More than Text: Multi-modal Chinese Word Segmentation

Dong Zhang[1], **Zheng Hu**[2], **Shoushan Li**[1]*, **Hanqian Wu**[2], **Qiaoming Zhu**[1], **Guodong Zhou**[1]
[1]School of Computer Science and Technology, Soochow University, China
[2]School of Computer Science and Engineering, Southeast University, China
{dzhang, lishoushan, qmzhu, gdzhou}@suda.edu.cn
{zhenghu, hanqian}@seu.edu.cn

Abstract

Chinese word segmentation (CWS) is undoubtedly an important basic task in natural language processing. Previous works only focus on the textual modality, but there are often audio and video utterances (such as news broadcast and face-to-face dialogues), where textual, acoustic and visual modalities normally exist. To this end, we attempt to combine the multi-modality (mainly the converted text and actual voice information) to perform CWS. In this paper, we annotate a new dataset for CWS containing text and audio. Moreover, we propose a time-dependent multi-modal interactive model based on Transformer framework to integrate multi-modal information for word sequence labeling. The experimental results on three different training sets show the effectiveness of our approach with fusing text and audio.

1 Introduction

Word segmentation is a fundamental task in Natural Language Processing (NLP) for those languages without word delimiters, e.g., Chinese and many other East Asian languages (Duan and Zhao, 2020). In this paper, we mainly take Chinese language as investigating object, namely CWS. As we know, CWS has been applied as an essential preprocessing step for many other NLP tasks (Zhou et al., 2019; Qiu et al., 2020), such as named entity recognition, sentiment analysis, machine translation, etc.

In the literature, some popular approaches to CWS systems report a high performance at the level of 96%–98%, and these systems typically require a large scale of pre-segmented textual dataset for training. However, the collection of a specific scenario on such large scale is very time-consuming and resource-intensive, such as video monologues and audio broadcast. In these scenarios, there are multiple modalities: text, audio and vision, thus if only using the text seems not a good choice. For example, as shown in Figure 1, if we only read the text"必须不忘初心牢记使命" with no punctuation, it is not easy to make word segmentation immediately. However, if there is the acoustic information, we can observe the obvious stop in spectrum and sonic wave at the middle of "心" and "牢", which provides the facility for CWS.

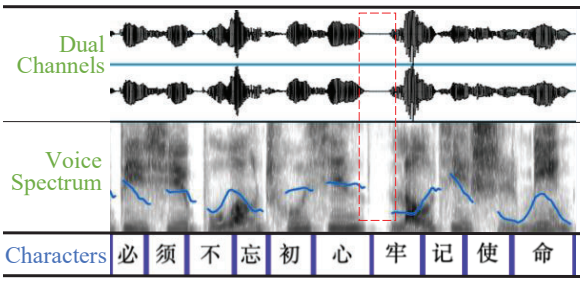

Figure 1: A multi-modal example for CWS. 必须(must) 不(not) 忘(forget) 初(original) 心(heart) 牢记(remember) 使命(mission).

Therefore, in this paper, we propose to performing CWS with multi-modality, namely MCWS, by a time-dependent multi-modal interactive network. Specifically, we first collect a new dataset from an audio and video news broadcast platform and annotate the word boundaries of audio transcription text. Second, we make the text and the audio align as the time stamp of each character, then encode both modalities [1] by Transformer-based framework to capture the intra-modal dynamics. Third, we design a time-dependent multi-modal interaction module for each character step to generate the multi-modal hybrid character representation.

*Corresponding author: lishoushan@suda.edu.cn

[1]Since each video in this platform mainly describe the specific news scene, not the face of the speaker, the visual modality is not useful for word segmentation. Therefore, for the sake of simplicity, we only utilize text and audio to perform CWS.

Finally, we leverage the CRF to perform sequence labeling on the basis of the above character representation.

We evaluate our approach on the newly annotated small-scale dataset with different size of training sets. The experimental results demonstrate that our approach performs significantly better than the single-modal state-of-the-art and the multi-modal approaches with early fused features of CWS.

2 Related Work

Xu (2003) first formalize CWS as a sequence labeling task, considering CWS as a supervised learning from annotated corpus with human segmentation. Peng et al. (2004) further adopt standard sequence labeling tool CRFs for CWS modeling, achieving a best performance in their same period. Then, a large amount of approaches based on above settings are proposed for CWS (Li and Sun, 2009; Sun and Xu, 2011; Mansur et al., 2013; Zhang et al., 2013).

Recently, deep neural approaches have been widely proposed to minimize the efforts in feature engineering for CWS (Zheng et al., 2013; Pei et al., 2014; Chen et al., 2015; Cai and Zhao, 2016; Zhou et al., 2017; Yang et al., 2017; Zhang et al., 2017; Ma et al., 2018; Li et al., 2019; Wang et al., 2019a; Fu et al., 2020; Ding et al., 2020; Tian et al., 2020a; Zhao et al., 2020). Among these studies, most of them follow the character-based paradigm to predict segmentation labels for each character in an input sentence. To enhance CWS with neural models, there were studies leverage external information, such as vocabularies from auto-segmented external corpus and weakly labeled data (Wang and Xu, 2017; Higashiyama et al., 2019; Gong et al., 2020).

To our best knowledge, we are first to perform CWS with multi-modality, which can deal with multi-modal scenarios and offers an alternative solution to robustly enhancing neural CWS models.

3 Data Collection and Annotation

We collect the multi-modal data for CWS from a Chinese news reporting platform "Xuexi"[2]. We mainly focus on the audios equipped with machine transcription text. In total, we crawl 120 short videos and segment them into about 2000 sentences. To avoid the contextual influence and augment the robust of designed computing model, we randomly select 250 sentences to annotate the word boundaries, and the remaining data are used to perform semi-supervised or unsupervised learning in the future.

[2]https://www.xuexi.cn/

Items	Size
Sentences	250
Avg. Length (Character)	50.56
Avg. Length (Word)	26.95
Avg. Length (Time)(s)	10.63
Max Length (Character)	382
Max Length (Word)	231
Max Length (Time)(s)	95.06
Total Characters	12640
Total Words	6736
Total Time(s)	2658.16

Table 1: The statistics summary for used data.

We annotate these Chinese audio transcriptions following the CTB word segmentation guidelines by Xia (2000). Two annotators are asked to annotate the data. Due to the clear annotation guideline, the annotation agreement is very high, reaching 98.3%. The disagreement instances are judged by an expert. The statistics of our annotated data are summarized in Table 1.

4 Time-dependent Multi-modal Interactive Network for CWS

In this section, we introduce our proposed multi-modal approach for CWS, namely Time-dependent Multi-modal interactive Network (TMIN), which can capture the interactive semantics between text and audio for better word segmentation. This approach mainly consists of three modules: time-dependent uni-modal interaction, time-dependent multi-modal interaction and CRF labeling. Figure 2 shows the overall architecture of our TMIN.

4.1 Time-dependent Uni-modal Interaction

To better capture the temporal correspondences between different modalities (Zhang et al., 2019; Ju et al., 2020), we first align two modalities by extracting the exact time stamp of each phoneme and character using Montreal Forced Aligner (McAuliffe et al., 2017).

For machines to understand human utterance, they must be first able to understand the intra-modal dynamics (Zadeh et al., 2018; Wang et al., 2019b; Tsai et al., 2019) in each modality, such as the word order and grammar in text, breathe and

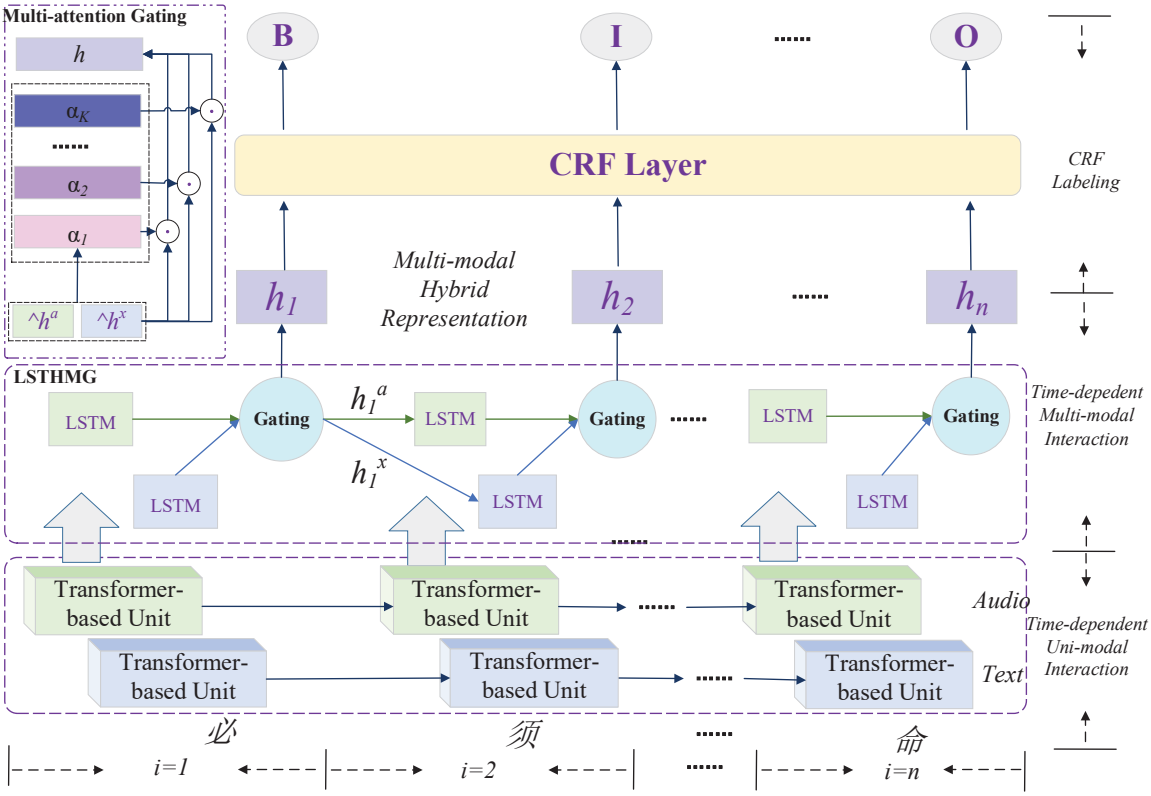

Figure 2: The overview of our proposed TMIN.

tone in audio.

Textual Modality. We use BERT (Devlin et al., 2019) as encoder to perform intra-modal interactions and obtain the contextual character representation. Then, each character of text transcripts can be represented as: $X = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^{n \times d_1}$.

Acoustic Modality. We use a famous audio processing tool, i.e., OpenSMILE (Eyben et al., 2010), to extract the MFCC, LP-coefficients, pure FFT spectrum, etc. from dual channels (Jayram et al., 2002; Sakran et al., 2017), and leverage multiple Transformer layers (Vaswani et al., 2017) to perform intra-modal interactions. Then, each character-level audio feature can be represented as: $A = (a_1, a_2, \cdots, a_n) \in \mathbb{R}^{n \times d_2}$.

4.2 Time-dependent Multi-modal Interaction

To better capture the cross-modal semantic correspondences (Wu et al., 2020; Zhang et al., 2020), we design a long- and short-term hybrid memory gating (LSTHMG) block, which is a extension of standard LSTM.

We first obtain the current memory of each character-level representation for both modalities.

$$\hat{h}_i^x, c_i^x = \text{LSTM}_i^x(x_i, h_{i-1}^x, c_{i-1}^x) \quad (1)$$

$$\hat{h}_i^a, c_i^a = \text{LSTM}_i^a(a_i, h_{i-1}^a, c_{i-1}^a) \quad (2)$$

where LSTM denotes the standard LSTM (Graves et al., 2013).

After current updating, we employ multi-attention to control the different contributions of each hidden state.

$$h_i = \hat{h}_i + \text{MA}(\hat{h}_i^x, \hat{h}_i^a) \quad (3)$$

$$= \hat{h}_i + \sum_{l=0}^{L} (\text{softmax}(\frac{Q^l (K^l)^\top}{\sqrt{d}}) V^l) \quad (4)$$

where MA denotes the multi-attention gating mechanism, which is considered to mine multiple potential dimension-aware importance for each modality (Zadeh et al., 2018). $\hat{h}_i \in \mathbb{R}^{(d_1+d_2)\times 1}$ is the unsqueezed concatenation of $\hat{h}_i^x$ and $\hat{h}_i^a$. L denotes the max times for attentions. The query Q^l, key K^l and value V^l at the l-th time are defined similarly to self-attention (Vaswani et al., 2017):

$$Q^l = \hat{h}_i W_q^l, K^l = \hat{h}_i W_k^l, V^l = \hat{h}_i W_v^l \quad (5)$$

Note that h_i denotes the sum of L times attentional state concatenation for multi-modal representation at character-level step i, which is then used to perform word segmentation by CRF. Besides, we split each part for each modality as its own dimension: h_i^x and h_i^a, and input them into the next LSTHMG step.

Model	50		100		150	
	F	R_{oov}	F	R_{oov}	F	R_{oov}
BC(Text)	93.13	93.15	94.29	95.19	95.29	96.15
BC(Audio)	30.26	36.56	34.63	37.11	33.65	36.75
BC(Text+Audio)	34.43	37.54	32.67	36.68	33.39	37.04
WMSEG(Text)	94.26	94.25	95.24	95.47	95.39	95.13
WMSEG(Audio)	69.34	70.29	70.46	71.00	71.20	77.17
WMSEG(Text+Audio)	63.29	53.21	69.71	69.20	70.37	70.44
TMIN(Ours)	**94.72**	**94.28**	**95.96**	**95.84**	**96.62**	**96.73**

Table 2: Performance (the overall F-score and the recall of OOV) comparison of different approaches on different training size. We perform a Friedman test on model- (row-) wise p-value< 0.05.

4.3 CRF Labeling

Since the textual and acoustic semantics of each character have been integrated by time-dependent uni-modal and multi-modal interactions, we allow h_i to perform conditional sequence labeling. Instead of decoding each label independently, we model them jointly using a CRF to consider the correlations between labels in neighborhoods. Formally,

$$p(y|\hat{X}) = \frac{\prod_{i=1}^{n} \mathcal{S}_i(y_{i-1}, y_i, \hat{X})}{\sum_{y' \in Y} \prod_{i=1}^{n} \mathcal{S}_i(y'_{i-1}, y'_i, \hat{X})} \quad (6)$$

where $\mathcal{S}_i(y_{i-1}, y_i, \hat{X})$ and $\mathcal{S}_i(y'_{i-1}, y'_i, \hat{X})$ are potential functions. $\hat{X}$ denotes the input of CRF. Y denotes the output label space.

We use the maximum conditional likelihood estimation for CRF training. The logarithm of likelihood is given by: $\sum_i \log p(y|\hat{X})$. In the inference phase, we predict the output sequence that obtains the maximum score given by: $\text{argmax}_{y' \in Y} p(y|\hat{X})$.

5 Experimentation

In this section, we provide the exploratory experimental results and a case analysis.

5.1 Experimental Setting

Data Split. We evaluate our approach on the different size of training sets and the same validation set and test set, i.e., 50, 100 and 150 sentences for training, the remaining 50 and 50 sentences for validation and test, respectively. For different training sets, the Out-of-vocabulary (OOV) rate in test set is 92.89%, 46.73% and 30.93%, respectively.

Implementation Details. The character embeddings of text X are initialized with the cased BERT_{base} model pre-trained with dimension of 768, and fine-tuned during training. The character-level embeddings of audio A are encoded by Transformer with dimension of 124. The learning rate, the dropout rate, and the tradeoff parameter are respectively set to 1e-4, 0.5, and 0.5, which can achieve the best performance on the development set of both datasets via a small grid search over the combinations of [1e-5, 1e-4], [0.1, 0.5], and [0.1, 0.9] on two pieces of NVIDIA GTX 2080Ti GPU with pytorch 1.7. Based on best-performed development results, the Transformer layers for audio encoding and the multi-attention times L in gating is set 2 and 4, respectively. To motivate future research, the dataset, aligned features and code will be released [3].

Baselines. For a thorough comparison, we implement the following approaches with F1 as metric: 1) BERT and CRF framework, **BC**: **BC(Text)** , **BC(Audio)**, and **BC(Text+Audio)**. 2) A representative state-of-the-art, **WMSEG** (Tian et al., 2020b): **WMSEG(Text)**, **WMSEG(Audio)**, and **WMSEG(Text+Audio)**. Note that the approaches with (Text) take character-level text as input, the approaches with (Audio) take character-level audio as input, and the approaches with (Text+Audio) take character-level concatenation of text and audio as input.

5.2 Main Results

Table 2 shows the performance of different baselines compared with our approach, where the overall F-score and the recall of OOV are reported. From this table we can see that:

1) **WMSEG** performs much better than the general framework **BC**. This indicates that it is effective for **WMSEG** to incorporate wordhood information with several popular encoder-decoder com-

[3] https://github.com/MANLP-suda/MCWS

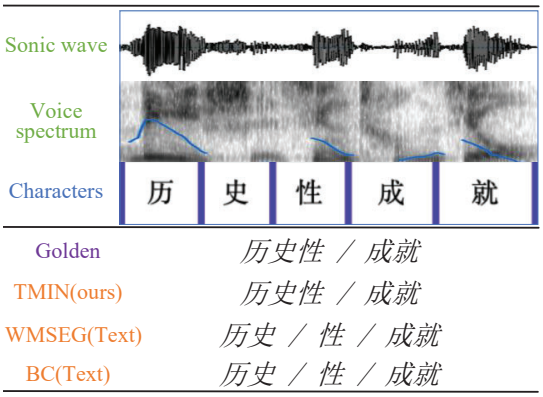

Figure 3: A case of the predicted results by different approaches. 1) 历史性(historic) 2) 历史(history), 性(sex) 3)成就(achievement).

binations and it is suitable as a competitive baseline.

2) The approach with only audio perform significantly worse than the approaches with text only, suggesting that it is confusing of the various acoustic features and we should utilize audio modality properly.

3) In most cases, the baselines with both text and audio bring in a poor performance compared with uni-modal approach, which suggests that simply concatenation of time-dependent character-level features for CWS seems a bad choice.

4) Among all approaches, our **TMIN** performs best, and significantly outperforms the competitive baselines (p-value<0.05). Moreover, with regard to R_{oov}, we can observe that our **TMIN** is able to recognize new words more accurately. This is mainly because our approach can obtain effective multi-modal information by time-dependent fusion against only textual, acoustic or early fused approaches.

5.3 Case Study

Figure 3 illustrates a real instance of the predicted boundaries by different approaches. From this figure, we can see that both **WMSRG** and **BC** give the wrong prediction of the boundary in "史" and "性" though they determine the correct segmentation for "历史" and "成就". However, our **TMIN** achieves all exact segmentation of this instance. This is mainly because it is very effective for audio, where there are a continuous breathing in the character " 性", thus "历史性" is a complete word.

6 Conclusion and Future Work

This paper proposes a new dataset for multi-modal Chinese word segmentation (MCWS), which is the first attempt to explore the multi-modality for traditional CWS. Meanwhile, we propose a time-dependent multi-modal interactive network (TMIN) to effectively integrate textual and acoustic features. The preliminary experimental results and case analysis demonstrate the reliability of our motivation and the effectiveness of the proposed approach.

In the future, we will annotate more samples at the current setting, and collect new samples with more modalities, such as visual information in social media, monologues and dialogues with continuous front face. Moreover, we will employ the neural active learning approaches for MCWS to reduce the annotation and achieve the best performance.

Acknowledgments

We thank all anonymous reviewers for their helpful comments. This work was supported by the National Key R&D Program of China under Grant No. 2020AAA0108600 and the NSFC grant No. 62076176. This work was also supported by a project funded by China Postdoctoral Science Foundation No. 2020M681713.

References

Deng Cai and Hai Zhao. 2016. Neural word segmentation learning for chinese. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. 2015. Long short-term memory neural networks for chinese word segmentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1197–1206.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ning Ding, Dingkun Long, Guangwei Xu, Muhua Zhu, Pengjun Xie, Xiaobin Wang, and Haitao Zheng. 2020. Coupling distant annotation and adversarial training for cross-domain chinese word segmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6662–6671.

Sufeng Duan and Hai Zhao. 2020. Attention is all you need for chinese word segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3862–3872.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Jinlan Fu, Pengfei Liu, Qi Zhang, and Xuanjing Huang. 2020. Rethinkcws: Is chinese word segmentation a solved task? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5676–5686.

Chen Gong, Zhenghua Li, Bowei Zou, and Min Zhang. 2020. Multi-grained chinese word segmentation with weakly labeled data. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2026–2036.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey E. Hinton. 2013. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 6645–6649.

Shohei Higashiyama, Masao Utiyama, Eiichiro Sumita, Masao Ideuchi, Yoshiaki Oida, Yohei Sakamoto, and Isaac Okada. 2019. Incorporating word attention into character-based word segmentation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2699–2709.

A. K. V. Sai Jayram, V. Ramasubramanian, and T. V. Sreenivas. 2002. Robust parameters for automatic segmentation of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2002, May 13-17 2002, Orlando, Florida, USA*, pages 513–514.

Xincheng Ju, Dong Zhang, Junhui Li, and Guodong Zhou. 2020. Transformer-based label set generation for multi-modal multi-label emotion detection. In *Proceedings of the 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 512–520.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is word segmentation necessary for deep learning of chinese representations? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3242–3252.

Zhongguo Li and Maosong Sun. 2009. Punctuation as implicit annotations for chinese word segmentation. *Comput. Linguistics*, 35(4):505–512.

Ji Ma, Kuzman Ganchev, and David Weiss. 2018. State-of-the-art chinese word segmentation with bi-lstms. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4902–4908.

Mairgup Mansur, Wenzhe Pei, and Baobao Chang. 2013. Feature-based neural language model and chinese word segmentation. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 1271–1277.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 498–502.

Wenzhe Pei, Tao Ge, and Baobao Chang. 2014. Max-margin tensor neural network for chinese word segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 293–303.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland*.

Xipeng Qiu, Hengzhi Pei, Hang Yan, and Xuanjing Huang. 2020. A concise model for multi-criteria chinese word segmentation with transformer encoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 2887–2897.

Alaa Ehab Sakran, Sherif Mahdy Abdou, Salah Eldeen Hamid, and Mohsen Rashwan. 2017. A review: Automatic speech segmentation. *International Journal of Computer Science and Mobile Computing*, 6(4):308–315.

Weiwei Sun and Jia Xu. 2011. Enhancing chinese word segmentation using unlabeled data. In *Proceedings of the 2011 Conference on Empirical Methods in*

Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 970–979.

Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. 2020a. Joint chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8286–8296.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020b. Improving Chinese word segmentation with wordhood memory networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285, Online. Association for Computational Linguistics.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6558–6569.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Chunqi Wang and Bo Xu. 2017. Convolutional neural network with word embeddings for chinese word segmentation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 163–172.

Xiaobin Wang, Deng Cai, Linlin Li, Guangwei Xu, Hai Zhao, and Luo Si. 2019a. Unsupervised learning helps supervised neural word segmentation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7200–7207.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019b. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7216–7223.

Liangqing Wu, Dong Zhang, Qiyuan Liu, Shoushan Li, and Guodong Zhou. 2020. Speaker personality recognition with multimodal explicit many2many interactions. In *Proceedings of IEEE International Conference on Multimedia and Expo, ICME 2020, London, UK, July 6-10, 2020*, pages 1–6.

Fei Xia. 2000. The segmentation guidelines for the penn chinese treebank (3.0).

Nianwen Xu. 2003. Chinese word segmentation as character tagging. *Int. J. Comput. Linguistics Chin. Lang. Process.*, 8(1).

Jie Yang, Yue Zhang, and Fei Dong. 2017. Neural word segmentation with rich pretraining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 839–849.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5642–5649.

Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multi-modal multi-label emotion detection with modality and label dependence. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3584–3593.

Dong Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Effective sentiment-relevant word selection for multi-modal sentiment analysis in spoken language. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 148–156.

Longkai Zhang, Houfeng Wang, Xu Sun, and Mairgup Mansur. 2013. Exploring representations from unlabeled data with co-training for chinese word segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 311–321.

Meishan Zhang, Guohong Fu, and Nan Yu. 2017. Segmenting chinese microtext: Joint informal-word detection and segmentation with neural networks. In

Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017, pages 4228–4234.

Xiaoyan Zhao, Min Yang, Qiang Qu, and Yang Sun. 2020. Improving neural chinese word segmentation with lexicon-enhanced adaptive attention. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1953–1956.

Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. 2013. Deep learning for chinese word segmentation and POS tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 647–657.

Hao Zhou, Zhenting Yu, Yue Zhang, Shujian Huang, Xin-Yu Dai, and Jiajun Chen. 2017. Word-context character embeddings for chinese word segmentation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 760–766.

Jianing Zhou, Jingkang Wang, and Gongshen Liu. 2019. Multiple character embeddings for chinese word segmentation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student Research Workshop*, pages 210–216.