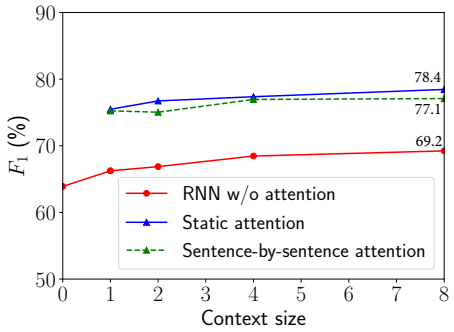

Figure 2: Effect of the context size and different attention mechanisms. The context size is the number of utterances, excluding the critical sentence, on each side of the decision point; a size of 0 refers to no context.

denoted as *non-hierarchical* in Table 1. Using contextual information yields an F_1 improvement of 2%. This controlled experiment validates the usefulness of context for SCD. The hierarchical RNN introduced in Sections 3.1 and 3.2 further improves the F_1-measure by 3%. With sentence-level static attention, our model achieves the highest performance of 89.2% accuracy and 78.4% F_1-measure.

We would like to have in-depth analysis regarding how the context size and different attention mechanisms affect our model. The context size was chosen by validation from $\{1, 2, 4, 8\}$.[4] As shown in Figure 2, even a single context sentence (on each side of the decision point) improves the performance by 2%; with more surrounding utterances, the performance grows gradually. Moreover, attention-based neural networks significantly outperform non-attention models by a margin of 10%. We also tried a dynamic sentence-by-sentence attention mechanism, similar to most existing work [1]. As analyzed in Section 3.3, such model buries critical sentences and thus slightly hurts the performance by 1–2% F_1-measure (green dashed line in Figure 2). The experiments verify the effectiveness of our hierarchical RNN with sentence-level, static attention.

Case study. Table 2 showcases a dialog snippet. In the example, our model makes an error as it fails to detect the change between Sentences 4–5. However, it is even hard for humans to judge this particular change, because the word *absolutely* also goes fluently into the next sentence. For other utterances with more substance, the neural network correctly joins Sentences 1–2 and 5–6, as well as segments Sentences 2–3 and 3–4, showing that our proposed model can effectively capture the semantics of these sentences.

5 CONCLUSION

In this paper, we proposed a static sentence-level attention LSTM-RNN for text-based speaker change detection. Our model uses an LSTM-RNN to encode each utterance into a vector, based on which another LSTM-RNN integrates contextual information, before and after a particular decision point, respectively. A static sentence-level attention mechanism is also applied to enhance information interaction. We crawled dialog transcripts from Cable News Network

[4]Due to efficiency concerns, we did not try larger context sizes.

ID	Utterances	Speaker Changes? Predicted	Truth	Correct?
1	there 's no question the deficit halts on both sides of the aisles .	No	No	✓
2	cbo , wall street , everyone will have a say into this , including workers and future retirees .	Yes	Yes	✓
3	and your saying it 's both parties ?	Yes	Yes	✓
4	absolutely .	No	Yes	✗
5	the key , though , is the first six months .	No	No	✓
6	you say it 's not going to get through in these first couple months .			

Table 2: Case study. Colors indicate speaker identities (they are used to infer groundtruth speaker changes, but cannot be seen during prediction).

TV talk shows for evaluation. Experimental results demonstrate the effectiveness of our approach. In particular, in-depth analysis validates that contextual information is indeed helpful for speaker change detection, and that our tailored model can make better use of context than other neural networks.

REFERENCES

[1] D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.
[2] S. Bowman, G. Angeli, C. Potts, and C. Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*. 632–642.
[3] Y. Chen, M. Sun, A. I Rudnicky, and A. Gershman. 2016. Unsupervised user intent modeling by feature-enriched matrix factorization. In *ICASSP*. 6150–6154.
[4] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel. 2000. Strategies for automatic segmentation of audio data. In *ICASSP*. 1423–1426.
[5] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL*. 110–119.
[6] J. Li, T. Luong, and D. Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *ACL-IJCNLP*. 1106–1115.
[7] R. Li, T. Schultz, and Q. Jin. 2009. Improving speaker segmentation via speaker identification and text segmentation.. In *INTERSPEECH*. 904–907.
[8] R. Lowe, N. Pow, L. Serban, and J. Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL*. 285–294.
[9] L. Lu and H. Zhang. 2005. Unsupervised speaker segmentation and tracking in real-time audio content analysis. *Multimedia Systems* 10, 4 (2005).
[10] Anguera M., S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals. 2012. Speaker diarization: A review of recent research. *TASLP* 20, 2 (2012), 356–370.
[11] T. Rocktäschel, E. Grefenstette, M. Hermann, T. Kočiskỳ, and P. Blunsom. 2016. Reasoning about entailment with neural attention. In *ICLR*.
[12] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *ICLR*.
[13] S. Shen and H. Lee. 2016. Neural Attention Models for Sequence Classification: Analysis and Application to Key Term Extraction and Dialogue Act Detection. In *INTERSPEECH*. 2716–2720.
[14] R. Sinha, S. Tranter, M. Gales, and P. Woodland. 2005. The Cambridge University March 2005 speaker diarisation system. In *INTERSPEECH*. 2437–2440.
[15] Y. Song, L. Mou, R. Yan, L. Yi, Z. Zhu, X. Hu, and M. Zhang. 2016. Dialogue session segmentation by embedding-enhanced TextTiling. In *INTERSPEECH*. 2706–2710.
[16] J. Tiedemann. 2009. News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In *RANLP*. 237–248.
[17] O. Vinyals and Q. Le. 2015. A neural conversational model. In *ICML Workshop*.
[18] P. Wang, L. Ji, J. Yan, L. Jin, and W. Ma. 2016. Learning to extract conditional knowledge for question answering using dialogue. In *CIKM*. 277–286.
[19] R. Yan, Y. Song, X. Zhou, and H. Wu. 2016. Shall I be your chat companion? Towards an online human-computer conversation system. In *CIKM*. 649–658.