Modeling Word Importance in Conversational Transcripts: Toward improved live captioning for Deaf and hard of hearing viewers

Akhter Al Amin
aa7510@rit.edu
Rochester Institute of Technology
Rochester, New York, USA

Saad Hassan
sh2513@rit.edu
Rochester Institute of Technology
Rochester, New York, USA

Matt Huenerfauth
matt.huenerfauth@rit.edu
Rochester Institute of Technology
Rochester, New York, USA

Cecilia Ovesdotter Alm
cecilia.o.alm@rit.edu
Rochester Institute of Technology
Rochester, New York, USA

ABSTRACT

Despite the recent improvements in automatic speech recognition (ASR) systems, their accuracy is imperfect in live conversational settings. Classifying the importance of each word in a caption transcription can enable evaluation metrics that best reflect Deaf and Hard of Hearing (DHH) readers' judgment of the caption quality. Prior work has proposed using word embeddings, e.g., word2vec or BERT embeddings, to model word importance in conversational transcripts. Recent work also disseminated a human-annotated word importance dataset. We conducted a word-token level analysis on this dataset and explored Part-of-Speech (POS) distribution. We then augmented the dataset with POS tags and reduced the class imbalance by generating 5% additional text using masking. Finally, we investigated how various supervised models learn the importance of words. The best performing model trained on our augmented dataset performed better than prior models. Our findings can inform the design of a metric for measuring live caption quality from DHH users' perspectives.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in accessibility**.

KEYWORDS

Accessibility, Word Importance, Augmentation

ACM Reference Format:
Akhter Al Amin, Saad Hassan, Matt Huenerfauth, and Cecilia Ovesdotter Alm . 2023. Modeling Word Importance in Conversational Transcripts: Toward improved live captioning for Deaf and hard of hearing viewers. In *20th International Web for All Conference (W4A '23), April 30–May 01, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3587281.3587290

W4A '23, April 30–May 01, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0748-3/23/04…$15.00
https://doi.org/10.1145/3587281.3587290

1 INTRODUCTION

Deaf and Hard of Hearing (DHH) users of captions are often dissatisfied with the quality of captioning provided in live contexts, which provide less time for caption production than pre-recorded contexts [4, 17]. Regulatory organization would benefit from metrics that give insights into DHH users' perception of the quality of imperfect captions.

Existing metrics used in transcription or captioning include Word Error Rate (WER) [2] and Number of Error in Recognition (NER) [21]. As noted by Kafle et al. [16], a major shortcoming of these metrics is that they do not consider the importance of individual words when measuring the accuracy of captioned transcripts (compared to the reference transcript) and assign equal weights to each word. DHH readers rely more heavily on important keywords while skimming through caption text [16].

Motivated by these shortcomings, prior work had proposed metrics which assign differential importance weights to individual words in captioned text when calculating an evaluation score [3, 13, 15, 25]. While that work had employed various methods for assigning the importance of words, prior research has revealed that DHH observers focus on keywords while reading a caption text to capture the main context [15, 16]. Since captions typically display only two or three lines of text at a time, the importance of each word judged by DHH viewers is largely driven by the localized contextual meaning of a word as shown in Table 1. Furthermore, methods that extract important words or phrases relying on an entire document or multiple documents might be unsuitable for capturing this aspect of caption-reading behavior.

Caption Text	That was pretty heartrending for her
Annotations	**That** was pretty **heartrending** for her

Table 1: A sample caption text where keywords are in bold font. For this text, excerpted from [14], annotators indicated *That* and *heartrending* were essential for comprehending the meaning.

We build on prior work and first investigate how existing state-of-the-art keyword-extraction methods perform when predicting the importance of words in a conversational transcript. Then, we

delve into the explainability of the performance of these models by investigating whether traditional Part-of-Speech (POS) tags of each token can explain importance of words in conversational text. Based on our initial findings, we then investigated how using both POS tags and word-embeddings as features affected the performance of a classical model. We then reduced the imbalanced distribution of word importance in the dataset by producing additional text using pre-trained BERT-embedding-based word masking method. We augmented the existing data and investigated how our data augmentations affected the performance of the models. Finally, we are releasing a dataset of word-embedding augmented with POS tag and augmented with the corpus generated using the masking method for future research in the field.

2 RELATED WORK

NLP researchers have explored approaches to determine **word-importance for various downstream tasks**, e.g. term weight determination when querying text [7], for text summarization [12] or text classification [25]. Prior research on identifying and scoring important words in a text has largely focused on the task of keyword or important-term or keyphrase extraction [5, 7, 11, 20, 22, 25]. This task involves in extracting keywords and keyphrases from a continuous description on a topic. In this space, supervised, unsupervised and semi-supervised approaches have been proposed and used, e.g., term frequency [20, 22, 23, 26, 27], statistical text features [6], traditional and static word-embedding using GloVe [15], BERT embedding [3, 11, 18], or BERT attention framework [9]. While the conceptualization of word importance as a keyword-extraction problem has enabled retrieving relevant information from large textual or multimedia data [7, 24], this approach may not generalize across domains, and functional and situational contexts of language use such as live broadcasts. For instance, given the meandering nature of topic transitions in television news broadcasts or talk shows [14, 15], when processing caption transcripts, a model of word importance that is more local may be more successful, rather than considering the entire transcript of the broadcast.

Focusing on understanding importance of words within conversational text, in our study, we use a previously released dataset from [14] consisting of 25,000-token subset of Switchboard corpus [10]. Kafle and Huenerfauth [14] developed this dataset by annotating importance of the words on range from 0.0 to 1.0, where 1.0 is most important. They used a partitioning protocol to divide the scores in 6 discrete classes: Class 1 [0-0.1), Class 2 [0.1-0.3), Class 3 [0.3-0.5), Class 4 [0.5-0.7), Class 5 [0.7-0.9), and Class 6 [0.9 - 1]. Finally, they used a sequence-to-sequence LSTM-based approach to predict importance category of each token. Another recent work demonstrated how importance of words can be learnt from BERT-contextualized embedding of each token [3]. In addition to releasing a BERT-embedding augmented dataset, their modeling experiments showed that logistic regression achieved the same accuracy reported by Kafle and Huenerfauth [14]. While both of these studies released datasets for use by other researchers, prior work has not focused on structural properties, e.g., POS tagging of these tokens at the transcript level. In our work, we have explored how to better explain the dataset, and we consider the shortcomings of this corpus and how limitations affects the performance of various machine-learning models. We perform two data augmentations to enable supervised models to learn word importance in live conversational transcripts.

3 EXPLAINABILITY

Our work evaluates the performance of three keyword or keyphrase extraction models: **1. NLTK-Rapid Automatic Keyword Extraction (RAKE) [22]**, **2. KeyBERT [11]** and **3. YAKE [6]**.

Method	Accuracy	*F1*	*RMS*
NLTK-RAKE	0.23	0.13	1.24
KeyBERT	0.23	0.15	2.63
YAKE	0.21	0.13	2.89

Table 2: Performance with three keyword extraction methods when estimating importance of the words from conversational transcripts.

As shown in Table 2, NLTK-RAKE, an existing unsupervised word frequency-based approach, and YAKE, a loosely unsupervised approach that relies on the frequency of keywords across the single or multiple documents, have been able to achieve F1 scores of 0.23, 0.21 respectively. The word-embedding-based method, KeyBERT achieved an F1 score of 0.15. To ensure word-by-word comparison, we have considered the phrase score as score of each word-token within a phrase.

While closely analyzing importance scores in transcripts, we observed that importance of words has been defined at sentence level. Since only two or three lines of caption text generally appears on screen at a given time, DHH viewers' assessment of importance of those words may be defined more by the text that appears on the screen and not the whole transcript. Therefore, existing methods that consider the entire document are less able to estimate the importance of words in this context as reflected by the low F1 score.

3.1 POS Tags and Word Importance

Based on the importance of POS tagging discussed in prior work [1], we explored how words possessing distinct POS tags were distributed across low- and high-importance classes. For this investigation, we define the low-importance words as words having scores less than or equal to 0.2, and higher-importance words, as having scores higher than or equal to 0.8. The frequency of POS tags across each class is shown in Figure 1.

After extracting the words that posses higher importance, we employed the NLTK POS-tagger [19] to tag POS of each word token. As shown in Fig. 1 (a), there is large proportion of words identified as nouns (NN, NNS and NNP), followed by adjectives (JJ).

Figure 1 (b) displays POS distribution among low-importance words. In this graph, words with NN POS-tags have the highest frequency, and the ranking of the next most frequent POS distributions were: determiner (DT), pronoun (PRP), or preposition or subordinating conjunction (IN). These differences in distributions motivate leveraging POS tags to better estimate word importance.

Since the above analysis 3.1 suggests that POS tagging is a potential indicator for differentiating high- and low-importance words,

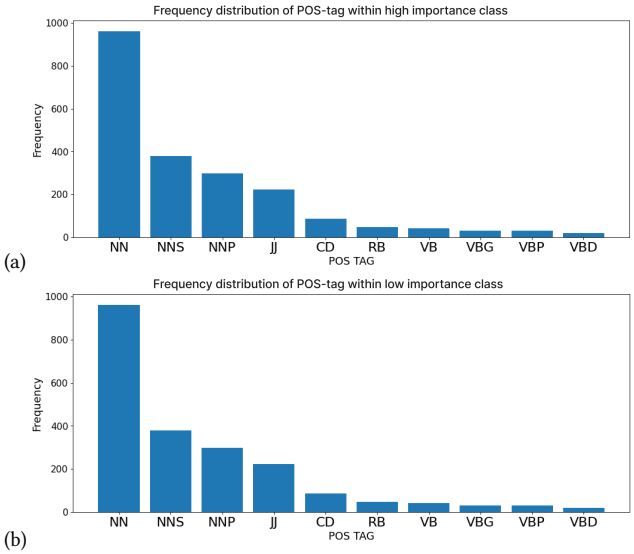

Figure 1: Distribution of Part-of-speech(POS) tag across (a) High Important words and (b) Low Important words. Graphs include the top-10 most frequent Part-of-speech (POS) tags in each class.

we employed POS tagging as a feature with the available BERT embedding dataset [3]. We added to the corpus POS tags based on the tagset corresponding to the UPENN corpus.

3.2 Masking to Synthesize Additional Text

Similar to Amin et al. [3], we also observed that the augmented dataset had an uneven word distribution across various classes (e.g., the percentage of words with importance classes 5 or 6 constitute only 6.3% of the data). Thus, we reduced the imbalance of this distribution in the dataset using a word-masking methodology [8], which allowed producing open-ended text with similar words. The objective of masking words and incorporating new samples was to increase the proportion of high importance words in our dataset.

We identified long sentences (15 or more words) from the corpus that contained high-importance words. We used *bert-large-cased-whole-word-masking* [8], a pretrained masking model, to generate more sentences. To minimize the risk of creating nonsense sentences, we selected sentences from the corpus that contained only one high-importance word. The generated sentences were manually checked. Here, we selected important words which had scores greater than or equal to 0.8. Table 3 displays examples of an actual sentence, masked sentence, generated sentence, and the corresponding scores.

Original Sent	so what do you do with your **time**
Masked Sent	so what do you do with your *[MASK]*
Generated Sent	so what do you do with your **life**
Importance Score	[0.1 0.2 0.1 0.2 0.1 0.2 0.4. 0.9]

Table 3: Example of using masking to producing sentences with high-importance words.

We generated 1296 new tokens which increased the words from high importance class by approximately 30%. Then, we generated corresponding BERT embedding and POS-tags to generate corresponding feature space. For the words we masked, we assumed that the generated word will also posses same importance score. Finally, we used this new text to augment the data.

4 PERFORMANCE ON THE AUGMENTED DATASET

Our findings in the previous section motivated us to investigate the performance of various supervised learning models for estimating word importance within conversational transcripts. We have employed simple logistic regression, which is the current best-performing model on the dataset we are using [3]. We used the same hyperparameters as in this prior work [3] to observe how the performance of the model was affected the by the new features. We applied this model to a version of the dataset with POS-tag features and to a version without such features, for comparison. We then investigated various configurations of neural network hyperparameters to study the performance of the models. These hyperparameters included various activation functions (tanh, relu), values for α, hidden layer sizes ((10, 30, 10, 18), (20,)), optimization algorithms or solvers ('sgd', 'adam', 'lbfgs'), and numbers of iteration (300-1500). During this process, we retained 10% of the data as a test set. To investigate how the addition of newly augmented data affected the performance of the model, we used the neural-network model with the best performing hyperparameter configuration [1] identified in the prior steps.

Method	Dataset	Accuracy	*F1*	*RMS*
Logistic Regression	Without POS	0.65	0.57	0.92
Logistic Regression	With POS	0.60	0.59	0.84
NN	With POS	0.72	0.61	0.77
NN	With POS and Additional Text	0.71	**0.64**	0.73

Table 4: Supervised classification performance showing macro-averaged F1 score and Root Mean Squared Error without and with POS-augmented dataset.

Classes	P	P (POS)	R	R (POS)	F1	F1 (POS)
6	0.56	**0.80**	0.83	0.60	0.67	**0.69**
5	0.40	**0.63**	0.29	0.64	0.33	**0.64**
3	0.48	0.37	0.38	0.34	0.42	0.36
4	0.48	0.41	0.33	0.42	0.39	0.41
2	0.64	**0.67**	0.68	0.71	0.66	**0.69**
1	0.69	**0.75**	0.72	0.74	0.71	**0.74**

Table 5: Precision, recall, and F1 performance of the best logistic regression model, for each importance class, both with and without using POS tag features.

[1] 'activation': 'relu', 'α': 0.0001, 'hidden_layer_sizes': (20,), 'solver': 'adam'.

Figure 4 illustrates how logistic regression and the neural network learned semantic features when POS tagging was used. Both approaches performed better the prior state-of-the-art method in terms of F1 score. We then examined the performance of the model across importance classes. Table 5 reveals that there was an improvement in precision for classes 1, 2, 5 & 6; this demonstrates the benefit of including POS-tag features and data augmentation.

5 LIMITATIONS

Even after data augmentation, our dataset is still small and does not support more advance feature learning approaches. Our automatic approach for data augmentation involved just masking a single word and ensuring that the output is sensible. Future work should employ human annotators or crowdsourcing to generate more data for live context (to enable deep learning approaches), and use more ecologically valid data annotation approaches where the annotators read the conversational transcriptions in blocks while assigning word importance. We only used one tagset in this paper.

6 DISCUSSION AND CONCLUSION

Our findings have revealed that traditional keyword or keyphrase extraction techniques are less suitable when users' perception of a text relatively depends on local characteristics of words. A followup analysis revealed that POS-tagging augmented with BERT contextualized word-embedding can better provide a meaningful score of word importance in conversational transcripts. Additionally, shallow neural networks were able to learn features and estimate the importance of words in the transcript. An in-depth analysis revealed that the model improves on identifying high-importance words. Given the accuracy on high-importance words, our proposed approach can be used to identify keywords within a small caption segment, which has been shown to benefit DHH viewers of caption texts [16]. We release the augmented dataset with this paper [2]. This dataset may benefit researchers who wish to measure the quality of captions from DHH users' perspective. Overall, our findings can support the design of automatic caption-evaluation metrics, to more accurately capture DHH viewers' judgments of the quality of an imperfect caption text. Wide availability of accurate automatic metrics may, in turn, help drive improvement in the quality of captions during live broadcasts.

ACKNOWLEDGMENTS

This material is partially based upon work supported by the National Science Foundation under Award No. DGE-2125362, 2235405, 2212303, and 1954284, and by the Department of Health and Human Services under Award No. 90DPCP0002-0100. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Department of Health and Human Services.

REFERENCES

[1] Mostafa Abdou, Artur Kulmizev, Vinit Ravishankar, Lasha Abzianidze, and Johan Bos. 2018. What can we learn from Semantic Tagging?. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4881–4889. https://doi.org/10.18653/v1/D18-1526

[2] Ahmed Ali and Steve Renals. 2018. Word Error Rate Estimation for Speech Recognition: e-WER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Melbourne, Australia, 20–24. https://doi.org/10.18653/v1/P18-2004

[3] Akhter Al Amin, Saad Hassan, Cecilia Alm, and Matt Huenerfauth. 2022. Using BERT Embeddings to Model Word Importance in Conversational Transcripts for Deaf and Hard of Hearing Users. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics, Dublin, Ireland, 35–40. https://doi.org/10.18653/v1/2022.ltedi-1.5

[4] Akhter Al Amin, Saad Hassan, and Matt Huenerfauth. 2021. Caption-Occlusion Severity Judgments across Live-Television Genres from Deaf and Hard-of-Hearing Viewers. In *Proceedings of the 18th International Web for All Conference* (Ljubljana, Slovenia) *(W4A '21)*. Association for Computing Machinery, New York, NY, USA, Article 26, 12 pages. https://doi.org/10.1145/3430263.3452429

[5] Gerlof Bouma. 2009. Normalized (Pointwise) Mutual Information in Collocation Extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*. Tubingen, 31–40.

[6] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences* 509 (2020), 257–289. https://doi.org/10.1016/j.ins.2019.09.013

[7] Zhuyun Dai and Jamie Callan. 2020. Context-Aware Document Term Weighting for Ad-Hoc Search. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1897–1907. https://doi.org/10.1145/3366423.3380258

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[9] Haoran Ding and Xiao Luo. 2021. AttentionRank: Unsupervised Keyphrase Extraction using Self and Cross Attentions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1919–1928. https://doi.org/10.18653/v1/2021.emnlp-main.146

[10] John Godfrey, E.C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. 517–520 vol.1. https://doi.org/10.1109/ICASSP.1992.225858

[11] Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT. https://doi.org/10.5281/zenodo.4461265

[12] Kai Hong and Ani Nenkova. 2014. Improving the Estimation of Word Importance for News Multi-Document Summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Gothenburg, Sweden, 712–721. https://doi.org/10.3115/v1/E14-1075

[13] Sushant Kafle, Cecilia Ovesdotter Alm, and Matt Huenerfauth. 2019. Fusion strategy for prosodic and lexical representations of word importance. In *Proc. Interspeech 2019*. 1313–1317. https://doi.org/10.21437/Interspeech.2019-1898

[14] Sushant Kafle and Matt Huenerfauth. 2018. A corpus for modeling word importance in spoken dialogue transcripts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan, 99–103. https://aclanthology.org/L18-1016

[15] Sushant Kafle and Matt Huenerfauth. 2019. Predicting the understandability of imperfect English captions for people who are deaf or hard of hearing. *ACM Trans. Access. Comput.* 12, 2, Article 7 (June 2019), 32 pages. https://doi.org/10.1145/3325862

[16] Sushant Kafle, Peter Yeung, and Matt Huenerfauth. 2019. Evaluating the benefit of highlighting key words in captions for people who are deaf or hard of hearing. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 43–55. https://doi.org/10.1145/3308561.3353781

[17] Raja Kushalnagar and Kesavan Kushalnagar. 2018. SubtitleFormatter: Making Subtitles Easier to Read for Deaf and Hard of Hearing Viewers on Personal Devices. In *Computers Helping People with Special Needs*, Klaus Miesenberger and Georgios Kouroupetroglou (Eds.). Springer International Publishing, Cham, 211–219.

[18] Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. 2021. Unsupervised Keyphrase Extraction by Jointly Modeling Local and Global Context. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 155–164. https://doi.org/10.18653/v1/2021.emnlp-main.14

[19] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. *CoRR* cs.CL/0205028 (2002). http://dblp.uni-trier.de/db/journals/corr/corr0205.html#cs-CL-0205028

[2] https://www.dropbox.com/s/ms239m9xqs2wi4f/Data%20for%20Published-20230311T200653Z-001.zip?dl=0

[20] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain, 404–411. https://aclanthology.org/W04-3252

[21] Pablo Romero-Fresco and Juan Martínez Pérez. 2015. *Accuracy Rate in Live Subtitling: The NER Model.* Audiovisual Translation in a Global Context. Palgrave Studies in Translating and Interpreting. Palgrave Macmillan, London.

[22] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. *Automatic Keyword Extraction from Individual Documents.* John Wiley & Sons, Ltd, Chapter 1, 1–20. https://doi.org/10.1002/9780470689646.ch1 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470689646.ch1

[23] Claude Sammut and Geoffrey I. Webb (Eds.). 2010. *TF–IDF.* Springer US, Boston, MA, 986–987. https://doi.org/10.1007/978-0-387-30164-8_832

[24] Chirag Shah and Pushpak Bhattacharyya. 2002. A Study for Evaluating the Importance of Various Parts of Speech (POS) for Information Retrieval. In *Proceedings of International Conference on Universal Knowledge, and Languages (ICUKL) 2002.*

[25] Imran Sheikh, Irina Illina, Dominique Fohr, and Georges Linares. 2016. Learning Word Importance with the Neural Bag-of-Words Model. In *ACL, Representation Learning for NLP (Repl4NLP) workshop (Proceedings of ACL 2016).* Berlin, Germany. https://hal.archives-ouvertes.fr/hal-01331720

[26] Mingyang Song, Liping Jing, and Lin Xiao. 2021. Importance Estimation from Multiple Perspectives for Keyphrase Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2726–2736. https://doi.org/10.18653/v1/2021.emnlp-main.215

[27] Asahi Ushio, Federico Liberatore, and Jose Camacho-Collados. 2021. Back to the Basics: A Quantitative Analysis of Statistical and Graph-Based Term Weighting Schemes for Keyword Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 8089–8103. https://doi.org/10.18653/v1/2021.emnlp-main.638