Xiaomingbot: A Multilingual Robot News Reporter

Runxin Xu[1,*], **Jun Cao**[2], **Mingxuan Wang**[2], **Jiaze Chen**[2], **Hao Zhou**[2]
Ying Zeng[2], **Yuping Wang**[2], **Li Chen**[2], **Xiang Yin**[2], **Xijin Zhang**[2]
Songcheng Jiang[2], **Yuxuan Wang**[2], and **Lei Li**[2†]
[1] School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] ByteDance AI Lab, Shanghai, China
`runxinxu@gmail.com`
`{caojun.sh, wangmingxuan.89, chenjiaze, zhouhao.nlp`
`zengying.ss, wangyuping, chenli.cloud,`
`yinxiang.stephen, zhangxijin, jiangsongcheng`
`wangyuxuan.11, lileilab}@bytedance.com`

Abstract

This paper proposes the building of Xiaomingbot, an intelligent, multilingual and multimodal software robot equipped with four integral capabilities: news generation, news translation, news reading and avatar animation. Its system summarizes Chinese news that it automatically generates from data tables. Next, it translates the summary or the full article into multiple languages, and reads the multilingual rendition through synthesized speech. Notably, Xiaomingbot utilizes a voice cloning technology to synthesize the speech trained from a real person's voice data in one input language. The proposed system enjoys several merits: it has an animated avatar, and is able to generate and read multilingual news. Since it was put into practice, Xiaomingbot has written over 600,000 articles, and gained over 150,000 followers on social media platforms.

1 Introduction

The wake of automated news reporting as an emerging research topic has witnessed the development and deployment of several robot news reporters with various capabilities. Technological improvements in modern natural language generation have further enabled automatic news writing in certain areas. For example, GPT-2 is able to create fairly plausible stories (Radford et al., 2019). Bayesian generative methods have been able to create descriptions or advertisement slogans from structured data (Miao et al., 2019; Ye et al., 2020). Summarization technology has been exploited to produce reports on sports news from human commentary text (Zhang et al., 2016).

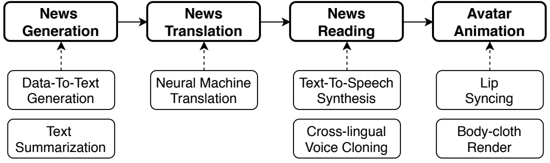


Figure 1: Xiaomingbot System Architecture

While very promising, most previous robot reporters and machine writing systems have limited capabilities reports on sports news that only focus on text generation. We argue in this paper that an intelligent robot reporter should acquire the following capabilities to be truly user friendly: *a)* it should be able to create news articles from input data; *b)* it should be able to read the articles with lifelike character animation like in TV broadcasting; and *c)* it should be multi-lingual to serve global users. None of the existing robot reporters are able display performance on these tasks that matches that of a human reporter. In this paper, we present Xiaomingbot, a robot news reporter capable of news writing, summarization, translation, reading, and visual character animation. In our knowledge, it is the first multilingual and multimodal AI news agent. Hence, the system shows great potential for large scale industrial applications.

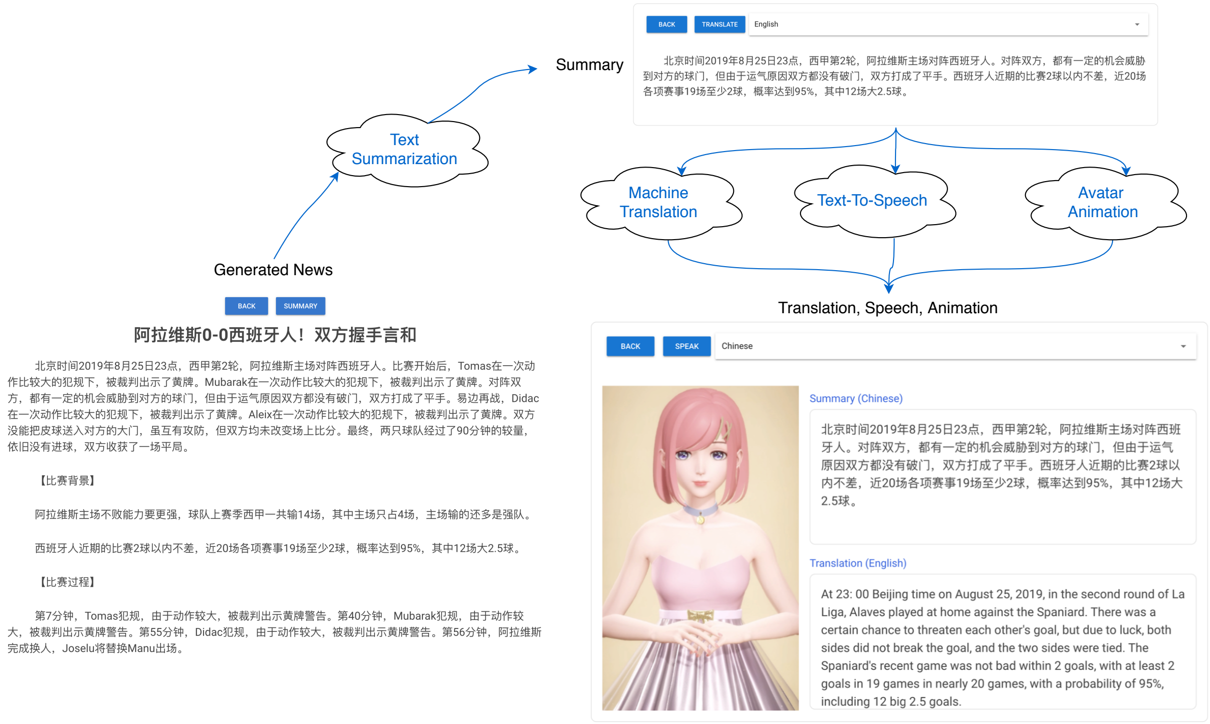
Figure 1 shows the capabilities and components of the proposed Xiaomingbot system. It includes four components: *a)* a news generator, *b)* a news translator, *c)* a cross-lingual news reader, and *d)* an animated avatar. The text generator takes input information from data tables and produces articles in natural languages. Our system is targeted for news area with available structure data, such as sports games and financial events. The fully automated news generation function is able to write and publish a story within mere seconds after the event took place, and is therefore much faster compared with manual writing. Within a few seconds after the events, it can accomplish the writing and publishing of a story. The system also uses a pretrained

---

[*] The work was done while the author was an intern at ByteDance AI Lab.

[†] Corresponding author.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1–8
July 5 - July 10, 2020. ©2020 Association for Computational Linguistics



北京时间2019年8月25日23点，西甲第2轮，阿拉维斯主场对阵西班牙人。对阵双方，都有一定的机会威胁到对方的球门，但由于运气原因双方都没有破门，双方打成了平手。西班牙人近期的比赛2球以内不差，近20场各项赛事19场至少2球，概率达到95%，其中12场大2.5球。

阿拉维斯0-0西班牙人！双方握手言和

北京时间2019年8月25日23点，西甲第2轮，阿拉维斯主场对阵西班牙人。比赛开始后，Tomas在一次动作比较大的犯规下，被裁判出示了黄牌。Mubarak在一次动作比较大的犯规下，被裁判出示了黄牌。对阵双方，都有一定的机会威胁到对方的球门，但由于运气原因双方都没有破门，双方打成了平手。易边再战，Didac在一次动作比较大的犯规下，被裁判出示了黄牌。Aleix在一次动作比较大的犯规下，被裁判出示了黄牌。双方没能把皮球送入对方的大门，虽互有攻防，但双方均未改变场上比分。最终，两只球队经过了90分钟的较量，依旧没有进球，双方收获了一场平局。

【比赛背景】

阿拉维斯主场不败能力要更强，球队上赛季西甲一共输14场，其中主场只占4场，主场输的还多是强队。

西班牙人近期的比赛2球以内不差，近20场各项赛事19场至少2球，概率达到95%，其中12场大2.5球。

【比赛过程】

第7分钟，Tomas犯规，由于动作较大，被裁判出示黄牌警告。第40分钟，Mubarak犯规，由于动作较大，被裁判出示黄牌警告。第55分钟，Didac犯规，由于动作较大，被裁判出示黄牌警告。第56分钟，阿拉维斯完成换人，Joselu将替换Manu出场。

北京时间2019年8月25日23点，西甲第2轮，阿拉维斯主场对阵西班牙人。对阵双方，都有一定的机会威胁到对方的球门，但由于运气原因双方都没有破门，双方打成了平手。西班牙人近期的比赛2球以内不差，近20场各项赛事19场至少2球，概率达到95%，其中12场大2.5球。

At 23: 00 Beijing time on August 25, 2019, in the second round of La Liga, Alaves played at home against the Spaniard. There was a certain chance to threaten each other's goal, but due to luck, both sides did not break the goal, and the two sides were tied. The Spaniard's recent game was not bad within 2 goals, with at least 2 goals in 19 games in nearly 20 games, with a probability of 95%, including 12 big 2.5 goals.

Figure 2: User Interface of Xiaomingbot. On the left is a piece of sports news, which is generated from a Table2Text model. On the top is the text summarization result. On the bottom right corner, Xiaomingbot produces the corresponding speech and visual effects.

text summarization technique to create summaries for users to skim through. Xiaomingbot can also translate news so that people from different countries can promptly understand the general meaning of an article. Xiaomingbot is equipped with a cross lingual voice reader that can read the report in different languages in the same voice. It is worth mentioning that Xiaomingbot excels at voice cloning. It is able to learn a person's voice from audio samples that are as short as only two hours, and maintain precise consistency in using that voice even when reading in different languages. In this work, we recorded 2 hours of Chinese voice data from a female speaker, and Xiaomingbot learnt to speak in English and Japanese with the same voice. Finally, the animation module produces an animated cartoon avatar with lip and facial expression synchronized to the text and voice. It also generates the full body with animated cloth texture. The demo video is available at `https://www.youtube.com/watch?v=zNfaj_DV6-E`. The home page is available at `https://xiaomingbot.github.io`.

The system has the following advantages: *a)* It produces timely news reports for certain areas and is multilingual. *b)* By employing a voice cloning model to Xiaomingbot's neural cross lingual voice reader, we've allowed it to learn a voice in different languages with only a few examples *c)* For better user experience, we also applied cross lingual visual rendering model, which generates synthesis lip syncing in consistent with the generated voice. *d)* Xiaomingbot has been put into practice and produced over $600,000$ articles, and gained over 150k followers in social media platforms.

2 System Architecture

The Xiaomingbot system includes four components working together in an pipeline, as shown in Figure 1. The system receives input from data table containing event records, which, depending on the domain, can be either a sports game with time-line information, or a financial piece such as tracking stock market. The final output is an animated avatar reading the news article with a synthesized voice. Figure 2 illustrates an example of our Xiaomingbot system. First, the text generation model generates a piece of sports news. Then, as is shown on the top of the figure, the text summarization module trims the produced news into a summary, which can be read by users who prefer a condensed abstract instead of the whole news. Next, the machine translation module will translate the summary into the language that the user specifies, as illustrated on the bottom right of the figure. Relying on the text to speech (TTS) module, Xiaomingbot can read both the summary and its translation in different languages using the same voice. Finally, the system can visualize an animated character with synchronized lip motion and facial expression, as well as lifelike body and clothing.

3 News Generation

In this section, we will first describe the automated news generation module, followed by the news summarization component.

3.1 Data-To-Text Generation

Our proposed Xiaomingbot is targeted for writing news for domains with structured input data, such as sports and finance. To generate reasonable text, several methods have been proposed(Miao et al., 2019; Sun et al., 2019; Ye et al., 2020). However, since it is difficult to generate correct and reliable content through most of these methods, we employ a template based on table2text technology to write the articles.

Table 1 illustrates one example of soccer game data and its generated sentences. In the example, Xiaomingbot retrieved the tabled data of a single sports game with time-lines and events, as well as statistics for each player's performance. The data table contains time, event type (scoring, foul, etc.), player, team name, and possible additional attributes. Using these tabulated data, we integrated and normalized the key-value pair from the table. We can also obtain processed key-value pairs such as "Winning team", "Lost team", "Winning Score" , and use template-based method to generate news from the tabulated result. Those templates are written in a custom-designed java-script dialect. For each type of the event, we manually constructed multiple templates and the system will randomly pick one during generation. We also created complex templates with conditional clauses to generate certain sentences based on the game conditions. For example, if the scores of the two teams differ too much, it may generate "Team A overwhelms Team B." Sentence generation strategy are classified into the following categories:

- **Pre-match Analysis.** It mainly includes the historical records of each team.
- **In-match Description.** It describes most important events in the game such as "someone score a goal", "someone received yellow card".
- **Post-match Summary.** It's a brief summary of this game , while also including predictions of the progress of the subsequent matches.

3.2 Text Summarization

For users who prefer a condensed summary of the report, Xiaomingbot can provide a short gist version using a pre-trained text summarization model. We choose to use the said model instead of generating the summary directly from the table data because the former can create more general content, and can be employed to process manually written reports as well. There are two approaches to summarize a text: extractive and abstractive summarization. Extractive summarization trains a sentence selection model to pick the important sentences from an input article, while an abstractive summarization will further rephrase the sentences and explore the potential for combining multiple sentences into a simplified one.

We trained two summarization models. One is a general text summarization using a BERT-based sequence labelling network. We use the TTNews dataset, a Chinese single document summarization dataset for training from NLPCC 2017 and 2018 shared tasks (Hua et al., 2017; Li and Wan, 2018). It includes 50,000 Chinese documents with human written summaries. The article is separated into a sequence of sentences. The BERT-based summarization model output 0-1 labels for all sentences.

In addition, for soccer news, we trained a special summarization model based on the commentary-to-summary technique (Zhang et al., 2016). It considers the game structure of soccer and handles important events such as goal kicking and fouls differently. Therefore it is able to better summarize the soccer game reports.

4 News Translation

In order to provide multilingual news to users, we propose using a machine translation system to translate news articles. In our system, we pre-trained several neural machine translation models, and employ state of the art Transformer Big Model as our NMT component. The parameters are exactly the same with (Vaswani et al., 2017). In order to further improve the system and speed up the inference, we implemented a CUDA based NMT system, which is 10x faster than the Tensorflow

Table 1: Examples of Sports News Generation

Time	Category	Player	Team	Generated Text	Translated Text
23’	Score	Didac	Espanyol	第23分钟，西班牙人迪达克打入一球。	In the 23rd minute, Espanyol Didac scored a goal.
35’	Yellow Card	Mubarak	Alavés	第35分钟，阿拉维斯穆巴拉克吃到一张黄牌。	In the 35th minute, Alavés Mubarak received a yellow card.

approach [1]. Furthermore, our machine translation system leverages named-entity (NE) replacement for glossaries including team name, player name and so on to improve the translation accuracy. It can be further improved by recent machine translation techniques (Yang et al., 2020; Zheng et al., 2020).

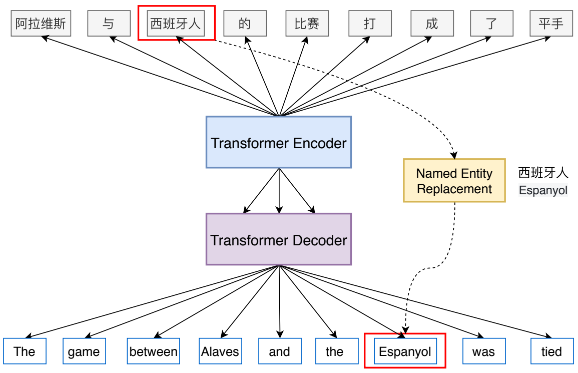


Figure 3: Neural Machine Translation Model.

We use the in-house data to train our machine translation system. For Chinese-to-English, the dataset contains more than 100 million parallel sentence pairs. For Chinese-to-Japanese, the dataset contains more than 60 million parallel sentence pairs.

5 Multilingual News Reading

In order to read the text of the generated and/or translated news article, we developed a text to speech synthesis model with multilingual capability, which only requires a small amount of recorded voice of a speaker in one language. We developed an additional cross-lingual voice cloning technique to clone the pronunciation and intonation. Our cross-lingual voice cloning model is based on Tacotron 2 (J. Shen, 2018), which uses an attention-based sequence-to-sequence model to generate a sequence of log-mel spectrogram frames from an input text sequence (Wang et al., 2017). The architecture is illustrated in Figure 4, we made the following augmentations on the base Tacotron 2 model:

- We applied an additional speaker as well as language embedding to support multi-speaker and multilingual input.
- We introduced a variational autoencoder-style residual encoder to encode the variational length mel into a fix length latent representation, and then conditioned the representation to the decoder.
- We used Gaussian-mixture-model (GMM) attention rather than location-sensitive attention.
- We used wavenet neural vocoder (Oord et al., 2016).

For Chinese TTS, we used hundreds of speakers from internal automatic audio text processing toolkit, for English, we used libritts dataset (Zen et al., 2019), and for Japanese we used JVS corpus which includes 100 Japanese speakers. As for input representations, we used phoneme with tone for Chinese, phoneme with stress for English, and phoneme with mora accent for Japanese. In our experiment, we recorded 2 hours of Chinese voice data from an internal female speaker who speaks only Chinese for this demo.

6 Synchronized Avatar Animation Synthesis

We believe that lifelike animated avatar will make the news broadcasting more viewer friendly. In this section, we will describe the techniques to render the animated avatar and to synchronize the lip and facial motions.

---

[1] https://github.com/bytedance/byseqlib

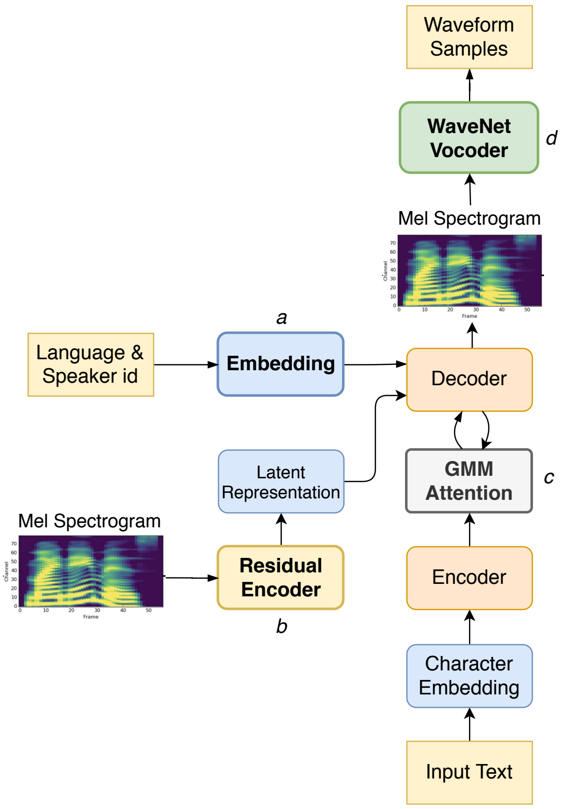


Figure 4: Voice Cloning for Cross-lingual Text-to-Speech Synthesis.

6.1 Lip Syncing

The avatar animation module produces a set of lip motion animation parameters for each video frame, which is synced with the audio synthesized by the TTS module and used to drive the character.

Since the module should be speaker agnostic and TTS-model-independent, no audio signal is required as input. Instead, a sequence of phonemes and their duration is drawn from the TTS module and fed into the lip motion synthesis module. This step can be regarded as tackling a sequence to sequence learning problem. The generated lip motion animation parameters should be able to be re-targeted to any avatar and easy to visualize by animators. To meet this requirement, the lip motion animation parameters are represented as blend weights of facial expression blendshapes. The blendshapes for the rendered character are designed by an animator according to the semantic of the blendshapes. In each rendered frame, the blendshapes are linear blended with the weights predicted by the module to form the final 3D mesh with correct mouth shape for rendering.

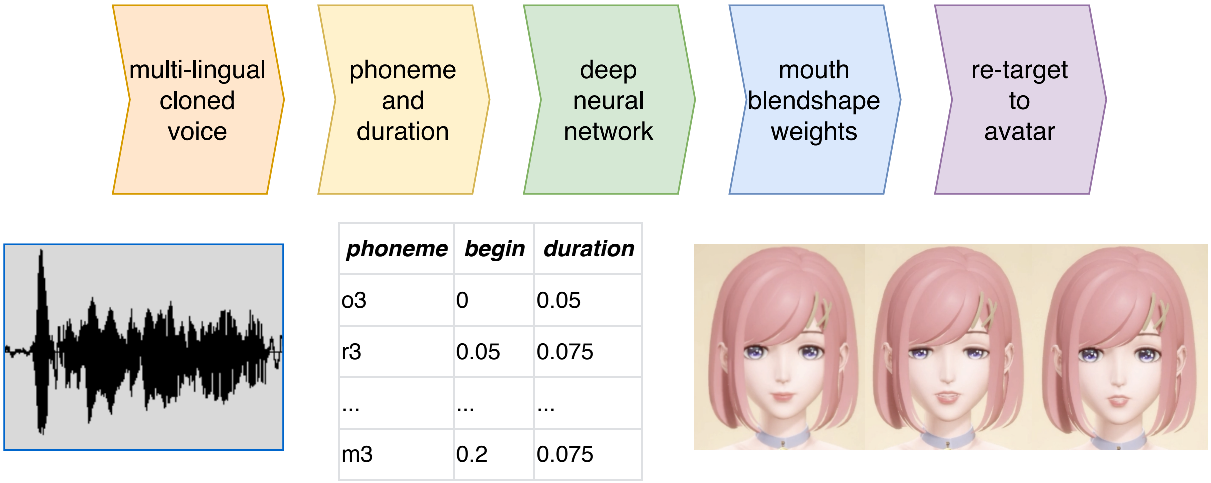
Since the module should produce high fidelity animations and run in real-time, a neural network model that has learned from real-world data is introduced to transform the phoneme and duration sequence to the sequence of blendshape weights. A sliding window neural network similar to Taylor et al. (2017), which is used to capture the local phonetic context and produce smooth animations. The phoneme and duration sequence is converted to fixed length sequence of phoneme frame according to the desired video frame rate before being further converted to one-hot encoding sequence which is taken as input to the neural network in a sliding widow the length of which is 11. Three are 32 mouth related blendshape weights predicted for each frame in a sliding window with length of 5. Following Taylor et al. (2017), the final blendshape weights for each frame is generated by blending every predictions in the overlapping sliding windows using the frame-wise mean.

The model we used is a fully connected feed forward neural network with three hidden layers and 2048 units per hidden layer. The hyperbolic tangent function is used as activation function. Batch normalization is used after each hidden layer (Ioffe and Szegedy, 2015). Dropout with probability of 0.5 is placed between output layer and last hidden layer to prevent over-fitting (Wager et al., 2013). The network is trained with standard mini-batch stochastic gradient descent with mini-batch size of 128 and learning rate of 1e-3 for 8000 steps.

The training data is build from 3 hours of video and audio of a female speaker. Different from Taylor et al. (2017), instead of using AAM to parameterize the face, the faces in the video frames are parameterized by fitting a blinear 3D face morphable model inspired by Cao et al. (2013) built from our private 3D capture data. The poses of the 3D faces, the identity parameters and the weights of the individual-specific blendshapes of each frame and each view angle are joint solved with a cost function built from reconstruction error of the facial landmarks. The identity parameters are shared within all frames and the weights of the blendshapes are shared through view angles which have the same timestamp. The phoneme-duration sequence and the blendshape weights sequence are used to train the sliding window neural network.

6.2 Character Rendering

Unity, the real time 3D rendering engine is used to render the avatar for Xiaomingbot.



phoneme	***begin***	***duration***
o3	0	0.05
r3	0.05	0.075
...	...	...
m3	0.2	0.075

Figure 5: Avatar animation synthesis: a) multi-lingual voices are cloned. b) A sequence of phonemes and their duration is drawn from the voices. c) A sequence of blendshape weights is transformed by a neural network model. d) Lip-motion is synthesized and re-targeted synchronously to avatar animation.

For eye rendering, we used Normal Mapping to simulate the the iris, and Parallax Mapping to simulate the effect of refraction. As for the highlights of the eyes, we used the GGX term in PBR for approximation. In terms of hair rendering, we used the kajiya-kay shading model to simulate the double highlights of the hair (Kajiya and Kay, 1989), and solved the problem of translucency using a mesh-level triangle sorting algorithm. For skin rendering, we used the Separable Subsurface Scattering algorithm to approximate the translucency of the skin (Jimenez et al., 2015). For simple clothing materials, we used the PBR algorithm directly. For fabric and silk, we used Disney's anisotropic BRDF (Burley and Studios, 2012).

Since physical-based cloth simulation algorithm is more expensive for mobile, we used the Spring-Mass System(SMS) for cloth simulation. The specific method is to generate a skeletal system and use SMS to drive the movement of bones (Liu et al., 2013). However, the above approach may cause the clothing to overlap the body. To address this problem, we deployed some new virtual bone points to the skeletal system, and reduced the overlay using the CCD IK method (Wang and Chen, 1991), which displayed great performance in most cases.

7 Conclusion

In this paper, we present Xiaomingbot, a multilingual and multi-modal system for news reporting. The entire process of Xiaomingbot's news reporting can be condensed as follows. First, it learns how to write news articles based on a template based table2text technology, and summarize the news through an extraction based method. Next, its system translates the summarization into multiple languages. Finally, the system produces the video of an animated avatar reading the news with synthesized voice. Owing to the voice cloning model that can learn from a few Chinese audio samples, Xiaomingbot can maintain consistency in intonation and voice projection across different languages. So far, Xiaomingbot has been deployed online and is serving users.

The system is but a first attempt to build a fully functional robot reporter capable of writing, speaking, and expressing with motion. Xiaomingbot is not yet perfect, and has limitations and room for improvement. One such important direction for future improvement is the expansion of areas that it can work in, which can be achieved through a promising approach of adopting model based technologies together with rule/template based ones. Another direction for improvement is to further enhance the ability to interact with users via a conversation interface.

Acknowledgments

We would like to thank Yuzhang Du, Lifeng Hua, Yujie Li, Xiaojun Wan, Yue Wu, Mengshu Yang, Xiyue Yang, Jibin Yang, and Tingting Zhu for helpful discussion and design of the system. The name Xiaomingbot was suggested by Tingting Zhu in 2016. We also wish to thank the reviewers for their insightful comments.

References

Brent Burley and Walt Disney Animation Studios. 2012. Physically-based shading at disney. In *ACM SIGGRAPH*, volume 2012, pages 1–7.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425.

Lifeng Hua, Xiaojun Wan, and Lei Li. 2017. Overview of the NLPCC 2017 shared task: Single document summarization. In *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings*, volume 10619 of *Lecture Notes in Computer Science*, pages 942–947. Springer.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 448–456.

R. J. Weiss M. Schuster N. Jaitly Z. Yang Z. Chen Y. Zhang Y. Wang R. Skerry-Ryan et al. J. Shen, R. Pang. 2018. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *ICASSP*.

Jorge Jimenez, Károly Zsolnai, Adrian Jarabo, Christian Freude, Thomas Auzinger, Xian-Chun Wu, Javier der Pahlen, Michael Wimmer, and Diego Gutierrez. 2015. Separable subsurface scattering. *Comput. Graph. Forum*, 34(6):188–197.

James T Kajiya and Timothy L Kay. 1989. Rendering fur with three dimensional textures. *ACM Siggraph Computer Graphics*, 23(3):271–280.

Lei Li and Xiaojun Wan. 2018. Overview of the NLPCC 2018 shared task: Single document summarization. In *Natural Language Processing and Chinese Computing - 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26-30, 2018, Proceedings, Part II*, volume 11109 of *Lecture Notes in Computer Science*, pages 457–463. Springer.

Tiantian Liu, Adam W Bargteil, James F O'Brien, and Ladislav Kavan. 2013. Fast simulation of mass-spring systems. *ACM Transactions on Graphics (TOG)*, 32(6):1–7.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. CGMH: Constrained sentence generation by metropolis-hastings sampling. In *the 33rd AAAI Conference on Artificial Intelligence (AAAI)*.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Zhaoyue Sun, Jiaze Chen, Hao Zhou, Deyu Zhou, Lei Li, and Mingmin Jiang. 2019. GraspSnooker: Automatic Chinese commentary generation for snooker videos. In *the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 6569–6571. Demos.

Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):1–11.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Stefan Wager, Sida Wang, and Percy S Liang. 2013. Dropout training as adaptive regularization. In *Advances in neural information processing systems*, pages 351–359.

L-CT Wang and Chih-Cheng Chen. 1991. A combined optimization method for solving the inverse kinematics problems of mechanical manipulators. *IEEE Transactions on Robotics and Automation*, 7(4):489–499.

Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 4006–4010.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020. Towards making the most of BERT in neural machine translation. In *the 34th AAAI Conference on Artificial Intelligence (AAAI)*.

Rong Ye, Wenxian Shi, Hao Zhou, Zhongyu Wei, and Lei Li. 2020. Variational template machine for data-to-text generation. In *International Conference on Learning Representations (ICLR)*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1526–1530.

Jianmin Zhang, Jin-ge Yao, and Xiaojun Wan. 2016. Towards constructing sports news from live text commentary. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1361–1371, Berlin, Germany. Association for Computational Linguistics.

Zaixiang Zheng, Hao Zhou, Shujian Huang, Lei Li, Xin-Yu Dai, and Jiajun Chen. 2020. Mirror-generative neural machine translation. In *International Conference on Learning Representations*.