for both manually and automatically word-aligned sentence pairs. This table allows two observations: Firstly, the evaluation of word order quality using automatic alignments shows good agreement with the evaluation using manually word-aligned sentences, thus highlighting that automatic alignments should suffice for this purpose in most cases. Secondly, we can observe that for all datasets presented in this table little is gained from increasing the number of extracted permutations beyond 100 predictions. We therefore apply a maximum number of 100 permutations per sentence in all experiments presented in the rest of this paper.

5 Translation Experiments

To evaluate the universal reordering model in a real-world task, we perform translation experiments on various language pairs. As a baseline system, we use a plain phrase-based machine translation system using a distortion-based reordering model with a distortion limit of 6. When applying the universal reordering model, we produce a lattice from each sentence's best 100 word order permutations. This lattice is then passed to the machine translation system and no additional reordering is allowed. During training, we choose the source sentence permutation closest to the gold word order determined via the word alignments (lattice silver training; Daiber et al., 2016). The word alignments for the preordered training corpus are then recreated from the original MGIZA alignments and the selected permutation.[4]

Translation experiments are performed with a phrase-based machine translation system, a version of Moses (Koehn et al., 2007) with extended lattice support.[5] We use the basic Moses features and perform 15 iterations of batch MIRA (Cherry and Foster, 2012). To control for optimizer instability, we perform 3 tuning runs for each system and report the mean BLEU score for these runs (Clark et al., 2011). As a baseline we use a translation system with distortion limit 6 and a distance-based reordering model. For each language pair, a 5-gram language model is estimated using *lmplz* (Heafield et al., 2013) on the target side of the parallel corpus.

5.1 Evaluating on a Broad Range of Languages

In order to test the ideas presented in this paper, we evaluate our model on a broad range of languages from various language families. While doing so, it is important to ensure that the results are not skewed by differences in the corpora used for training and testing each language pair. We therefore build translation systems from the same corpus and domain for every language pair. We use the 2012 OpenSubtitles corpus[6] (Tiedemann, 2012) to extract 800,000 parallel sentences for each language pair, ensuring that every sentence pair contains only a single source sentence and that every source sentence contains at least 10 tokens. For each language pair, 10,000 parallel sentences are retained for tuning and testing. We use English as the source language in all language pairs. Table 2 summarizes properties of the data

[4]To keep the experiments manageable, we opted not to re-align the preordered training corpus using MGIZA. Re-alignment often leads to improved translation results, therefore we are likely underestimating the potential preordered translation quality.

[5]Made available at `https://github.com/wilkeraziz/mosesdecoder`.

[6]`http://opus.lingfil.uu.se/OpenSubtitles2012.php`

	BLEU	Δ BLEU				BLEU	Δ BLEU		
Language	Baseline	No WALS	WALS ↓	Gold	Language	Baseline	No WALS	WALS ↓	Gold
Dutch	13.76	+0.11	+0.79	+3.44	Greek	7.22	−0.02	+0.01	+0.49
Italian	23.59	+0.04	+0.48	+1.83	Arabic	5.36	−0.10	−0.01	+0.36
Turkish	5.89	−0.36	+0.43	+0.80	Swedish	25.60	−0.14	−0.03	+2.04
Spanish	23.82	−0.27	+0.29	+1.98	Slovenian	10.56	−0.35	−0.10	+1.21
Portuguese	25.94	−0.48	+0.21	+1.64	Slovak	15.56	−0.09	−0.13	+1.98
Finnish	9.95	+0.13	+0.16	+0.51	Icelandic	14.97	−0.31	−0.14	+0.66
Hebrew	11.64	+0.30	+0.11	+2.24	Polish	17.68	−0.45	−0.16	+0.40
Romanian	16.11	+0.11	+0.11	+1.14	Russian	20.12	−0.47	−0.17	+0.92
Hungarian	8.26	−0.10	+0.10	+0.61	German	17.08	−0.21	−0.19	+3.31
Danish	26.36	−0.13	+0.08	+1.56	Czech	12.81	−0.47	−0.21	+0.70
Chinese	11.09	−0.32	+0.05	+0.44	French	19.92	−0.70	−0.23	+1.20

Table 3: Translation experiments with parallel subtitle corpora.

used in these experiments. Apart from the average sentence length and the number of training examples, we report Bilingual Head Direction Entropy, BiHDE (Daiber et al., 2016), which indicates the difficulty of predicting target word order given the source sentence and its syntactic analysis. The language pairs in Table 2 are sorted by their BiHDE score, meaning that target languages whose word order is more deterministic are listed first. For each language pair, we train four translation systems:

Baseline The baseline system is a standard phrase-based machine translation system with a distance-based reordering model, a distortion limit of 6, and a maximum phrase length of 7.

Gold The gold system provides an indication for the upperbound achievable translation quality using preordering. In this system, the tuning and test sets are word-aligned along with the training portion of the corpus and the word alignments are then used to determine the optimal source word order. While this system provides an indication for the theoretically achievable improvement, this improvement may not be achievable in practice since not all information required to determine the target word order may be available on the source side (e.g. morphologically rich languages can allow several interchangeable word order variations). Apart from the source word order, the gold system is equivalent to the Baseline system.

No WALS As a baseline for our preordering systems, we create a translation system that differs from our universal reordering model only in the lack of WALS information. The preordering model is trained using the standard set of features described in Section 4.1 with only a single additional feature: the name of the target language. As in the WALS system, this system is applied by generating a minimized lattice from the 100-best permutations of each sentence and restricting the decoder's search space to this lattice. This system therefore isolates two potential sources of improvement: (1) improvement due to restricting the search space by the source dependency tree and (2) improvement from the preordering model itself, independent of the typology information provided by WALS.

WALS The WALS system applies the universal reordering model introduced in Section 4.2. For each language pair, the preordering model is provided with the target language and all the WALS features available for this language. The MT system's search space is then restricted using the minimized lattice of the 100-best word order permutations for each sentence and no additional reordering within the MT decoder is allowed.

The results of the translation experiments using the OpenSubtitles corpora are presented in Table 3. BLEU scores for the No WALS, WALS and Gold systems are reported as absolute improvement over the Baseline system (Δ BLEU). Over the three tuning runs performed for each system, we observe minor variance in BLEU scores (mean standard deviations: Baseline 0.04, No WALS 0.05, WALS 0.05, Gold 0.07), thus we report the mean BLEU score for each system's three runs.

While performing monotone decoding (i.e., allowing no reordering on top of the input lattice), the universal reordering model (WALS) enables improvements or comparable performance for the majority

	Dataset				Baseline	WALS
Language	Domain	# sent.	# tok.	BiHDE	BLEU	Δ BLEU ↓
Turkish	News	0.20m	23.54	0.73	8.27	+0.34
Spanish	Parl. + News	1.73m	23.47	0.58	24.34	+0.18
Italian	Parl. + News	1.67m	24.49	0.61	24.83	+0.13
Portuguese	Parl. + News	1.73m	23.67	0.58	32.13	−0.08
Hungarian	Parl. + News	1.41m	17.11	0.70	7.63	−0.19

Table 4: Translation experiments with varying training data and domains.

of the language pairs we evaluated while the No WALS system performs worse for most language pairs. This suggests that the improvements are not due to the neural preordering model or the lattice-based translation alone, but that the WALS information is crucial in enabling these results.

5.2 Influence of Domain and Data Size

While the experiments using the subtitle corpora presented in the previous section allow a fair comparison of a large number of language pairs, they also exhibit certain restrictions: (1) all experiments are limited to a single domain, (2) the source sentences are fairly short, and (3) to ensure consistent corpus sizes, a limited number of 800k sentence pairs had to be used. Therefore, we perform an additional set of experiments with data from different domains, longer sentences and a larger number of sentence pairs.

To train the translation systems for these experiments, we use the following training data: For En–It, En–Es and En–Pt, we train systems on Europarl v7 (Koehn, 2005). En–Hu uses the WMT 2008 training data,[7] En–Tr the SETIMES2 corpus (Tiedemann, 2009). Tuning is performed on the first 1512 sentences of newssyscomb2009+newstest2009 (En–It), newstest2009 (En–Es), newsdev2016 (En–Tr), newstest2008 (En–Hu), and the first 3000 sentences of news commentary v11 (En–Pt). As test sets we use the rest of newssyscomb2009+newstest2009 (En–It), newstest2013 (En–Es), newstest2016 (En–Tr), newstest2009 (En–Hu), and the first 3000 sentences of news commentary v11 not used in the dev set (En–Pt). All datasets are filtered to contain sentences up to 50 words long, and tokenization and truecasing is performed using the Moses tokenzier and truecaser. Statistics about each dataset and the dataset's domains, as well as translation results for the baseline system and the universal reordering model are summarized in Table 4. The results indicate that despite the longer sentences and different domains, the universal reordering model performs similarly as in the experiments performed in Section 5.1.

Our intrinsic evaluation (Section 4.3) as well as the extrinsic evaluation on a translation task (Section 5) indicate that a universal reordering model is not only feasible but can also provide good results on a diverse set of language pairs. The performance difference between the No WALS baseline and the universal reordering model (cf. Table 3) further demonstrates that the typological data points provided by WALS are the crucial ingredient in enabling this model to work.

6 Conclusion

In this paper, we show that linguistics in the form of linguistic typology and modern methods in natural language processing in the form of neural networks are not rivaling approaches but can come together in a symbiotic manner. In the best case, combining both approaches can yield the best of both worlds: the generalization power of linguistic descriptions and the good empirical performance of statistical models. Concretely, we have shown in this paper that it is possible to use linguistic typology information as input to a preordering model, thus enabling us to build a single model with a single set of model parameters for a diverse range of languages. As an empirical result, our findings provide support for the adequacy of the language descriptions found in linguistic typology. Additionally, they open the way for more compact and universal models of word order that can be especially beneficial for machine translation between language pairs with little parallel data. And finally, our results suggest that target-language typological features could play a key role in building better preordering models.

[7] http://www.statmt.org/

Acknowledgements

We thank the three anonymous reviewers for their constructive comments and suggestions. This work received funding from EXPERT (EU FP7 Marie Curie ITN nr. 317471), NWO VICI grant nr. 277-89-002, DatAptor project STW grant nr. 12271 and QT21 project (H2020 nr. 645452).

References

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. OpenFst: A general and efficient weighted finite-state transducer library. In *Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA 2007)*, volume 4783 of *Lecture Notes in Computer Science*, pages 11–23. Springer. `http://www.openfst.org`.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March.

Arianna Bisazza and Marcello Federico. 2016. A survey of word reordering in statistical machine translation: Computational models and language phenomena. *Computational Linguistics*, 42(2):163–205, June.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada, June.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press, Cambridge, MA.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA, June.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 531–540, Ann Arbor, Michigan, June.

Bernard Comrie. 1981. *Language Universals and Linguistic Typology: Syntax and Morphology*. Blackwell, Oxford.

Joachim Daiber, Miloš Stanojević, Wilker Aziz, and Khalil Simaan. 2016. Examining the relationship between preordering and word order freedom in machine translation. In *Proceedings of the First Conference on Machine Translation (WMT16)*, Berlin, Germany, August. Association for Computational Linguistics.

Adrià De Gispert, Gonzalo Iglesias, and Bill Byrne. 2015. Fast and accurate preordering for SMT using neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1012–1017, Denver, Colorado, May–June.

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

M. Amin Farajian, Nicola Bertoldi, and Marcello Federico. 2014. Online word alignment for online adaptive machine translation. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 84–92, Gothenburg, Sweden, April.

João Graça, Joana Paulo Pardal, and Luísa Coheur. 2008. Building a golden collection of parallel multi-language word alignments.

Joseph H. Greenberg. 1966. *Universals of Language*. The MIT Press, Cambridge, MA, 2nd edition.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August.

Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2010. Head finalization: A simple reordering rule for sov languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 244–251, Uppsala, Sweden, July.

Laura Jehl, Adrià de Gispert, Mark Hopkins, and Bill Byrne. 2014. Source-side preordering for translation using logistic regression and depth-first branch-and-bound search. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 239–248, Gothenburg, Sweden.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, pages 81–93.

Maxim Khalilov and Khalil Sima'an. 2012. Statistical translation after source reordering: Oracles, context-aware models, and empirical analysis. *Natural Language Engineering*, 18(4):491–519.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, volume 5, pages 79–86.

Uri Lerner and Slav Petrov. 2013. Source-side classifier preordering for machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP '13)*.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72, Edinburgh, Scotland, UK., July.

Graham Neubig, Taro Watanabe, and Shinsuke Mori. 2012. Inducing a discriminative parser to optimize machine translation reordering. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 843–853, Jeju Island, Korea, July.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Robert Östling. 2015. Word order typology through multilingual word alignment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 205–211, Beijing, China, July.

Sebastian Padó and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1161–1168, Sydney, Australia, July.

Miloš Stanojević and Khalil Sima'an. 2015. Reordering grammar induction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 44–54, Lisbon, Portugal, September.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Edo S van der Poort, Marek Libura, Gerard Sierksma, and Jack A.A van der Veen. 1999. Solving the k-best traveling salesman problem. *Computers & Operations Research*, 26(4):409 – 425.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with large-scale neural language models improves translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1387–1392, Seattle, Washington, USA, October.

Fei Xia and Michael McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP 2008 Workshop on NLP for Less Privileged Languages*, pages 35–42, Hyderabad, India. International Institute of Information Technology.