Ariadne's Thread: Using Text Prompts to Improve Segmentation of Infected Areas from Chest X-ray Images

Yi Zhong, Mengqiu Xu, Kongming Liang, Kaixin Chen, and Ming Wu[✉]

Beijing University of Posts and Telecommunications, Beijing, China
{xiliang2017,xumengqiu,liangkongming,chenkaixin,wuming}@bupt.edu.cn

Abstract. Segmentation of the infected areas of the lung is essential for quantifying the severity of lung disease like pulmonary infections. Existing medical image segmentation methods are almost uni-modal methods based on image. However, these image-only methods tend to produce inaccurate results unless trained with large amounts of annotated data. To overcome this challenge, we propose a language-driven segmentation method that uses text prompt to improve to the segmentation result. Experiments on the QaTa-COV19 dataset indicate that our method improves the Dice score by 6.09% at least compared to the uni-modal methods. Besides, our extended study reveals the flexibility of multi-modal methods in terms of the information granularity of text and demonstrates that multi-modal methods have a significant advantage over image-only methods in terms of the size of training data required.

Keywords: Multi-modal Learning · Medical Image Segmentation

1 Introduction

Radiology plays an important role in the diagnosis of some pulmonary infectious diseases, such as the COVID-19 pneumonia outbreak in late 2019 [1]. With the development of deep learning, deep neural networks are more and more used to process radiological images for assisted diagnosis, such as disease classification, lesion detection and segmentation, etc. With the fast processing of radiological images by deep neural networks, some diagnoses can be obtained immediately, such as the classification of bacterial or viral pneumonia and the segmentation mask for pulmonary infections, which is important for quantifying the severity of the disease as well as its progression [2]. Besides, these diagnoses given by the AI allow doctors to predict risks and prognostics in a "patient-specific" way [3]. Radiologists usually take more time to complete lesion annotation than AI, and annotation results can be influenced by individual bias and clinical experience [4].

Ariadne's thread, the name comes from ancient Greek myth, tells of Theseus walking out of the labyrinth with the help of Ariadne's golden thread.

Therefore, it is of importance to design automatic medical image segmentation algorithms to assist clinicians in developing accurate and fast treatment plans.

Most of the biomedical segmentation methods [5–9] are improved based on U-Net [10]. However, the performance of these image-only methods is constrained by the training data, which is also a dilemma in the medical image field. Radford et al. proposed CLIP [11] in 2021, where they used 4M image-text pairs for contrastive learning. With the rise of multi-modal learning in the recent years, there are also methods [12–16] that focus on vision-language pretraining/processing and applying them on local tasks. Li et al. proposed a language-driven medical image segmentation method LViT [16], using a hybrid CNN-Transformer structure to fuse text and image features. However, LViT uses an early fusion approach and the information contained in the text is not well represented. In this paper, we propose a multi-modal segmentation method that using independent text encoder and image encoder, and design a GuideDecoder to fuse the features of both modalities at decoding stage. Our main contributions are summarized as follow:

- We propose a language-driven segmentation method for segmenting infected areas from lung x-ray images. Source code of our method see: https://github.com/Junelin2333/LanGuideMedSeg-MICCAI2023
- The designed GuideDecoder in our method can adaptively propagate sufficient semantic information of the text prompts into pixel-level visual features, promoting consistency between two modalities.
- We have cleaned the errors contained in the text annotations of QaTa-COV19 [17] and contacted the authors of LViT to release a new version.
- Our extended study reveals the impact of information granularity in text prompts on the segmentation performance of our method, and demonstrates the significant advantage of multi-modal method over image-only methods in terms of the size of training data required.

2   Method

The overview of our proposed method is shown in Fig. 1(a). The model consists of three main components: Image Encoder, Text Encoder and GuideDecoder that enables multi-modal information fusion. As you can see, our proposed method uses a modular design. Compared to early stage fusion in LViT, our proposed method in modular design is more flexible. For example, when our method is used for brain MRI images, thanks to the modular design, we could first load pre-trained weights trained on the corresponding data to separate visual and text encoders, and then only need to train GuideDecoders.

Visual Encoder and Text Encoder. The Visual Encoder used in the model is ConvNeXt-Tiny [18]. For an input image $I \in \mathbb{R}^{H \times W \times 1}$, we extract multiple visual features from the four stages of ConvNeXt-Tiny, which are defined as $f_4 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C_1}$, $f_8 \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times C_2}$, $f_{16} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_3}$ and $f_{32} \in \mathbb{R}^{\frac{H}{32} \times \frac{W}{32} \times C_4}$,
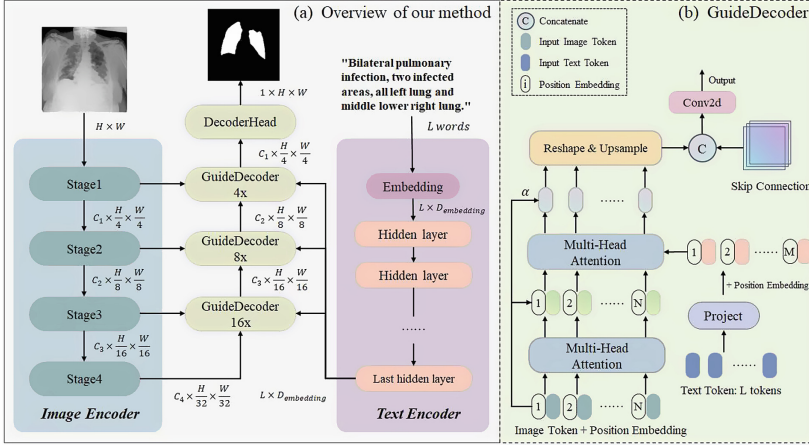
Fig. 1. The overview of the our proposed method (a) and the detail of the GuideDecoder in out method. Our proposed approach uses a modular design where the model consists mainly of an image encoder, a text encoder and several GuideDecoders. The GuideDecoder is used to adaptively propagate semantic information from textual features to visual features and output decoded visual features.

Note that C is the feature dimension, H and W are the height and width of the original image. For an input text prompt $T \in \mathbb{R}^L$, We adopt the CXR-BERT [19] to extract text features $g_t \in \mathbb{R}^{L \times C}$. Note that C is the feature dimension, L is the length of the text prompt.

GuideDecoder. Due to our modular design, visual features and textual features are encoded independently by different encoders. Therefore, the design of the decoder is particularly important, as we can only fuse multi-modal features from different encoders in post stage. The structure of GuideDecoder is shown in Fig. 1(b). The GuideDecoder first processes the input textual features and visual features before performing multi-modal interaction.

The input textual features first go through a projection module (i.e. Project in the figure) that aligns the dimensionality of the text token with that of the image token and reduces the number of text tokens. The projection process is shown in Eq. 1.

$$f_t = \sigma(Conv(TW_T)) \tag{1}$$

where W_T is a learnable matrix, $Conv(\cdot)$ denotes a 1×1 convolution layer, and $\sigma(\cdot)$ denotes the ReLU activation function. Given an input feature $T \in \mathbb{R}^{L \times D}$, the output projected features is $f_t \in \mathbb{R}^{M \times C_1}$, where M is the number of tokens after projection and C_1 is the dimension of the projected features, consistent with the dimension of the image token.

For the input visual features $I \in \mathbb{R}^{H \times W \times C_1}$, after adding the position encoding we use self-attention to enhance the visual information in them to obtain the

evolved visual features. The process is shown in Eq. 2.

$$f_i = I + LN(MHSA(I)) \tag{2}$$

where $MHSA(\cdot)$ denotes Multi-Head Self-Attention layer, $LN(\cdot)$ denotes Layer Normalization, and finally the evolved visual features $f_i \in \mathbb{R}^{H \times W \times C_1}$ with residuals could be obtained.

After those, the multi-head cross-attention layer is adopted to propagate fine-grained semantic information into the evolved image features. To obtain the multi-modal feature $f_c \in \mathbb{R}^{H \times W \times C_1}$, the output further computed by layer normalization and residual connection:

$$f_c = f_i + \alpha(LN(MHCA(f_i, f_t))) \tag{3}$$

where $MHCA(\cdot)$ denotes multi-head cross-attention and α is a learnable parameter to control the weight of the residual connection.

Then, the multi-modal feature $f_c \in \mathbb{R}^{(H \times W) \times C_1}$ would be reshaped and upsampling to obtain $f_c' \in \mathbb{R}^{H' \times W' \times C_1}$. Finally the f_c' is concatenated with $f_s \in \mathbb{R}^{H' \times W' \times C_2}$ on the channel dimension, where f_s is the low-level visual feature obtained from visual encoder via skip connection. The concatenated features are processed through a convolution layer and a ReLU activation function to obtain the final decoded output $f_o \in \mathbb{R}^{H' \times W' \times C_2}$

$$f_c' = Upsample(Reshape(f_c))$$
$$f_o = \sigma(Conv([f_c', f_s'])) \tag{4}$$

where $[\cdot, \cdot]$ represents the concatenate operation on the channel dimension.

3   Experiments

3.1   Dataset

The dataset used to evaluate our method performance is the QaTa-COV19 dataset [17], which is compiled by researchers from Qatar University and Tampere University. It consists of 9258 COVID-19 chest radiographs with pixel-level manual annotations of infected lung areas, of which 7145 are in the training set and 2113 in the test set. However, the original QaTa-COV19 dataset does not contain any matched text annotations.

Li et al. [16] have made significant contributions by extending the text annotations of the dataset, their endeavors are worthy of commendation. We conducted a revisitation of the text annotations and found several notable features. Each sentence consists of three parts, containing position information at different granularity. However, these sentences cannot be considered as medical reports for lacking descriptions of the disease, we consider them as a kind of "text prompt" just as the title of the paper states.

Besides, we found some obvious errors (e.g. misspelled words, grammatical errors and unclear referents) in the extended text annotations. We have fixed these identified errors and contacted the authors of LViT to release a new version of the dataset. Dataset see Github link: https://github.com/HUANGLIZI/LViT.

3.2   Experiment Settings

Following the file name of the subjects in the original train set, we split the training set and the validation set uniformly in the ratio of 80% and 20%. Therefore, the training set has a total of 5716 samples, the validation set has 1429 samples and the test set has 2113 samples. All images are cropped to 224×224 and the data is augmented using a random zoom with 10% probability.

We used a number of open source libraries including but not limited to PyTorch, MONAI [20] and Transformers [21] to implement our method and baseline approach. We use PyTorch Lightning for the final training and inference wrapper. All the methods are training on one NVIDIA Tesla V100 SXM3 32 GB VRAM GPU. We use the Dice loss plus Cross-entropy loss as the loss function, and train the network using AdamW optimization with a batch size of 32. We utilize the cosine annealing learning rate policy, the initial learning rate is set to 3e−4 and the minimal learning rate is set to 1e−6.

We used three metrics to evaluate the segmentation results objectively: Accuracy, Dice coefficient and Jaccard coefficient. Both Dice and Jaccard coefficient calculate the intersection regions over the union regions of the given predicted mask and ground truth, where the Dice coefficient is more indicative of the segmentation performance of small targets.

3.3   Comparison Experiments

We compared our method with common mono-modal medical image segmentation methods and with the LViT previously proposed by Li et al. The quantitative results of the experiment are shown in Table 1. UNet++ achieves the best performance of the mono-modal approach. Comparing to UNet++, our method improves accuracy by 1.44%, Dice score by 6.09% and Jaccard score by 9.49%. Our method improves accuracy by 1.28%, Dice score by 4.86% and Jaccard coefficient by 7.66% compared to the previous multi-modal method LViT. In general, using text prompts could significantly improve segmentation performance.

Table 1. Comparisons with some mono-modal methods and previous multi-modal method on QaTa-COV19 test set. '*' denotes these methods use text prompt and CXR-BERT as the text embedding encoder.

Method		Acc	Dice	Jaccard
Mono-Modal	Unet	0.9584	0.8299	0.7092
	Unet++	**0.9608**	**0.8369**	**0.7196**
	Attention Unet	0.9567	0.8240	0.7006
	Swin UNETR	0.9511	0.8002	0.6669
Multi-Modal*	LViT	0.9624	0.8492	0.7379
	Our Method	**0.9752**	**0.8978**	**0.8145**

The results of the qualitative experiment are shown in Fig. 2. The image-only mono-modal methods tend to generate some over-segmentation, while

the multi-modal approach refers to the specific location of the infected region through text prompts to make the segmentation results more accurate.
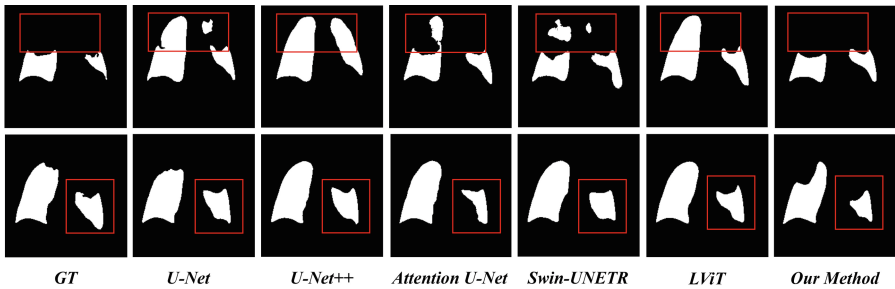

| GT | U-Net | U-Net++ | Attention U-Net | Swin-UNETR | LViT | Our Method |

Fig. 2. Qualitative results on QaTa-COV19. The sample image in the first row is from 'sub-S09345_ses-E16849_run-1_bp-chest_vp-ap_dx.png' and The sample image in the second row is from 'sub-S09340_ses-E17282_run-1_bp-chest_vp-ap_cr.png'.

3.4   Ablation Study

Our proposed method introduces semantic information of text in the decoding process of image features and designs the GuideDecoder to let the semantic information in the text guide the generation of the final segmentation mask. We performed an ablation study on the number of GuideDecoder used in the model and the results are shown in the Table 2.

Table 2. Ablation studies on QaTa-COV19 test set. We used different numbers (0–3) of GuideDecoders in the model to verify the effectiveness of the GuideDecoder. Note that the GuideDecoder in the model is replaced in turn by the Decoder in the UNet, 'w/o text' means without text and the model use UNet Decoders only.

Method	Acc	Dice	Jaccard
w/o text	0.9610	0.8414	0.7262
1 layer	0.9735	0.8920	0.8050
2 layers	0.9748	0.8963	0.8132
3 layers	**0.9752**	**0.8978**	**0.8144**

As can be seen from the Table 2, the segmentation performance of the model improves as the number of GuideDecoders used in the model increases. The effectiveness of GuideDecoder could be proved by these results.

3.5   Extended Study

Considering the application of the algorithm in clinical scenarios, we conducted several interesting extension studies based on the QaTa-COV19 dataset with the text annotations. It is worth mentioning that the following extended studies were carried out on our proposed method.

Impact of Text Prompts at Different Granularity on Segmentation Performance. In Sect. 3.1 we mention that each sample is extended to a text annotation with three parts containing positional information at different granularity, as shown in the Fig. 3. Therefore we further explored the impact of text prompts at different granularity on segmentation performance of our method and the results are shown in Table 3.
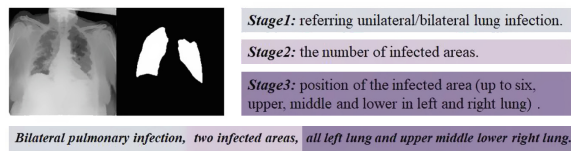

Fig. 3. The Split Example of Different Stages in Text Annotation. The chest x-ray, mask and text annotation correspond to 'covid_1.png'. We divide the sentence into three stages and distinguished them with a background colour, the darker the background colour the more detailed the stage has position information.

Table 3. Study of text prompts at different granularity and segmentation performance. The term *w/o text* in the table means *without text*, while *Stage1*, *Stage2*, and *Stage3* represent the three parts of each text annotation. See Fig. 3 for examples of the different Stages.

Stage of Text Prompt	Acc	Dice	Jaccard
w/o text	0.9610	0.8414	0.7262
stage1 + stage2	0.9648	0.8557	0.7479
stage3	**0.9753**	**0.8981**	**0.8151**
stage1 + stage2 + stage3	0.9752	0.8978	0.8145

The results in the table show that the segmentation performance of our proposed method is driven by the granularity of the position information contained in the text prompt. Our proposed method achieved better segmentation performance when given a text prompt with more detailed position information. Meanwhile, we observed that the performance of our method is almost identical

when using two types of text prompts, i.e. *Stage3* alone and *Stage1 + Stage2 + Stage3*. It means the most detailed position information in the text prompt plays the most significant role in improving segmentation performance. But this does not mean that other granularity of position information in the text prompt does not contribute to the improvement in segmentation performance. Even when the input text prompts contain only the coarsest location information (*Stage1 + Stage2* items in the Table 3), our proposed method yielded a 1.43% higher Dice score than the method without text prompt.

Impact of the Size of Training Data on Segmentation Performance. As shown in Table 4, our proposed method demonstrates highly competitive performance even with a reduced amount of training data. With only a quarter of the training data, our proposed method achieves a 2.69% higher Dice score than UNet++, which is the best performing mono-modal model trained on the full dataset. This provides sufficient evidence for the superiority of multi-modal approaches and the fact that suitable text prompts could significantly help improve the segmentation performance.

Table 4. Study of the size of training data and segmentation of performance. Note that the performance of UNet++ is used as a comparative reference.

Method	Acc	Dice	Jaccard
Unet++ (using 10% training data)	0.9425	0.7706	0.6268
Ours (using 10% training data)	0.9574	0.8312	0.7111
Unet++ (using 100% training data)	0.9608	0.8369	0.7196
Ours (using 15% training data)	0.9636	0.8503	0.7395
Ours (using 25% training data)	0.9673	0.8638	0.7602
Ours (using 50% training data)	0.9719	0.8821	0.7891
Ours (using 100% training data)	**0.9752**	**0.8978**	**0.8145**

We observed that when the training data was reduced to 10%, our method only began to exhibit inferior performance compared to UNet++, which was trained with all available data. Similar experiments could be found in the LViT paper. Therefore, it can be argued that multi-modal approaches require only a small amount of data (less than 15% in the case of our method) to achieve performance equivalent to that of mono-modal methods.

4   Conclusion

In this paper, we propose a language-driven method for segmenting infected areas from lung x-ray images. The designed GuideDecoder in our method can

adaptively propagate sufficient semantic information of the text prompts into pixel-level visual features, promoting consistency between two modalities. The experimental results on the QaTa-COV19 dataset indicate that the multi-modal segmentation method based on text-image could achieve better performance compared to the image-only segmentation methods. Besides, we have conducted several extended studies on the information granularity of the text prompts and the size of the training data, which reveals the flexibility of multi-modal methods in terms of the information granularity of text and demonstrates that multi-modal methods have a significant advantage over image-only methods in terms of the size of training data required.

References

1. Yin, S., Deng, H., Xu, Z., Zhu, Q., Cheng, J.: SD-UNet: a novel segmentation framework for CT images of lung infections. Electronics **11**(1), 130 (2022)
2. Oulefki, A., Agaian, S., Trongtirakul, T., Laouar, A.K.: Automatic COVID-19 lung infected region segmentation and measurement using CT-scans images. Pattern Recogn. **114**, 114 (2021)
3. Mu, N., Wang, H., Zhang, Y., Jiang, J., Tang, J.: Progressive global perception and local polishing network for lung infection segmentation of COVID-19 CT images. Pattern Recogn. **120**, 108168 (2021)
4. Fan, D.P., et al.: Inf-Net: automatic COVID-19 lung infection segmentation from CT images. IEEE Trans. Med. Imaging **39**(8), 2626–2637 (2020). https://doi.org/10.1109/TMI.2020.2996645
5. Cao, H., et al.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds.) Computer Vision-ECCV 2022 Workshops: Tel Aviv, Israel, 23–27 October 2022, Proceedings, Part III, pp. 205–218. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-25066-8_9
6. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: Crimi, A., Bakas, S. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, 27 September 2021, Revised Selected Papers, Part I, pp. 272–284. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-08999-2_22
7. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging **39**(6), 1856–1867 (2019)
8. Oktay, O., et al.: Attention U-Net: learning where to look for the pancreas. CoRR abs/1804.03999 (2018). http://arxiv.org/abs/1804.03999
9. Hatamizadeh, A., et al.: UNETR: transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 574–584 (2022)

10. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds.) Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015, Proceedings, Part III 18, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

11. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)

12. Wang, Z., et al.: CRIS: clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11686–11695 (2022)

13. Rao, Y., et al.: DenseCLIP: language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18082–18091 (2022)

14. Bhalodia, R., et al.: Improving Pneumonia localization via cross-attention on medical images and reports. In: de Bruijne, M., et al. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021, Proceedings, Part II 24, pp. 571–581. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_53

15. Müller, P., Kaissis, G., Zou, C., Rueckert, D.: Radiological reports improve pre-training for localized imaging tasks on chest X-rays. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention-MICCAI 2022: 25th International Conference, Singapore, 18–22 September 2022, Proceedings, Part V, pp. 647–657. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16443-9_62

16. Li, Z., et al.: LViT: language meets vision transformer in medical image segmentation. arXiv preprint arXiv:2206.14718 (2022)

17. Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: OSegNet: operational segmentation network for Covid-19 detection using chest X-ray images. In: 2022 IEEE International Conference on Image Processing (ICIP), pp. 2306–2310. IEEE (2022)

18. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)

19. Boecking, B., et al.: Making the most of text semantics to improve biomedical vision-language processing. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022, Proceedings, Part XXXVI, pp. 1–21. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-20059-5_1

20. Cardoso, M.J., et al.: MONAI: an open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701 (2022)

21. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics, Online, October 2020. https://www.aclweb.org/anthology/2020.emnlp-demos.6