Low-Resource Machine Translation Using Cross-Lingual Language Model Pretraining

Francis Zheng, Machel Reid, Edison Marrese-Taylor, Yutaka Matsuo
Graduate School of Engineering
The University of Tokyo
{francis, machelreid, emarrese, matsuo}@weblab.t.u-tokyo.ac.jp

Abstract

This paper describes UTokyo's submission to the AmericasNLP 2021 Shared Task on machine translation systems for indigenous languages of the Americas. We present a low-resource machine translation system that improves translation accuracy using cross-lingual language model pretraining. Our system uses an mBART implementation of FAIRSEQ to pretrain on a large set of monolingual data from a diverse set of high-resource languages before finetuning on 10 low-resource indigenous American languages: Aymara, Bribri, Asháninka, Guaraní, Wixarika, Náhuatl, Hñähñu, Quechua, Shipibo-Konibo, and Rarámuri. On average, our system achieved BLEU scores that were 1.64 higher and CHRF scores that were 0.0749 higher than the baseline.

1 Introduction

Neural machine translation (NMT) systems have produced translations of commendable accuracy under large-data training conditions but are data-hungry (Zoph et al., 2016) and perform poorly in low-resource languages, where parallel data is lacking (Koehn and Knowles, 2017).

Many of the indigenous languages of the Americas lack adequate amounts of parallel data, so existing NMT systems have difficulty producing accurate translations for these languages. Additionally, many of these indigenous languages exhibit linguistic properties that are uncommon in high-resource languages, such as English or Chinese, that are used to train NMT systems.

One striking feature of many indigenous American languages is their polysynthesis (Brinton, 1885; Payne, 2014). Polysynthetic languages display high levels of inflection and are morphologically complex. However, NMT systems are weak in translating "low-frequency words belonging to highly-inflected categories (e.g. verbs)" (Koehn and Knowles, 2017). Quechua, a low-resource, polysynthetic American language, has on average twice as many morphemes per word compared to English (Ortega et al., 2020b), which makes machine translation difficult. Mager et al. (2018b) shows that information is often lost when translating polysynthetic languages into Spanish due to a misalignment of morphemes. Thus, existing NMT systems are not appropriate for indigenous American languages, which are low-resource, polysynthetic languages.

Despite the scarcity of parallel data for these indigenous languages, some are spoken widely and have a pressing need for improved machine translation. For example, Quechua is spoken by more than 10 million people in South America, but some Quechua speakers are not able to access health care due to a lack of Spanish ability (Freire, 2011).

Other languages lack a large population of speakers and may appear to have relatively low demand for translation, but many of these languages are also crucial in many domains such as health care, the maintenance of cultural history, and international security (Klavans, 2018). Improved translation techniques for low-resource, polysynthetic languages are thus of great value.

In light of this, we participated in the AmericasNLP 2021 Shared Task to help further the development of new approaches to low-resource machine translation of polysynthetic languages, which are not commonly studied in natural language processing. The task consisted of producing translations from Spanish to 10 different indigenous American languages.

In this paper, we describe our system designed for the AmericasNLP 2021 Shared Task, which achieved BLEU scores that were 1.64 higher and CHRF scores that were 0.0749 higher than the baseline on average. Our system improves translation accuracy by using monolingual data to improve understanding of natural language before finetuning

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 234–240

for each of the 10 indigenous languages.

2 Methods

2.1 Data

Our model employs two types of data:

1. 13 GB of monolingual data from Bulgarian, English, French, Irish, Korean, Latin, Spanish, Sundanese, Vietnamese, and Yoruba

2. 140 MB of parallel data between Spanish and Aymara, Bribri, Asháninka, Guaraní, Wixarika, Náhuatl, Hñähñu, Quechua, Shipibo-Konibo, and Rarámuri

2.1.1 Monolingual Data

We selected a variety of widely-spoken languages across the Americas, Asia, Europe, Africa, and Oceania for the monolingual data we used during our pretraining, allowing our model to learn from a wide range of language families and linguistic features. These monolingual data were acquired from CC100[1] (Wenzek et al., 2020; Conneau et al., 2020). We use these monolingual data as part of our pretraining, as this has been shown to improve results with smaller parallel datasets (Conneau and Lample, 2019; Liu et al., 2020; Song et al., 2019).

2.1.2 Parallel Data

The parallel data between Spanish and the indigenous American languages were provided by AmericasNLP 2021 (Mager et al., 2021).

We have summarized some important details of the training data and development/test sets (Ebrahimi et al., 2021) below. More details about these data can be found in the AmericasNLP 2021 official repository[2].

Aymara The Aymara–Spanish data came from translations by Global Voices and Facebook AI. The training data came primarily from Global Voices[3] (Prokopidis et al., 2016; Tiedemann, 2012), but because translations were done by volunteers, the texts have potentially different writing styles. The development and test sets came from translations from Spanish texts into Aymara La Paz jilata, a Central Aymara variant.

Bribri The Bribri–Spanish data (Feldman and Coto-Solano, 2020) came from six different sources (a dictionary, a grammar, two language learning textbooks, one storybook, and transcribed sentences from a spoken corpus) and three major dialects (Amubri, Coroma, and Salitre). Two different orthographies are widely used for Bribri, so an intermediate representation was used to facilitate training.

Asháninka The Asháninka–Spanish data[4] were extracted and pre-processed by Richard Castro (Cushimariano Romano and Sebastián Q., 2008; Ortega et al., 2020a; Mihas, 2011). Though the texts came from different pan-Ashaninka dialects, they were normalized using **AshMorph** (Ortega et al., 2020a). The development and test sets came from translations of Spanish texts done by Feliciano Torres Ríos.

Guaraní The Guaraní–Spanish data (Chiruzzo et al., 2020) consisted of training data from web sources (blogs and news articles) written in a mix of dialects and development and test sets written in pure Guaraní. Translations were provided by Perla Alvarez Britez.

Wixarika The Wixarika–Spanish data came from Mager et al. (2018a). The training, development, and test sets all used the same dialect (Wixarika of Zoquipan) and orthography, though word boundaries were not consistent between the development/test and training sets. Translations were provided by Silvino González de la Crúz.

Náhuatl The Náhuatl–Spanish data came from Gutierrez-Vasques et al. (2016). Náhuatl has a wide dialectal variation and no standard orthography, but most of the training data were close to a Classical Náhuatl orthographic "standard." The development and test sets came from translations made from Spanish into modern Náhuatl. An orthographic normalization was applied to these translations to make them closer to the Classical Náhuatl orthography found in the training data. This normalization was done by employing a rule-based approach based on predictable orthographic changes between modern varieties and Classical Náhuatl. Translations were provided by Giovany Martinez Sebastián, José Antonio, and Pedro Kapoltitan.

[1] http://data.statmt.org/cc-100/

[2] https://github.com/AmericasNLP/americasnlp2021/blob/main/data/information_datasets.pdf

[3] https://opus.nlpl.eu/GlobalVoices.php

[4] https://github.com/hinantin/AshaninkaMT

Hñähñu The Hñähñu–Spanish training data came from translations into Spanish from Hñähñu text from a set of different sources[5]. Most of these texts are in the Valle del Mezquital dialect. The development and test sets are in the Ñûhmû de Ixtenco, Tlaxcala variant. Translations were done by José Mateo Lino Cajero Velázquez.

Quechua The training set for Quechua–Spanish data (Agić and Vulić, 2019) came from Jehova's Witnesses texts (available in OPUS), sentences extracted from the official dictionary of the Minister of Education (MINEDU) in Peru for Quechua Ayacucho, and dictionary entries and samples collected and reviewed by Diego Huarcaya. Training sets were provided in both the Quechua Cuzco and Quechua Ayacucho variants, but our system only employed Quechua Ayacucho data during training. The development and test sets came from translations of Spanish text into Quechua Ayacucho, a standard version of Southern Quechua. Translations were provided by Facebook AI.

Shipibo-Konibo The training set of the Shipibo-Konibo–Spanish data (Galarreta et al., 2017) was obtained from translations of flashcards and translations of sentences from books for bilingual education done by a bilingual teacher. Additionally, parallel sentences from a dictionary were used as part of the training data. The development and test sets came from translations from Spanish into Shipibo-Konibo done by Liz Chávez.

Rarámuri The training set of the Rarámuri–Spanish data came from a dictionary (Brambila, 1976). The development and test sets came from translations from Spanish into the highlands Rarámuri by María del Cármen Sotelo Holguín. The training set and development/test sets use different orthographies.

2.2 Preprocessing

We tokenized all of our data together using SentencePiece (Kudo and Richardson, 2018) in preparation for our multilingual model. We used a vocabulary size of 8000 and a character coverage of 0.9995, as the wide variety of languages cover a rich character set.

Then, we sharded our data for faster processing. With our SentencePiece model and vocabulary, we used FAIRSEQ[6] (Ott et al., 2019) to build vocabularies and binarize our data.

2.3 Pretraining

We pretrained our model on the 20 languages described in 2.1 with an mBART (Liu et al., 2020) implementation of FAIRSEQ (Ott et al., 2019). We pretrained on 32 NVIDIA V100 GPUs for three hours.

Balancing data across languages

Due to the large variability in text data size between different languages, we used the exponential sampling technique used in Conneau and Lample (2019); Liu et al. (2020), where the text is resampled according to smoothing parameter α as follows:

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \quad (1)$$

In equation 1, q_i refers to the resample probability for language i, given multinomial distribution $\{q_i\}_{i=1...N}$ with original sampling probability p_i.

As we want our model to work well with the low-resource languages, we chose a smoothing parameter of $\alpha = 0.25$ (compared with $\alpha = 0.7$ used in mBART (Liu et al., 2020)) to alleviate model bias towards the higher proportion of data from high-resource languages.

Hyperparameters

We used a six-layer Transformer with a hidden dimension of 512 and feed-forward size of 2048. We set the maximum sequence length to 512, with a batch size of 1024. We optimized the model using Adam (Kingma and Ba, 2015) using hyperparameters $\beta = (0.9, 0.98)$ and $\epsilon = 10^{-6}$. We used a learning rate of 6×10^{-4} over 10,000 iterations. For regularization, we used a dropout rate of 0.5 and weight decay of 0.01. We also experimented with lower dropout rates but found that a higher dropout rate gave us a model that produces better translations.

2.4 Finetuning

Using our pretrained model, we performed finetuning on each of the 10 indigenous American languages with the same hyperparameters used during pretraining. For each language, we conducted our finetuning using four NVIDIA V100 GPUs for three hours.

[5] https://tsunkua.elotl.mx/about/

[6] https://github.com/pytorch/fairseq

Language	Baseline[1]		Dev[2]		Test1[3]		Test2[4]	
	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF
Aymara (aym)	0.01	0.157	2.84	0.2338	1.17	0.214	1.03	0.209
Bribri (bzd)	0.01	0.058	1.22	0.1203	1.7	0.143	1.29	0.131
Asháninka (cni)	0.01	0.102	0.48	0.2188	0.2	0.216	0.45	0.214
Guaraní (gn)	0.12	0.193	3.64	0.2492	3.21	0.265	3.16	0.254
Wixarika (hch)	2.2	0.126	4.89	0.2093	7.09	0.238	6.74	0.229
Náhuatl (nah)	0.01	0.157	0.3	0.253	0.55	0.239	1.2	0.238
Hñähñu (oto)	0	0.054	0.04	0.1035	2.45	0.152	1.28	0.133
Quechua (quy)	0.05	0.304	1.46	0.3155	2.35	0.332	2.47	0.33
Shipibo-Konibo (shp)	0.01	0.121	0.49	0.176	0.33	0.163	0.71	0.175
Rarámuri (tar)	0	0.039	0.12	0.1163	0.1	0.122	0.06	0.123

[1] Baseline test results provided by AmericasNLP 2021, from a system where the development set was not used for training
[2] Our own results on the development set
[3] Our official test results for our system where the development set was used for training
[4] Our official test results for our system where the development set was not used for training

Table 1: Results

2.5 Evaluation

Using the SacreBLEU library[7] (Post, 2018), we evaluated our system outputs with detokenized BLEU (Papineni et al., 2002; Post, 2018). Due to the polysynthetic nature of the languages involved in this task, we also used CHRF (Popović, 2015) to measure performance at the character level and better see how well morphemes or parts of morphemes were translated, rather than whole words. For these reasons, we focused on optimizing the CHRF score.

3 Results

We describe our results in Table 1. Our test results (**Test1** and **Test2**) show considerable improvements over the baseline provided by AmericasNLP 2021. We also included our own results on the development set (**Dev**) for comparison. The trends we saw in the **Dev** results parallel our test results; languages for which our system achieved high scores in **Dev** (e.g. Wixarika and Guaraní) also demonstrated high scores in **Test1** and **Test2**. Likewise, languages for which our system performed relatively poorly in **Dev** (e.g. Rarámuri, whose poor performance may be attributed to the difference in orthographies between the training set and development/test sets) also performed poorly in **Test1** and **Test2**. This matches the trend seen in the baseline scores.

The baseline results and **Test2** results were both produced using the same test set and by systems where the development set was not used for training. Thus, the baseline results and **Test2** results can be directly compared. On average, our system used to produce the **Test2** results achieved BLEU scores that were 1.54 higher and CHRF scores that were 0.0725 higher than the baseline. On the same test set, our **Test1** system produced higher BLEU and CHRF scores for nearly every language. This is expected, as the system used to produce **Test1** was trained on slightly more data; it used the development set of the indigenous American languages provided by AmericasNLP 2021 in addition to the training set.

If we factor in our results from **Test1** to our **Test2** results, we achieved BLEU scores that were 1.64 higher and CHRF scores that were 0.0749 higher than the baseline on average. Overall, we attribute this improvement in scores primarily to the cross-lingual language model pretraining (Conneau and Lample, 2019) we performed, allowing our model to learn about natural language from the monolingual data before finetuning on each of the 10 indigenous languages.

4 Conclusions and Future Work

We described our system to improve low-resource machine translation for the AmericasNLP 2021 Shared Task. We constructed a system using the mBART implementation of FAIRSEQ to translate from Spanish to 10 different low-resource indigenous languages from the Americas. We demon-

[7] https://github.com/mjpost/sacrebleu

strated strong improvements over the baseline by pretraining on a large amount of monolingual data before finetuning our model for each of the low-resource languages.

We are interested in using dictionary augmentation techniques and creating pseudo-monolingual data to use during the pretraining process, as we have seen improved results with these two techniques when translating several low-resource African languages. We can also incorporate these two techniques in an iterative pretraining procedure (Tran et al., 2020) to produce more pseudo-monolingual data and further train our pretrained model for potentially better results.

Future research should also explore using probabilistic finite-state morphological segmenters, which may improve translations by exploiting regular agglutinative patterns without the need for much linguistic knowledge (Mager et al., 2018a) and thus may work well with the low-resource, polysynthetic languages dealt with in this paper.

References

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

David Brambila. 1976. *Diccionario Raramuri–Castellano (Tarahumara)*. Obra Nacional de la Buena Prensa, México.

D.G. Brinton. 1885. *On Polysynthesis and Incorporation: As Characteristics of American Languages*. McCalla & Stavely.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Rubén Cushimariano Romano and Richer C. Sebastián Q. 2008. Ñaantsipeta asháninkaki birakochaki. diccionario asháninka-castellano. versión preliminar. http://www.lengamer.org/publicaciones/diccionarios/.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, Gustavo A. Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando A. Coto Solano, Ngoc Thang Vu, and Katharina Kann. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Germán Freire. 2011. *Perspectivas en salud indígena: cosmovisión, enfermedad y políticas públicas*. Ediciones Abya-Yala.

Ana-Paula Galarreta, Andrés Melgar, and Arturo Oncevay. 2017. Corpus creation and initial SMT experiments between Spanish and Shipibo-konibo. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 238–244, Varna, Bulgaria. INCOMA Ltd.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Judith L. Klavans. 2018. Computational challenges for polysynthetic languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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, Vancouver. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018a. Probabilistic finite-state morphological segmenter for Wixarika (Huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Manuel Mager, Elisabeth Mager, Alfonso Medina-Urrea, Ivan Vladimir Meza Ruiz, and Katharina Kann. 2018b. Lost in translation: Analysis of information loss during machine translation between polysynthetic and fusional languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 73–83, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Anna Currey, Vishrav Chaudhary, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager, Ngoc Thang Vu, Graham Neubig, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas. In *Proceedings of the The First Workshop on NLP for Indigenous Languages of the Americas*, Online. Association for Computational Linguistics.

Elena Mihas. 2011. *Añaani katonkosatzi parenini, El idioma del alto Perené*. Milwaukee, WI: Clarks Graphics.

John Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2020a. Overcoming resistance: The normalization of an Amazonian tribal language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 1–13, Suzhou, China. Association for Computational Linguistics.

John E. Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020b. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

D.L. Payne. 2014. *Morphological Characteristics of Lowland South American Languages*. University of Texas Press.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. 2016. Parallel Global Voices: a collection of multilingual corpora with citizen media stories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 900–905, Portorož, Slovenia. European Language Resources Association (ELRA).

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: masked sequence to sequence pre-training for language generation. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages

4003–4012, Marseille, France. European Language Resources Association.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.