Identifying Nuanced Dialect for Arabic Tweets with Deep Learning and Reverse Translation Corpus Extension System

Rawan Tahssin **Youssef Kishk** **Marwan Torki**

Faculty of Engineering, Alexandria University
{eng-rowan.tarek1520, es-Youssef.Aly20}@alexu.edu.eg
mtorki@alexu.edu.eg

Abstract

In this paper, we present our work for the NADI Shared Task (Abdul-Mageed et al., 2020): Nuanced Arabic Dialect Identification for Subtask-1: country-level dialect identification. We introduce a Reverse Translation Corpus Extension Systems (RTCES) to handle data imbalance along with reported results on several experimented approaches of word and document representations and different models architectures. The top scoring model was based on the Transformer-based Model for Arabic Language Understanding (AraBERT) (Antoun et al., 2020), with our modified extended corpus based on reverse translation of the given Arabic tweets. The selected system achieved a macro average F1 score of 20.34% on the test set, which places our team CodeLyoko as the 7th out of 18 teams in the final ranking Leaderboard.

1 Introduction

Arabic is one of the most complex languages, which presents significant challenges for natural language processing. Like other languages, Arabic has a number of dialectal varieties. Many of these varieties of Arabic have started being widely represented in the written form with the emergence of social media. Arabic language speakers use Modern Standard Arabic (MSA) as the official language in very formal situations , while they use an Arabic Dialect for everyday conversation. Dialect identification is the task of detecting the source variety of a given text or speech segment automatically. Previous work on Arabic dialect identification has focused on country-level varieties such as the Arabic Fine-Grained Dialect Identification task (MADAR) co-located with The Fourth Arabic Natural Language Processing Workshop (WANLP 2019) (Bouamor et al., 2019). The classification task remains challenging as it covers 21 different Arabic dialects with high similarities and common words. Throughout the paper, we propose an approach for data balancing and augmentation without using any external manually-labelled data sets. We also report the different systems that were experimented in feature extraction and word embedding such as Term Frequency-Inverse Document Frequency (TF-IDF) and fastText (Mikolov et al., 2018). For the tweets classification, Logistic Regression, Bi-directional Long Short Term Memory (LSTM) (Graves and Schmidhuber, 2005) and AraBERT were evaluated to reach the top score.

2 Data

2.1 Dataset Description

The data used in all of the proposed systems is based on the official available dataset for Subtask-1 with no external data sets used. Table 1 shows the distribution of available data across different sets.

	Train	Dev	Test
# Tweets	21000	4957	5000

Table 1: Available dataset distribution

Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 288–294
Barcelona, Spain (Online), December 12, 2020

The available data is covering the dialects of 21 Arab countries with the distribution in Figure 1 for the training set.

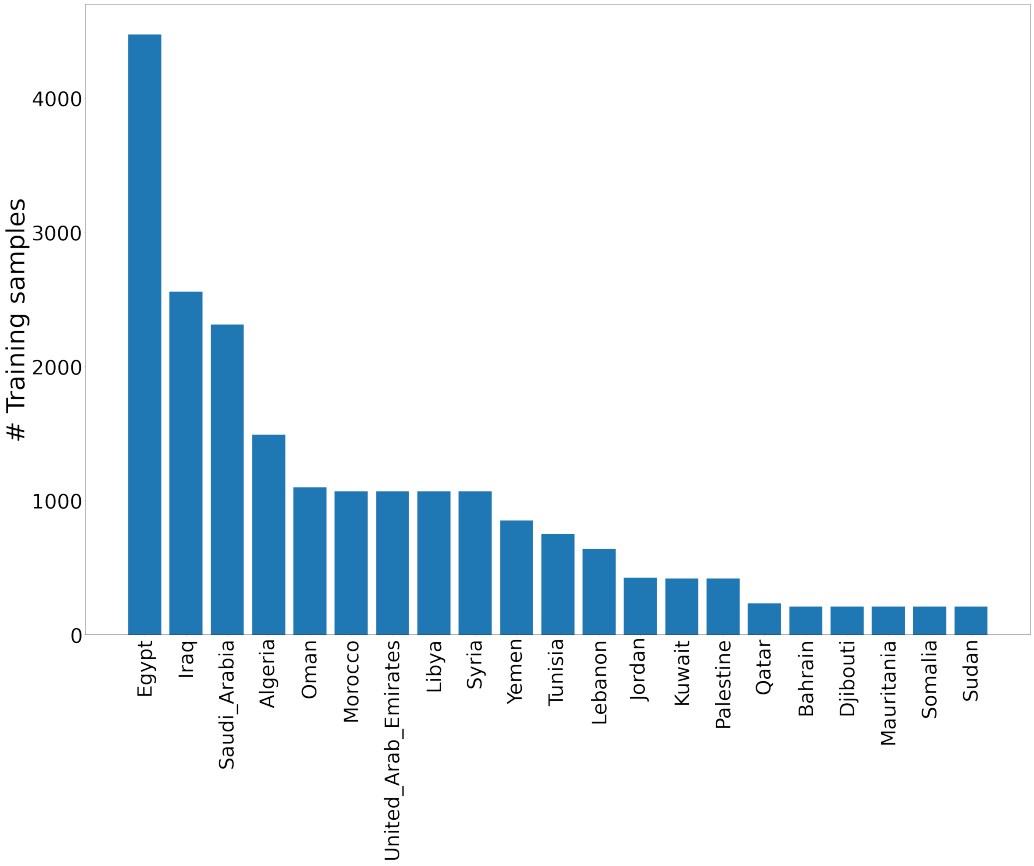


Figure 1: Training set classes distribution

2.2 Data Preprocessing

As the available dataset was collected from general tweets, thus, it required a generic transformation before its usage as an input to our systems. A pre-processing phase (Shoukry and Rafea, 2012) was implemented to remove punctuation, vowel elongation, URLs, mentions and diacritization. English and French words along with emojis were kept to be used as features.

2.3 Reverse Translation Corpus Extension System (RTCES)

The presence of class imbalance between countries labels within the training corpus was highly noticed as shown in Figure 1. Accordingly, a reverse translation approach was taken to handle this imbalance and augmentation. The approach consisted of a number of steps, starting from pre-processing module till the new generated sentence as shown in Figure 2.

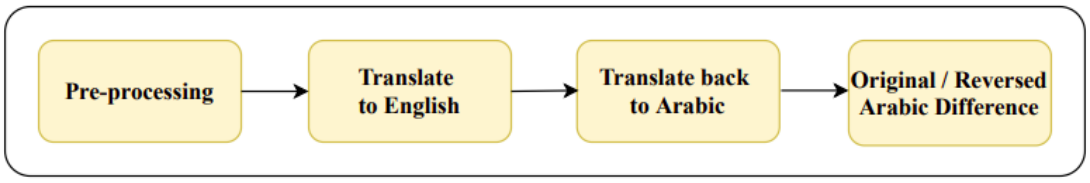


Figure 2: Reverse Translation Corpus Extension System

First, the entire pre-processed data is translated to English using Google's NMT API (Wu et al., 2016) to provide an equivalent corpus in English. The next step is the reverse translation of the newly created

English corpus to translate back the whole data to Arabic. As a final step, the extraction of the difference between the original Arabic tweet and the newly generated Arabic tweet from the reverse translation;to create a new sentence. An example of the steps applied on a tweet form the corpus is shown in Table 2.

Steps	**Output Sentence**
Step 1 (Pre-processing)	فارقه كتير اوى راجع فرق الزمن وظروف الحياة وهتعرف
Step 2 (English translated)	Too much difference, review the difference in time and circumstances of life, and you will know
Step 3 (Reverse Translated)	الكثير الاختلاف راجع الفرق الوقت وظروف الحياه وستعرف
Step 4 (Sentence Difference)	فارقه كتير اوي فرق الزمن وهتعرف

Table 2: RTCES applied on an Example from training set

One of the main observations that made this approach interesting, was the ability to filter out parts of the words based on Modern Standard Arabic (MSA) and keep the words reflecting the Arabic dialects of each country. This filtering served the purpose of our task and allowed the formation of new sentences for the classes with lower occurrences.

For our explored document and word representations as well as the classification model approaches, an extended corpus has been used. The new extended corpus consisted of the initial training set, added to it the new sentences generated from RTCES excluding the classes with higher occurrences (Egypt, Iraq and Saudi Arabia) to provide a more balanced distribution of classes. Figure 2 shows the complete system architecture. Finally, our extended training corpus is composed of 32,417 training sentences whose distribution is shown in Figure 3.

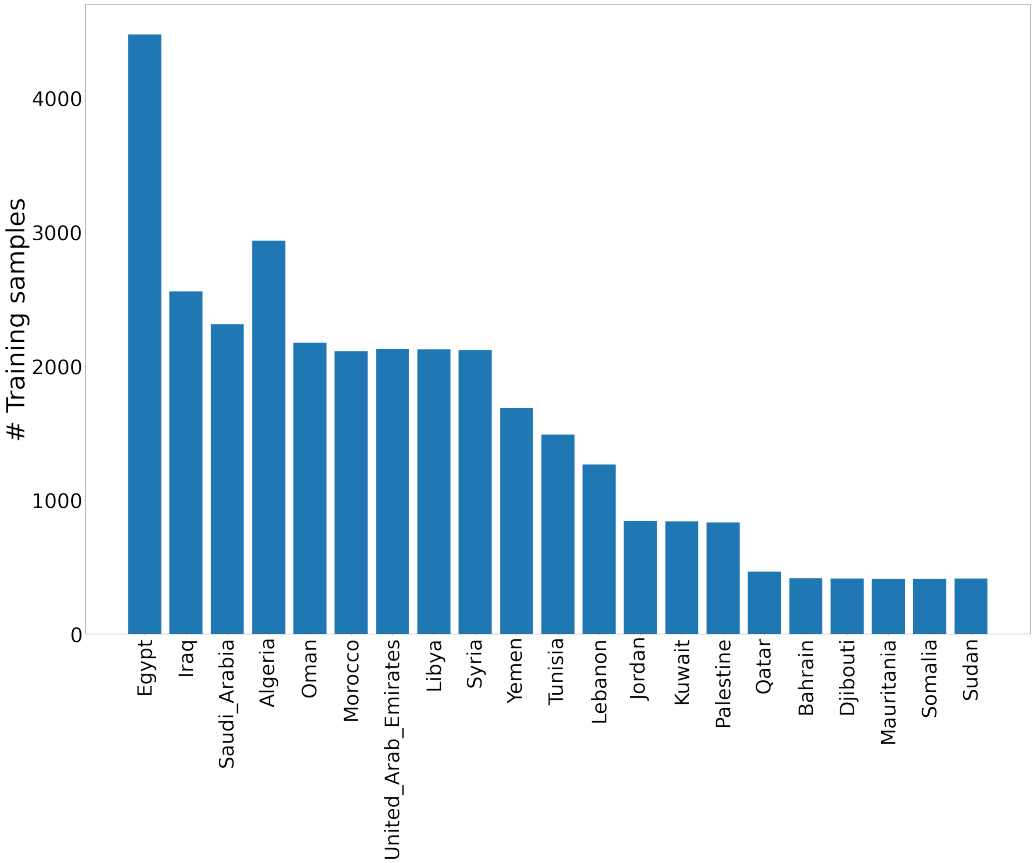


Figure 3: Modified corpus classes distribution

3 Systems

The aim of this work is to design a system that can classify 21 different Arabic dialects efficiently. In this section, we describe some selected experimented approaches and architectures out of various attempts to reach the goal of NADI Shared Task (Abdul-Mageed et al., 2020). All these approaches were applied to the output of the RTCES and their results were reported.

3.1 TF-IDF with Logistic Regression Model

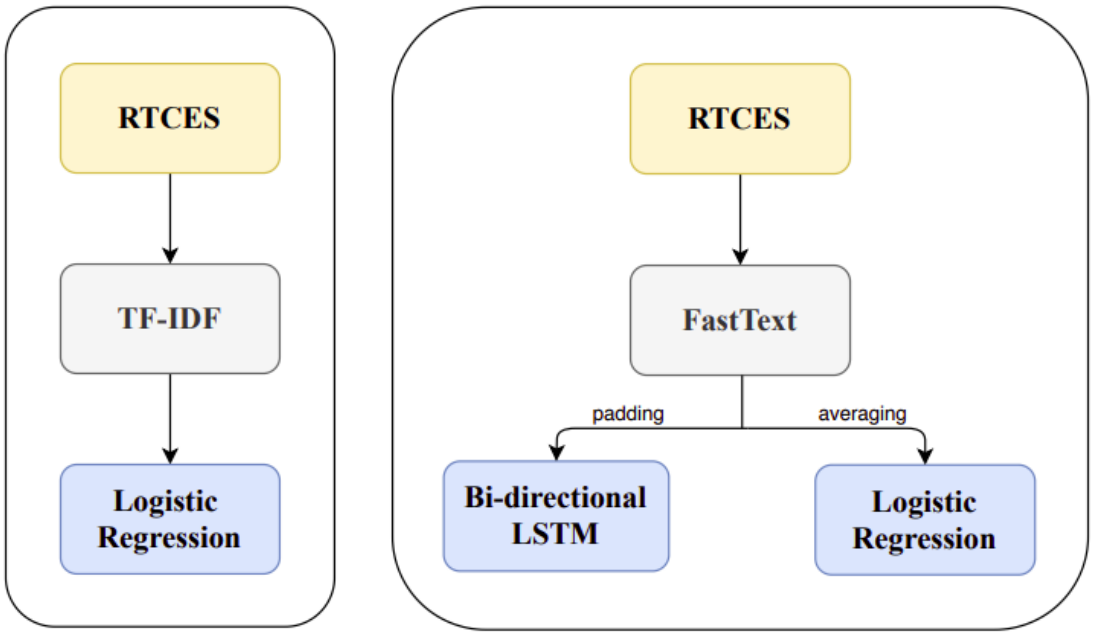
Extracted features using TF-IDF and Logistic Regression Model from (Pedregosa et al., 2011) with the tuned parameters from k-fold cross validation as shown in Figure 4 (a).

3.2 FastText averaged word embeddings with Logistic Regression Model

fastText (Mikolov et al., 2018) is a deep learning-based approach for efficient learning of word representations. It was selected as it returns a vector representation to non-existing words in its vocabulary by computing the closest word based on the character level n-gram. The implementation of Gensim (Řehůřek and Sojka, 2010) was used. We trained fastText over the extra data corpus of 10M unlabelled tweets, after the pre-processing phase shown in section 2.1; to obtain efficient vector representations for each word. The returned vectors were averaged to obtain a representation suitable for the Logistic Regression input, the obtained result are shown in Figure 4(b).

3.3 FastText word embeddings with Bi-directional LSTM

The word vectors and labels were sequenced, padded and passed to a bi-directional LSTM (Graves and Schmidhuber, 2005) model which is able to exploit previous and future context of a given word and calculated the loss from the concatenation of the last hidden layer in both directions as shown in Figure 4 (b). The bi-directional LSTM model was built using Keras (Chollet, 2015).



(a) TF-IDF based system Architecture (b) fastText-based system Architecture

Figure 4: Implemented approaches Architectures

3.4 Fined tuned AraBERT

Our top submission model was based on AraBERT (Antoun et al., 2020), which is an Arabic language model based on Multilingual Bidirectional Encoder Representations from Transformers (BERT) trained on 70M sentences or 23GB of Arabic text with 3B words from a collection of publically available large

scale raw Arabic text. We applied first a tokenization and segmentation phase using Fast and Accurate Arabic Word Segmenter (Farasa) (Abdelali et al., 2016) on the extended corpus described in section 2.3. The pre-trained AraBERT model is fine-tuned with one additional output layer of 21 classes, then the model is trained on our corpus which reaches a 20.34 test-set F1 score.

4 Results and Discussion

Interesting observations at the beginning of conducting our work showed that TF-IDF and a simple Logistic Regression model performed better than NN-based models. However, with more experiments, NN-based models outperformed it. The results of the approaches described in the previous section are shown in Table 3.

Model	Macro Avg. F1
TF-IDF + Logistic Regression	17.34
fastText + Logistic Regression	14.24
fastText + Bi-directional LSTM	18.33
Fine-tuned AraBERT	20.73

Table 3: Results (in %) on Dev. set

The fine-tuned AraBERT score was the highest. Accordingly, its predictions were selected to be our final submission for the NADI Shared Task 2020 Subtask-1. Moreover, one of the observations of our approach was that our results on the development set are quite close to those on the test set. This indicates that no over-fitting took place as shown in Table 4. The final Macro average F1, accuracy, precision and recall scores for the best-performing model were addressed in section 3.4.

Subtask #	Set	F1-score	Accuracy	Precision	Recall
1	Test	20.34	36.26	27.83	20.56
1	Dev	20.73	36.59	29.29	19.97

Table 4: Final Submitted results (in %) of AraBERT on Test and Dev. sets

4.1 Two-level Hierarchical Prediction Structure

In an attempt to improve the reported results, the 21 countries were clustered to 5 super classes inspired by (Fares et al., 2019) according to the origin of the dialect labels as shown in Table 5.

Maghreb	Egypt_Sudan	Gulf	Levant	Others
Morocco	Egypt	Iraq	Jordan	Somalia
Algeria	Sudan	UAE	Syria	Djibouti
Tunisia		Saudi Arabia	Palestine	
Libya		Qatar	Lebanon	
Mauritania		Kuwait		
		Bahrain		
		Yemen		
		Oman		

Table 5: Two-level classes distribution

A two-level hierarchical prediction structure inspired by (de Francony et al., 2019) was implemented. The predicted labels from the first five super classes level are passed to the second level to output the prediction of the corresponding countries in each of the five origins. This is a general structure that can be used on different models. However, we reported its results on the first level classes using system

explained in Section 3.1 using TF-IDF with Logistic Regression Model as shown in Figure 5. The result of the second level were close to that obtained from the system explained in Section 3.2 which is 14.241%. We aim to enhance this structure and report its results on other models as a future work.

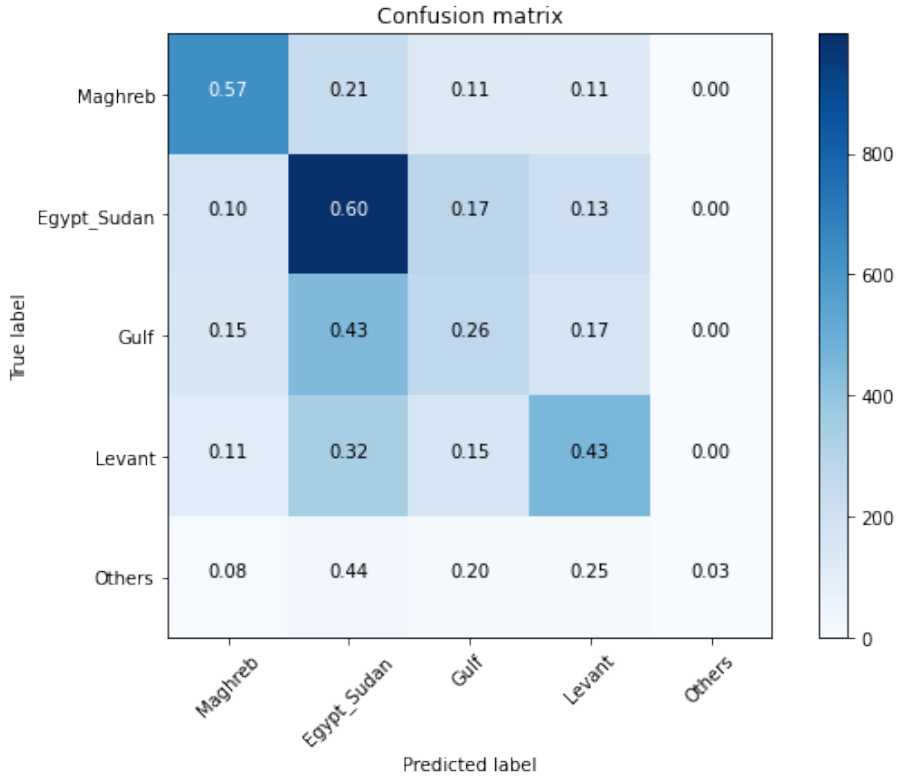


Figure 5: Confusion matrix for structure in 4.1 on first level classes

5 Conclusion

We introduced a reverse Arabic translation solution to handle unbalanced data and small data set, a hierarchical architecture to enhance the efficiency and deal with the 21 classes classification, several neural network based models built on different word and document representations. Future work will include trying to ensemble the mentioned models, enhance the two-level hierarchical prediction structure and exploring the effect of adding a named entity recognition system module for better focus on highly effective words that identifies each country such as places, food, public figures, etc. Moreover, we will examine more data augmentation methods such as suggested in (Fares et al., 2019; Ibrahim et al., 2018; Ibrahim et al., 2020).

Acknowledgements

We would like to thank Ms. Samaa Abdelaal, our language editor for her dedicated work and efforts on this paper.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and H. Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *HLT-NAACL Demos*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy, August. Association for Computational Linguistics.

François Chollet. 2015. keras. `https://github.com/fchollet/keras`.

Gael de Francony, Victor Guichard, Praveen Joshi, Haithem Afli, and Abdessalam Bouchekif. 2019. Hierarchical deep learning for Arabic dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 249–253, Florence, Italy, August. Association for Computational Linguistics.

Youssef Fares, Zeyad El-Zanaty, Kareem Abdel-Salam, Muhammed Ezzeldin, Aliaa Mohamed, Karim El-Awaad, and Marwan Torki. 2019. Arabic dialect identification with deep learning and hybrid frequency based features. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 224–228, Florence, Italy, August. Association for Computational Linguistics.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. 2018. Imbalanced toxic comments classification using data augmentation and deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 875–878, Dec.

Mai Ibrahim, Marwan Torki, and Nagwa El-Makky. 2020. Alexu-backtranslation-tl at semeval-2020 task [12]: Improving offensive language detection using data augmentation and transfer learning. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830, November.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Amira Shoukry and Ahmed Rafea. 2012. Preprocessing egyptian dialect tweets for sentiment mining. 11.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.