Multilingual Unsupervised NMT using Shared Encoder and Language-Specific Decoders

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Patna
Patna, India
`{sukanta.pcs15,kamal.pcs17,asif,pb}@iitp.ac.in`

Abstract

In this paper, we propose a multilingual unsupervised NMT scheme which jointly trains multiple languages with a shared encoder and multiple decoders. Our approach is based on denoising autoencoding of each language and back-translating between English and multiple non-English languages. This results in a universal encoder which can encode any language participating in training into an interlingual representation, and language-specific decoders. Our experiments using only monolingual corpora show that multilingual unsupervised model performs better than the separately trained bilingual models achieving improvement of up to 1.48 BLEU points on WMT test sets. We also observe that even if we do not train the network for all possible translation directions, the network is still able to translate in a many-to-many fashion leveraging encoder's ability to generate interlingual representation.

1 Introduction

Neural machine translation (NMT) (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014; Bahdanau et al., 2015) has become a dominant paradigm for machine translation achieving state-of-the-art results on publicly available benchmark datasets. An effective NMT system requires supervision of a huge amount of high-quality parallel data which is not easily available for many language pairs. In absence of such huge amount of parallel data, NMT systems tend to perform poorly (Koehn and Knowles, 2017). However, NMT without using any parallel data such as bilingual translations, bilingual dictionary or comparable translations, has recently become reality and opened up exciting opportunities for future research (Lample et al., 2018; Artetxe et al., 2018; Yang et al., 2018). It completely eliminates the need of any kind of parallel data and depends heavily on cross-lingual embeddings and iterative back-translations (Sennrich et al., 2016) between the source and target language using monolingual corpora. On the architectural point of view, the approaches combine one encoder and one (Lample et al., 2018) or two (Artetxe et al., 2018) decoders. In supervised NMT settings, combining multiple languages to jointly train an NMT system has been found to be successful in improving the performance (Dong et al., 2015; Firat et al., 2016; Johnson et al., 2017). However, to the best of our knowledge, this is the very first attempt which aims at combining multiple languages in an unsupervised NMT training.

To translate between many languages using bilingual version of unsupervised NMT, we require an encoder and one (Lample et al., 2018) or two (Artetxe et al., 2018) decoders for each pair of languages. However, we may not need separate decoders depending on the source language. We can train source-independent, target-specific decoders, wherein each decoder will take an intermediate representation of a source sentence obtained from the shared encoder to translate into their corresponding language. Also, to translate in many-to-many direction for n languages using bilingual unsupervised NMT (Artetxe et al., 2018), we may need n autoencodings and $n * (n - 1)$ back-translations in each iteration during training.

In this work, we propose to combine multiple languages in an unsupervised NMT training using a shared-encoder and language-specific decoders through one source to many targets and many targets to one source translations. Our proposed approach needs only $2 * (n - 1)$ back-translations in each iteration during training. Specifically, we train an NMT system, using only monolingual corpora, for 6 translation directions using 4 languages (English, French, German and Spanish) to perform translation in 12 directions. We take En-

glish as the anchor language and map three non-English languages' embeddings into the English embedding space. We train the network to denoise all the four languages and back-translate between English and non-English languages. We evaluate on *newstest13* and *newstest14* using BLEU (Papineni et al., 2002) score. We find that the multilingual model outperforms the bilingual models by up to 1.48 BLEU points. We also find that the network learns to translate between the non-English (French, German and Spanish) language pairs as well even though it does not explicitly see these pairs during training. To translate between a non-English language pair, no modification to the network is required at inference time. We also evaluate the performance of the non-English language pairs and achieve a maximum BLEU score of 13.92.

The key contributions of our current work are as follows: (i) we propose a strategy to train multilingual unsupervised NMT for one source to many targets and many targets to one source translations; (ii) we empirically show that jointly training multiple languages improves separately trained bilingual models; and (iii) we also show that without training the network for many-to-many translations, the network can translate between all the languages participating in the training.

2   Related Work

Training multiple languages using a single network is a well known approach in NMT. All the previous works in this line were carried out by using parallel data only. Dong et al. (2015) introduced one-to-many translation using a single encoder for the source language and a decoder for each target language. Firat et al. (2016) proposed multi-way multilingual NMT using multiple encoders and decoders with a single shared attention mechanism. Johnson et al. (2017) came up with a simpler but effective approach that needed only a single encoder and a single decoder, in which all the parallel data were merged into a single corpus after appending some special tokens at the beginning of each sentence. Our multilingual unsupervised translation approach is inspired by Artetxe et al. (2018). We use single encoder which is shared by all languages and a decoder for each language.

3   Background

In this section, we briefly describe the basic unsupervised NMT model as proposed in Artetxe et al. (2018). The architecture has one shared encoder and two language specific decoders, and uses following two strategies to train the NMT system in an unsupervised manner:

Denoising Autoencoding: The shared encoder takes a noisy (noise through random swaps between two adjacent words) sentence in a given language, initialized with cross-lingual embeddings, encodes into an intermediate representations, and the decoder of that specific language reconstructs the original sentence from that intermediate representations.

Back-translation: Training strategy with denoising involves one language at a time, thus it is nothing more than a copying task. In order to perform actual translation without violating the constraint of using nothing but monolingual corpora, back-translation approach is adapted to generate synthetic parallel sentences. At first, for a given sentence in one language, authors (Artetxe et al., 2018) use the system in inference mode to translate it in another language using greedy decoding. Then, the system is trained to predict the original sentence from this synthetic sentence.

4   Proposed Approach

Our proposed approach comprises mainly two steps: in the first step, we map multiple languages into a shared latent space through cross-lingual embedding mapping, and in the second step, using the shared representation we train NMT using only monolingual corpora with the help of a shared encoder and language-specific decoders through denoising and back-translation.

4.1   Cross-lingual Embedding

For creating cross-lingual embedding, we follow the work by Conneau et al. (2018), which is a fully unsupervised approach to aligning monolingual word embeddings and is based on the existing work of Mikolov et al. (2013). At first, two monolingual embedding spaces X and Y are learned. Then using adversarial training (Ganin et al., 2016), a translation matrix W is learned to map X into Y. A discriminator is trained to discriminate between WX and Y, while W is trained

to prevent the discriminator from doing so by making WX and Y as similar as possible. Using W, a small bilingual dictionary of frequent words is learned. A new translation matrix W that translates between X and Y space is induced by solving the Orthogonal Procrustes problem:

$$W^* = argmin_W ||WX - Y||_F = UV^T \quad (1)$$
$$\text{s.t} \quad WW^T = I, U\Sigma V^T = SVD(YX^T) \quad (2)$$

This step can be iterated multiple times by using new W to extract new translation pairs. New translation pairs between the two languages are produced using cross-domain similarity local scaling (CSLS) (Conneau et al., 2018).

4.2 Multilingual Embeddings

In general, for n languages, we choose one language L_1 as anchor to map other $n - 1$ languages into its embedding space. To do so, we first train monolingual word embeddings for each of n languages. Then one by one, we map each of $n - 1$ languages' embedding into embedding space of L_1. In our experiments, we consider 4 languages, namely English, French, Spanish and German. We create three cross-lingual embeddings for French, Spanish, and German by keeping English embedding fixed.

4.3 Multilingual NMT Training

NMT systems are ideally trained to predict a target sentence given a source sentence. However, in case of unsupervised version of NMT training, we only have monolingual corpora. In absence of a true source-target pair, we depend on synthetic source-target pair having a authentic monolingual sentence at the target side and synthetic equivalent of target at the source side.

Our proposed multilingual unsupervised NMT training strategy is inspired by a recent work of Artetxe et al. (2018), which has mainly two steps, *viz.* (i) denoising autoencoding of the sentences of source and target; and (ii) back-translation between source and target. For n languages $L_1, L_2, ..., L_n$, in each iteration, we perform denoising of n languages, back-translation from L_1 to the other $n - 1$ languages, and back-translation of $n - 1$ languages to L_1. Figure 1 shows the block-diagrammatic representation. In our experimental setting, we have 4 languages and L_1 is English. In denoising autoencoding step, sentences in one language are corrupted by some random

shuffle of words and the decoder is trained to predict the original sentences. In back-translation step, to train the system for a source-to-target direction, first a target sentence is translated to a source sentence using the system in inference mode (using the shared encoder and the source language decoder) to generate pseudo source-target parallel sentence and then this pseudo parallel sentence is used to train the network for source-to-target direction. Similarly for a target-to-source training, the process is analogous to the above approach.
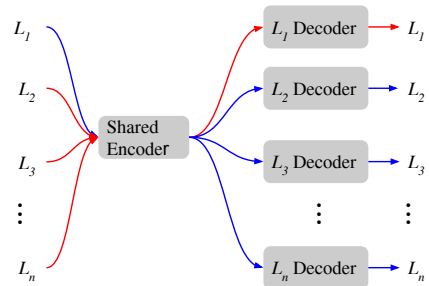

Figure 1: Block diagrammatic view of the proposed network. The shared encoder and decoders of each language are 2-layered bidirectional GRUs. In each iteration of the training: 1. we denoise all languages (L_1, L_2, L_3, ..., L_n); 2. back-translate from each L_i to L_1 as shown using red arrows; 3. back-translate from L_1 to each L_i as shown using blue arrows, where $i \in \{2, 3, ..., n\}$.

5 Datasets and Experimental Setup

5.1 Datasets

We use monolingual English, French, and German news corpora from WMT 2014[1] (Bojar et al., 2014) and Spanish from WMT 2013[2] (Bojar et al., 2013) for the experiments. The number of tokens for English, German, French and Spanish are 495.5, 622.6, 224.3 and 122.9 millions, respectively. For English-{French, German}, we use *newstest2013* and *newstest2014*, and for English-Spanish, we use *newstest2013*. We do not use any parallel data to train, or development set to tune a model. We tokenize and truecase the data using Moses tokenizer[3] and truecaser scripts.

---

[1]http://www.statmt.org/wmt14/translation-task.html
[2]http://www.statmt.org/wmt13/translation-task.html
[3]https://github.com/moses-smt/mosesdecoder/blob/RELEASE-3.0/scripts/tokenizer/tokenizer.perl

5.2 Experimental Setup

Monolingual embeddings are trained using fast-Text[4] using the skip-gram model with vector dimension of 300. For other hyperparameters, we keep default values of fastText (Bojanowski et al., 2017). After getting monolingual embedding for each language, we map every non-English embedding into the embedding space of English using the cross-lingual embedding mapping code MUSE[5] by Conneau et al. (2018). For mapping, we use no bilingual data. We implement the proposed multilingual NMT architecture using Py-Torch[6], and is based on the implementation of Artetxe et al. (2018). The encoder and decoders are 2-layered bidirectional gated recurrent units (Cho et al., 2014). We keep the maximum sentence length to 50 tokens. For training, we keep embedding dimension of 300 and hidden dimension of 600, vocabulary size 50K, learning rate 0.0002 with Adam optimizer (Kingma and Ba, 2015). As we do not use any development set, we run all the models (bilingual as well as multilingual) for 200k iterations keeping batch size of 50 sentences, and take the final models for evaluation.

6 Results and Analysis

We train bilingual models for English↔{French, German, Spanish} as the baselines following Artetxe et al. (2018). We present the BLEU score for each translation direction using bilingual and multilingual models in Table 1. From Table 1, we observe that proposed multilingual model outperforms the separately trained bilingual models for all translation directions on both test sets with a maximum improvement of 1.48 BLEU points for for Spanish to English on *newstest2013*. As the parameters are shared at only encoder side and a separate decoder is used for each target language, multilingual training provides an improved performance for all the language pairs without loosing their own linguistic characteristics.

Though, for one translation direction (En→Fr), the improvement on *newstest2014* is only 0.12 BLEU points. The proposed method is still useful as our method shows consistent improvements over all the baseline models. In supervised multilingual NMT, specifically for one-to-many translation directions, this consistency is absent in some

existing works (Dong et al., 2015; Firat et al., 2016; Johnson et al., 2017). However, in this work, we find that using shared encoder with fixed cross-lingual embedding improves performance in all the translation directions. Though, it may not be fair to compare this unsupervised approach with the supervised ones, but this suggests that supervised multilingual NMT can be improved with cross-lingual embeddings. We leave it for future work.

We also study the outputs produced by the different models. We find that multilingual models are better than bilingual models at lexical selection. For example, French words *préparation* and *payons* are translated as *build-up* and *owe* by bilingual model. However, the correct translations *preparation* and *pay* are generated by the multilingual model. For more examples and the quality of outputs, refer to Table 3 in Appendix A.

System	newstest2013			newstest2014		
	Base	Multi	▲	Base	Multi	▲
Fr→En	13.81	14.47	+0.66	14.98	15.76	+0.78
Es→En	13.97	15.45	+1.48	-	-	-
En→Fr	13.28	13.71	+0.43	14.57	14.69	+0.12
En→Es	14.01	14.82	+0.81	-	-	-
De→En	11.30	11.94	+0.64	10.48	11.21	+0.73
En→De	7.24	8.09	+0.85	6.24	6.77	+0.53

Table 1: BLEU scores on *newstest2013* and *newstest2014*. ▲ shows improvements over bilingual models. Spanish (Es) is not part of the *newstest2014* test set. **Base**: Baseline. **Multi**: Multingual

6.1 Translation between Unseen Language Pairs

In Table 2, we show the results of the language pairs never seen explicitly during training. During training, we only back-translate between English and non-English (Spanish, French, German) languages, but the network learns to translate between the non-English language pairs as well. For example, to translate from Spanish to French, we encode a Spanish sentence and the encoded output of the encoder is decoded by the French decoder. For evaluation, we use the *newstest2013*[7] test set for Spanish-French, Spanish-German, and French-German language pairs. From Table 2, we see translations between French and Spanish achieve very encouraging BLEU scores of 13.87 and 13.92, and pairs involving German achieve

---

moderate BLEU score of up to 7.40 considering the fact that the network is not trained for these pairs. For sample outputs, refer to Table 4 in Appendix A.

→	Es	Fr	De
Es	-	13.92	4.78
Fr	13.87	-	4.59
De	7.40	6.78	-

Table 2: BLEU scores of translation between non-English languages on *newstest2013*. Consider rows are source and columns are target. The network is not trained for these language pairs and still it is possible to translate between these pairs by using the shared encoder and language specific decoders.

6.2 Interlingual Representations

Though the network is not trained for many-to-many translation direction, it is still able to translate in all directions. In multilingual training, the encoder is shared by all the languages while each language has a separate decoder. The hidden vectors generated by the shared encoder is consumed by a language-specific decoder to generate the translation in that specific language. The network learns to translate between the non-English languages as well, though the network is not trained to do so. It may happen that the encoder generates an interlingual representation from which a language-specific decoder is able to generate the translation. To see if the encoded representations share any pattern, we project them using t-SNE[8] (Maaten and Hinton, 2008) for some sentences in all the four languages. From the projection as shown in Figure 6.2, we see that there are well-formed clusters, each representing a sentence in four languages. It means that for a sentence, the shared encoder generates approximately the same hidden contexts for all the four languages.

7 Conclusion

In this paper, we propose a multilingual unsupervised NMT framework to jointly train multiple languages using a shared encoder and language-specific decoders. Our approach is based on denoising autoencoding of all languages and back-translating between English and non-English languages. Our approach shows consistent improvement over the baselines in all the translation di-

---

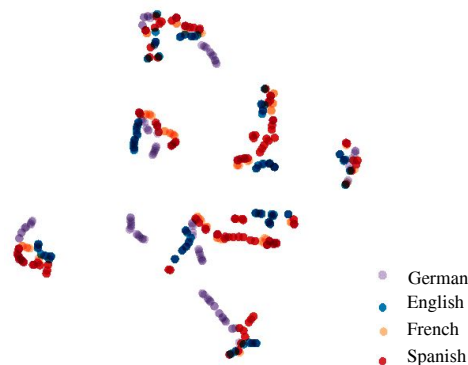[8]https://projector.tensorflow.org

Figure 2: t-SNE projection of hidden vectors obtained from the shared encoder for some sentences in four languages. Each cluster indicates one sentence in four languages. Dots are the words in a sentence. Color represents the languages.

rections with a maximum improvement of 1.48 BLEU points. We also observe that the network learns to translate between unseen language pairs. This is due to the ability of the shared encoder in our proposed network to generate language-independent representation. In future, we would like to explore other languages with diverse linguistic characteristics.

Acknowledgments

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representation (ICLR 2015)*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and

Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics (TACL)*, 5:135–146.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT 2013)*, pages 1–44.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the ninth workshop on statistical machine translation (WMT 2014)*, pages 12–58.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-Task Learning for Multiple Language Translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015) (Volume 1: Long Papers)*, pages 1723–1732.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*, pages 866–875.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-Adversarial Training of Neural Networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics (TACL)*, 5:339–351.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1700–1709.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representation (ICLR 2015)*.

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised Machine Translation using Monolingual Corpora Only. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation. *arXiv preprint arXiv:1309.4168*.

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 meeting on association for computational linguistics (ACL 2002)*, pages 311–318.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016) (Volume 1: Long Papers)*, pages 86–96.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of Advances in neural information processing systems (NIPS 2014)*, pages 3104–3112.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised Neural Machine Translation with Weight Sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 46–55.

A Sample Outputs

We present sample outputs, generated by bilingual and proposed multilingual models, in Table 3. We find that multilingual models are better at lexical selection (see the underlined words in Table 3). Table 4 shows sample outputs on *news2013* for unseen language pairs.

Source	Reference	Bilingual	Multilingual
	French→English		
La préparation à gérer une classe dans un contexte nord-américain, québécois.	Preparation to manage a class in a North-American and Quebec context.	The build-up to manage a class in a Australian, Australian.	The preparation to handle a class in a Latin American context.
Il va y avoir du changement dans la façon dont nous payons ces taxes.	There is going to be a change in how we pay these taxes.	There will be the change in the course of whom we owe these bills.	There will be the change in the way we pay these taxes.
	German→English		
Auch diese Frage soll letztlich Aufschluss darüber geben, welche Voraussetzungen es für die Entstehung von Leben gibt.	This question should also provide information regarding the preconditions for the origins of life.	This question will also ultimately give clues about what there are for the evolution of life.	This question will ultimately give clues to how there is conditions for the emergence of life.
Ihm werde weiterhin vorgeworfen, unerlaubt geheime Informationen weitergegeben zu haben.	He is still accused of passing on secret information without authorisation.	Him will continue to be accused of stealing unlawful information.	Him would continue to be accused of illegally of leaking secret information.
	Spanish→English		
Los estudiantes, por su parte, aseguran que el curso es uno de los más interesantes.	Students, meanwhile, say the course is one of the most interesting around.	The students, by their part, say the practice is one of the most intriguing.	The students, by their part, say the course is one of the most interesting.
No duda en contestar que nunca aceptaría una solicitud de una persona desconocida.	He does not hesitate to reply that he would never accept a request from an unknown person.	No doubt ever answering doubt it would never accept an argument an unknown person.	No doubt in answer that he would never accept a request of a unknown person.

Table 3: Sample outputs for bilingual and multilingual models on *newstest2013* test set. We observe that the multilingual model is better at lexical selection. Underlined words are some examples of our observation.

Source	Reference	Multilingual
	French→Spanish	
Les dirigeants républicains justifièrent leur politique par la nécessité de lutter contre la fraude électorale.	Los dirigentes republicanos justificaron su política por la necesidad de luchar contra el fraude electoral.	Los dirigentes republicanos <OOV> su política por la necesidad de luchar contra la fraude electoral.
	French→German	
Chacun sait que son livre fait partie de cet édifice.	Jeder weiß , dass sein Buch Teil dieses Gebäudes ist.	Jeder weiß , dass sein Buch Teil seines Gebäudes machte.
	German→Spanish	
Seine Zahlen auf Ebene der internationalen Turniere sind beeindruckend.	Sus números a nivel de torneos internacionales son impresionantes.	Sus cifras sobre el nivel de torneos internacionales son impresionantes.
	German→French	
Diese Einschränkungen sind nicht ohne Folgen.	Ces restrictions ne sont pas sans conséquence.	Ces restrictions ne sont pas sans conséquences.
	Spanish→German	
Tomemos por caso la elección directa del presidente , que ha sido un logro de la presión pública.	Nehmen Sie nur einmal die direkte Wahl des Präsidenten, die ein Verdienst des öffentlichen Drucks war.	Nehmen Sie über die direkte Wahl des Präsidenten, hat dies ein Erfolg ein der öffentlichen Druck.
	Spanish→French	
Las inversiones en la materia superan los 1.5 billones de dólares.	Les investissements dans ce domaine dépassent les 1,5 milliards de dollars.	Les investissements dans la matière dépassent les 1,5 milliards de dollars.

Table 4: Sample outputs for unseen language pairs on *newstest2013* test set.