Making Sense of Subtitles: Sentence Boundary Detection and Speaker Change Detection in Unpunctuated Texts

Gregor Donabauer
gregor.donabauer@stud.uni-regensburg.de
University of Regensburg
Regensburg, Germany

Udo Kruschwitz
udo.kruschwitz@ur.de
University of Regensburg
Regensburg, Germany

David Corney
david.corney@fullfact.org
Full Fact
London, UK

ABSTRACT

The rise of deep learning methods has transformed the research area of natural language processing beyond recognition. New benchmark performances are reported on a daily basis ranging from machine translation to question-answering. Yet, some of the unsolved practical research questions are not in the spotlight and this includes, for example, issues arising at the interface between spoken and written language processing.

We identify sentence boundary detection and speaker change detection applied to automatically transcribed texts as two NLP problems that have not yet received much attention but are nevertheless of practical relevance. We frame both problems as binary tagging tasks that can be addressed by fine-tuning a transformer model and we report promising results.

ACM Reference Format:
Gregor Donabauer, Udo Kruschwitz, and David Corney. 2021. Making Sense of Subtitles: Sentence Boundary Detection and Speaker Change Detection in Unpunctuated Texts. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3442442.3451894

1 INTRODUCTION

Text and speech processing are closely related research areas, yet one still gets the impression that research is conducted in two separate communities (and if you add video as another mode, then you get another research community). Some of the interesting problems can therefore be found at the boundary of the different fields. While our research is firmly rooted in text processing, we see our work as a contribution to help bridge the gap between work conducted on written and spoken language.

The immediate motivation for our work comes from the domain of *fact checking*. Fact checkers monitor the media to identify potentially harmful or misleading claims. It is important for them to know who said what and when in order to find claims worth investigating. To cope with the volume of potential claims, and the limited time available, fact checkers are increasingly turning to technology to help, including NLP [1]. These tools can help identify claims worth checking, find repeats of claims that have already been checked or even assist in the verification process directly. Most such tools rely on text as input and require the text to be split into sentences.

Some media sources, such as official parliamentary reports, are very rich, providing marked-up text showing sentence and speech boundaries and tag each speaker with a unique identifier. Newspapers and social media usually give some information about speakers though often implicitly or ambiguously. In contrast, audio and video feeds – including TV and radio news broadcasts and videos shared on YouTube or Facebook – do not usually contain explicit information about speakers. In some cases, automatic captioning may be used to generate a transcript, or subtitles may be made available by broadcasters. But in many cases, using post-hoc speech-to-text processing is the only way to extract text.

There is thus a need to bridge the gap between large volumes of audio-visual content and the existing text-based tools that fact checkers use. Our work addresses two aspects of this gap, namely detecting sentence boundaries in transcripts of speech and detecting when the speaker changes, such as during an interview or debate.

Figure 1 illustrates the absence of text structure (including capitalisation and punctuation) as well as conversational structure, as the result of automatic transcription.[1]

> the taxi drivers are on strike again what for they want the government to reduce the price of the gasoline it is really a hot potato we've managed to reduce our energy consumption in our factory by about 15 per cent in the last two years that's excellent how have you managed that mainly because we've invested in a heat recovery system what does that mean exactly well, we use the exhaust gases from our printing presses to provide energy to heat our dryers

Figure 1: Auto-generated subtitles on YouTube.

Figure 2 shows the same example dialogue as in Figure 1 but with the full sentence and conversational structure in place, making it far easier to read and process.

The problem we address is the restoration of some fundamental structure from unpunctuated text data, particularly in the context of transcribed speech and conversation data. In a first step we *restore sentence boundary information*. Sentences are generally considered a fundamental information unit of written text, e.g. [7, 9]. Therefore, this task has been well-studied, e.g. in the context of automatic speech recognition [6, 23–25, 28]. Frequently, the problem of sentence boundary detection in unpunctuated text is treated as a tagging task tackled using IOB sequence labeling [4, 8] as also used in named entity recognition (NER) [19].

[1] This example is not taken from a fact-checking use case but adopted from one of the benchmark collections we use in our experimental work.

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3451894

P1: The taxi drivers are on strike again.
P2: What for?
P1: They want the government to reduce the price of the gasoline.
P2: It is really a hot potato.
P1: We've managed to reduce our energy consumption in our factory by about 15 per cent in the last two years.
P2: That's excellent. How have you managed that?
P1: Mainly because we've invested in a heat recovery system.
P2: What does that mean exactly?
P1: Well, we use the exhaust gases from our printing presses to provide energy to heat our dryers.

Figure 2: Sample from the DailyDialog dataset.

As a subsequent task we want to *restore information on speaker changes* based on the previously identified sentences – as much transcribed data is based on more than a single speaker (as seen with the earlier example). We therefore want to detect whether the next sentence was uttered by the same person or not, an important step in the context of dialogue data restoration and necessary for further postprocessing in this area, e.g. [27].

Given the impressive advances in a variety of NLP tasks using a transformer-based architecture, e.g. [3], we use this approach to tackle the problem at hand. More specifically, we treat both steps as sequence tagging tasks using binary labels by fine-tuning BERT and we compare our work against strong baselines on previously used benchmarks.

By making all our resources (code and test collections) available our aim is to provide a solid reference point and a strong benchmark for future work.

2 RELATED WORK

We will briefly discuss each of the two problems in turn, i.e. Sentence Boundary Detection (SBD) and Speaker Change Detection (SCD).

2.1 Sentence Boundary Detection (SBD)

SBD is an important and well-studied text processing step but it typically relies on the presence of punctuation within the input text [7]. Even with such punctuation it can be a difficult task, e.g. [5, 20], and traditional approaches use a variety of architectures including CRFs [12] and combinations of HMMs, maximum likelihood as well as maximum entropy approaches [11]. With unpunctuated texts (and lack of word-casing information) it becomes a lot harder as even humans find it difficult to determine sentence boundaries in this case [23], as illustrated in Figure 1. Song et al. [22] simplify the problem we are addressing by aiming to detect the sentence boundary within a 5-word chunk – using YouTube subtitle data. Using LSTMs they report an F1 of 81.43% at predicting the position of the sample's sentence boundaries but did not consider any chunks without sentence boundary. Le [8] presents a hybrid model (using BiLSTMs and CRFs) originally used for NER that was evaluated on SBD in the context of conversational data by preprocessing the CornellMovie-Dialogue and the DailyDialog datasets to obtain samples that neither contain sentence boundary punctuation nor word-casing information (they also predict whether the sentence is a statement or a question). They report F1-scores of 81.62% for questions and 91.90% for statements on the CornellMovie-Dialogue data and 94.66% (questions) and 96.29% (statements) on DailyDialog. To the best of our knowledge, only Du et al. [4] present a transformer-based approach to the problem, but they assume partially punctuated text and word-casing information.

Hence, Le [8] and Song et al. [22] appear to be the strongest baselines to compare our approach against.

2.2 Speaker Change Detection (SCD)

Most related work in this area is concerned with audio-based SCD [2, 13, 14, 18] with the exception of Meng et al. [16] who collected transcribed conversations. The text-data is pre-processed to lowercase and contains punctuation. They compare different deep learning approaches with the best-performing being a RNN with LSTM layers, hierarchical context and static attention giving an F1-score of 78.4%. Apart from this work, there are other approaches that treat the topic of text-based SCD, though not explicitly, e.g. Serban and Pineau [21], or they aim at assigning specific speaker ids to sentences [15].

In conclusion, Meng et al. [16] appears to be the most plausible baseline to choose. We will also adopt their benchmark corpus for comparison.

3 METHODOLOGY AND EXPERIMENTAL SETUP

We treat both tasks, SBD and SCD, as sequence labeling tasks. More specifically, we apply IO tagging to label sequences of tokenized text data adopted from NER [7]. In both cases two distinct labels are sufficient to identify whether a token marks the sentence boundary or the start of an utterance by a different speaker, respectively. We use a pre-trained transformer-based language model and fine-tune it on each of the two tasks. The resulting IO sequence taggers allow us to deduce sentence boundaries (SBD-TT) and speaker changes (SCD-TT).

For the experimental setup we opted to fine-tune **BERT-base-uncased** (given our input is expected to be in lowercase, we do not need casing information within the language model). The model training and evaluation are implemented using the PyTorch[2] version of the Python huggingface[3] transformers library. The model's output is produced utilizing a dense layer as classification head. Using the argmax operator, we can deduce labels for resulting vectors in the same dimension as the label list for each introduced token. The processes are executed using three Nvidia GeForce RTX 2080 GPUs with an overall memory size of 24GB. Most experiments are executed in 3 epochs, using a batch-size of 16. The number of epochs is set according to the recommendation of Devlin et al. [3]. Unless specified further down, we refer to our GitHub repository[4] for task-specific sequence lengths, deviations from our parameter settings, all source code, data, models and additional information.

Where appropriate we apply paired t-tests for significance testing (at $p < 0.01$).

[2] https://pytorch.org/
[3] https://huggingface.co/
[4] https://github.com/doGregor/SBD-SCD-pipeline

4 DATASETS

For fair comparison we adopt datasets proposed in prior work. For SBD we use a *Stanford Lectures Dataset* reproduced from Song et al. [22], the *DailyDialog Dataset* proposed by Li et al. [10] and applied in [8]. In addition we also experiment with a hybrid set. For SCD we use the dataset introduced by Meng et al. [16] (we refer to it as *MengCorpus*).

4.1 Stanford Lectures

Song et al. [22] collected the hand-transcribed lecture subtitles provided by Stanford University on YouTube using the text data associated with the lecture series "Natural Language Processing with Deep Learning" and "Human Behavioral Biology". We replicated this process. In addition to that we identified five more lecture series Stanford University provides subtitles for and collected the accompanying text data (resulting in a corpus about 4 times as big). Details on the exact lectures, their source as well as the data themselves can be found on our GitHub repository. For further data pre-processing, we basically adopt the methods introduced by Song et al. [22]. The punctuated transcripts provide ground truth information. We transform all text data to lower-case and tokenize the data using NLTK[5]. Sentences with fewer than 7 or more than 70 words are discarded, and any punctuation is removed. Finally, all tokens including sentence boundary positions are tagged. The text is then split into chunks of 64 token-tag pairs. Sentence boundary tags can appear anywhere within those chunks (which is more generic than the 5-word chunk approach by Song et al. [22]).

The preprocessing steps applied in our work are depicted in Figure 3.

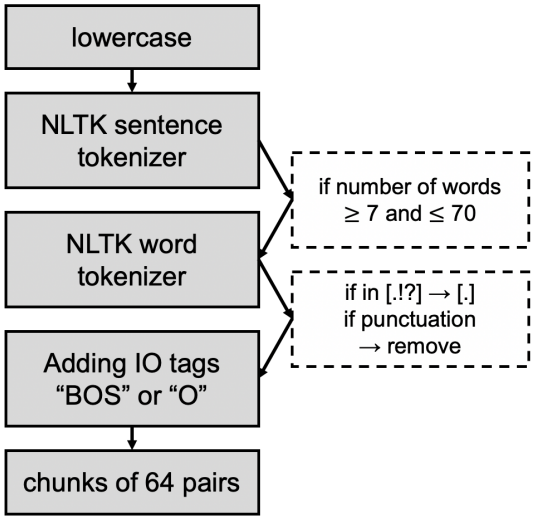

Figure 3: Data Preprocessing for SBD.

In line with convention, we split the data into training (80%), development (10%) and test set (10%) [7]. All samples are saved in CoNLL-2003 format [26] as is frequently used for tagging tasks like NER.

[5]https://www.nltk.org/index.html

Dataset	Train	Dev	Test
Stanford Lectures	19,285	2,411	2,411
DailyDialog	15,259	1,374	1,405
Hybrid Dataset	34,156	3,848	4,142
MengCorpus	174,702	22,065	21,918

Table 1: Number of samples per dataset

	Accuracy	F1-Score
Song et al. [22]	70.84%	81.43%
Le [8]	89.80%	93.07%
SBD-TT	92.49%	93.68%

Table 2: Sentence Boundary Detection applied to *Stanford Lectures* as described by Song et al. [22]

4.2 DailyDialog

The second dataset used, originally introduced by Li et al. [10] captures daily communication with a wide variety of daily life's topics. Since the complete text-data is human written, it is expected to be less noisy than for example automatically transcribed conversational data. The dataset was used for SBD by Le [8] and comes with a 80:10:10 split.

4.3 Hybrid Dataset

To be able to train one single model that can predict sentence boundaries within conversational data as well as a single person's speech data, we create a mixture of the two datasets introduced above. Given the conversational structure of the text we cannot simply randomize development and test sets. Instead we split the data into chunks of 10 sentences each, which are subsequently shuffled. They are concatenated and afterwards split into samples of length 64. Thereby, the structure of subsequent sentences as well as dialogues should be preserved. The basic properties of each dataset are listed in Table 1.

4.4 MengCorpus

For SCD, we use the dataset introduced by Meng et al. [16]. It is a collection of 3,000 hours of hand-transcribed CNN talk-shows. The transcripts provide speaker change information through assigned speaker IDs and comprise approximately 1.5 million utterances. They are split into train, development and test set by an 80:10:10 ratio. The data are provided in form of one sentence per line. As before we use NLTK to perform tokenization and then mark speaker changes accordingly. We split the text into samples of 7 successive sentences to include as much context as possible and satisfy the maximum sequence length of BERT (512 tokens). Hence, the resulting samples have differing lengths in terms of occurring token-tag pairs. They are saved in CoNLL-2003 format. Basic properties are included in Table 1.

Data	Accuracy	F1-Score
Stanford Lectures (big)	97.98%	79.83%
Hybrid Dataset	97.64%	85.31%

Table 3: Sentence Boundary Detection benchmarks applied to our own datasets

5 RESULTS

We will first report the experimental results and then discuss those further in the next Section. All metrics are calculated on a token-based level – in line with what had been adopted in the work we compare our results against.

5.1 Sentence Boundary Detection

Since neither code nor data were available for Song et al. [22], we simply reproduce their accuracy and F1 measures in Table 2. However, as described we also replicated the data collection and processing steps and run 5-fold cross-validation to compare our approach (SBD-TT) against Le [8]. The results can also be found in Table 2. A paired t-test reveals that SBD-TT outperforms Le [8] in terms of both accuracy and F1 (significant at $p < 0.01$).

We also trained and tested SBD-TT on the *DailyDialog* dataset and get an F1 for statements of 97.19% (vs. 96.29% reported by Le [8]) and for questions: 95.64% (vs. 94.66%).

We observe that our SBD-TT approach for sentence boundary detection outperforms state-of-the-art methods and conclude that our vanilla transformer-based approach using BERT leaves scope for further advances.

As an additional contribution and to foster reproducibility we also provide benchmarks obtained from the two corpora we introduced in this paper (and which are available on our GitHub account): the larger Stanford Lectures Dataset and the Hybrid Dataset, and these are reported in Table 3. Note that for these experiments we use sequence lengths of 64 words (unlike the much shorter 5-word chunks used to compare against baselines in Table 2).

Going back to the discussion of related work, one might ask why we did not compare our results against those reported by Du et al. [4]? That is because they treat a similar but different problem. They have word-casing information available and more than 90% of the "end of sentence" tokens. Therefore it was not a suitable comparison.

5.2 Speaker Change Detection

We have two different results to report and to compare against the baseline scores achieved by Meng et al. [16], since we use two different approaches for evaluation. The first evaluation method simply uses 7 successive sentences at a time and tags those with the speaker change labels. The second method also uses 7 successive sentences as an input, but only takes into account the predictions for the middle sentence. All other sentences are seen as context. This sliding-window evaluation is executed with a stride of one sentence at a time. Table 4 presents the results of Meng et al. [16] in comparison to our approach, SCD-TT.

We note that our straightforward fine-tuning approach is competitive for speaker change detection, and for the sliding-window-based evaluation we even achieve a 0.4 percentage point improvement in F1 compared to the best score of Meng et al. [16]. In general, the results show that context is important for the model to predict speaker changes. This confirms the findings described by Meng et al. [16]. While they used 8 sentences of context on each side of the evaluated sentence, we are limited by the maximum sequence length of 512 tokens that can be used as input of our BERT-based model. Therefore we only used 3 sentences on each side of the evaluated sentence as context though were still able to achieve F1 scores slightly higher than those reported by Meng et al. [16].

6 DISCUSSION

There are a number of discussion points emerging from our experimental setup and the results we obtained.

First of all, why did we only test for statistical significance for SBD-TT? The reason is that while we were able to reproduce the work of Le [8], for Song et al. [22] we did not have the code nor the exact data. We requested both from the authors but did not get a response (hence the comparison against reported results only). Unlike for SBD, where we achieved a new state-of-the-art performance, in SCD our results are on par with the best-performing alternative so we only compare against the performance reported by Meng et al. [16].

Also, why did we not combine the two methods, to first detect sentence boundaries and then detect speaker change over that result? Our overall aim was to demonstrate the general suitability of our approach and provide strong benchmarks for each of the two problems rather than providing the best possible model that combines both. As such it is possible to use the models separately where necessary, e.g., when processing a single person's transcript we do not need the SCD model. Obviously this leaves plenty of room for future investigations, and the provision of all resources on GitHub will support this.

Another question arising from our strong performance against state-of-the-art approaches (keeping in mind that we are using a relatively straightforward architecture) is to ask what kind of knowledge does our approach encode that the other approaches don't? We would argue that BERT clearly encodes exactly the type of contextual information that is needed for the two tasks. This information is captured implicitly and obtained partly during fine-tuning but also in pre-training. Note again that our aim was to demonstrate the general suitability of a transformer-based approach. Using other BERT-based models as well as better fine-tuning can result in further improvements.

Finally, one might ask whether a performance that is on par with other state-of-the-art approaches (as is the case for speaker change detection) gives us any benefit. Well, in addition to the points just raised we should also point out that we get better (SBD) or similar (SCD) results for both tasks with a much simpler model. Given we only apply a very simple setup (e.g. only using BERT-base) there is potential to push the effectiveness without losing the overall simplicity.

Coming back to the initial example, Figure 4 demonstrates the output generated by applying each of our two models. While the

Model	Accuracy	Precision	Recall	F1-Score
Random guess	61.8%	26.0%	25.0%	25.4%
Logistic Regression w/ (uni+bi)-gram	80.5%	73.0%	39.0%	50.9%
DNN w/ (uni+bi)-gram	76.6%	54.4%	58.8%	56.5%
CNN w/o context	77.8%	56.8%	58.9%	57.8%
RNN w/o context	83.3%	72.5%	57.1%	63.9%
RNN w/ context (non-hierarchical)	83.7%	72.6%	60.0%	65.7%
RNN w/ context (hierarchical)	85.1%	74.6%	64.6%	69.2%
SCD-TT w/o sliding window evaluation	82.4%	76.2%	72.1%	74.1%
RNN w/ context (hierarchical) + static attention	**89.2%**	**81.5%**	75.6%	78.4%
SCD-TT w/ sliding window evaluation	85.4%	80.1%	**77.6%**	**78.8%**

Table 4: Results of Speaker Change Detection in comparison to scores reported by Meng et al. [16]

sentence segmentation works perfectly in this case, we see that speaker change detection (predicted by the label *True*) leaves room for improvement: lines 4, 8, 9 and 10 are incorrectly classified.

```
('the taxi drivers are on strike again.', True)
('what for.', True)
('they want the government to reduce the price of the gasoline.', True)
('it is really a hot potato.', False)
("we've managed to reduce our energy consumption in our factory by about 15 per
cent in the last two years.", True)
("that's excellent.", True)
('how have you managed that.', True)
("mainly because we've invested in a heat recovery system.", False)
('what does that mean exactly.', False)
('well we use the exhaust gases from our printing presses to provide energy to heat
our dryers.', True)
```

Figure 4: Restored structure of initial example.

7 CONCLUSION

With a bit of delay when compared to image processing, natural language processing has now also witnessed a paradigm shift from traditional statistical approaches to deep learning architectures. This has resulted in some staggering performance improvements across a wide range of applications. However, there are still plenty of open problems – often based on practical use cases. The rapidly evolving mix of different types of media and new forms of interaction highlights the fact that at the interface between different communities, such as those working with spoken and those with written textual data, there are opportunities to make rapid progress. This can be achieved by adopting paradigms that have already been shown to push the state of the art forward elsewhere, most prominently transformer-based architectures.

In this paper we identify the detection of sentence boundaries and speaker changes in unpunctuated text as problems of natural language processing that sit at the interface between spoken and written text, and which have attracted little interest before now. By making our methods available to fact checkers, they may find it as easy to identify and analyse claims made during televised debates or news interviews as claims made in online textual news sites. This will help ensure that no matter where harmful or misleading information is shared, it can also be identified and challenged rapidly to limit its spread. Beyond the work of fact checkers we envisage the proposed steps to be also incorporated in NLP pipelines that will automatically flag up such harmful or misleading information.[6]

We should note that the two tasks could be seen as individual NLP tasks or combined as a sequence of two steps. In our work we frame both tasks as an IO tagging problem that is addressed using fine-tuning of a BERT-based language model.

The results we report demonstrate that the problems at hand are yet another pair of examples where the transformer-based paradigm outperforms existing baselines. There is much scope to push the effectiveness even further as we have only experimented with basic models.

To foster further research we also provide a range of corpora and benchmarks that can be used as future reference points.

ACKNOWLEDGMENTS

This work was supported by the project *COURAGE: A Social Media Companion Safeguarding and Educating Students* funded by the Volkswagen Foundation, grant number 95564.

REFERENCES

[1] Phoebe Arnold. 2020. *The challenges of online fact checking.* Technical Report. Full Fact, London, UK. https://fullfact.org/media/uploads/coof-2020.pdf

[2] S. Chen and P.S. Gopalakrishnan. 1998. Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion. In *DARPA Broadcast News Transcription and Understanding Workshop.* Landsdowne, USA, 127–132.

[3] 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).* Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[4] Jinhua Du, Yan Huang, and Karo Moilanen. 2019. AIG Investments.AI at the FinSBD Task: Sentence Boundary Detection through Sequence Labelling and BERT Fine-tuning. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing.* Association for Computational Linguistics, Macao, China, 81–87. https://www.aclweb.org/anthology/W19-5513

[5] Dan Gillick. 2009. Sentence Boundary Detection and the Problem with the U.S.. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers* (Boulder, Colorado) *(NAACL-Short '09).* Association for Computational Linguistics, USA, 241–244.

[6] As part of the COURAGE research project we are exploring ways to help teenagers manage social media exposure by providing a virtual companion that would, among other things, automatically identify examples of hate speech, bullying or other toxic content [17].

[6] C. C. Juin, R. X. J. Wei, L. F. D'Haro, and R. E. Banchs. 2017. Punctuation prediction using a bidirectional recurrent neural network with part-of-speech tagging. In *TENCON 2017 - 2017 IEEE Region 10 Conference*. IEEE, Penang, Malaysia, 1806–1811.

[7] Daniel Jurafsky and James Martin. 2020. *Speech and Language Processing: An Introduction to Natural Language Processing,Computational Linguistics, and Speech Recognition* (third (draft) ed.). https://web.stanford.edu/~jurafsky/slp3/

[8] The Anh Le. 2020. Sequence Labeling Approach to the Task of Sentence Boundary Detection. In *Proceedings of the 4th International Conference on Machine Learning and Soft Computing* (Haiphong City, Viet Nam) *(ICMLSC 2020)*. ACM, New York, NY, USA, 144–148. https://doi.org/10.1145/3380688.3380703

[9] Joan Persily Levinson. 1985. *Punctuation and the orthographic sentence: a linguistic analysis.* Doctoral dissertation. City University of New York.

[10] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 986–995. https://www.aclweb.org/anthology/I17-1099

[11] Yang Liu, Andreas Stolcke, Elizabeth Shriberg, and Mary Harper. 2004. Comparing and Combining Generative and Posterior Probability Models: Some Advances in Sentence Boundary Detection in Speech. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain, 64–71. https://www.aclweb.org/anthology/W04-3209

[12] Yang Liu, Andreas Stolcke, Elizabeth Shriberg, and Mary Harper. 2005. Using Conditional Random Fields for Sentence Boundary Detection in Speech. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (Ann Arbor, Michigan) *(ACL '05)*. Association for Computational Linguistics, USA, 451–458. https://doi.org/10.3115/1219840.1219896

[13] Lie Lu and Hong-Jiang Zhang. 2002. Speaker Change Detection and Tracking in Real-Time News Broadcasting Analysis. In *Proceedings of the Tenth ACM International Conference on Multimedia* (Juan-les-Pins, France) *(MULTIMEDIA '02)*. Association for Computing Machinery, New York, NY, USA, 602–610. https://doi.org/10.1145/641007.641127

[14] Lie Lu and Hong-Jiang Zhang. 2005. Unsupervised Speaker Segmentation and Tracking in Real-Time Audio Content Analysis. *Multimedia Syst.* 10, 4 (April 2005), 332–343. https://doi.org/10.1007/s00530-004-0160-5

[15] Kaixin Ma, Catherine Xiao, and Jinho D. Choi. 2017. Text-based Speaker Identification on Multiparty Dialogues Using Multi-document Convolutional Neural Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Student Research Workshop*, Allyson Ettinger, Spandana Gella, Matthieu Labeau, Cecilia Ovesdotter Alm, Marine Carpuat, and Mark Dredze (Eds.). Association for Computational Linguistics, Vancouver, Canada, 49–55. https://doi.org/10.18653/v1/P17-3009

[16] Zhao Meng, Lili Mou, and Zhi Jin. 2017. Hierarchical RNN with Static Sentence-Level Attention for Text-Based Speaker Change Detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (Singapore, Singapore) *(CIKM '17)*. ACM, New York, NY, USA, 2203–2206. https://doi.org/10.1145/3132847.3133110

[17] Dimitri Ognibene, Davide Taibi, Udo Kruschwitz, Rodrigo Souza Wilkens, Davinia Hernandez-Leo, Emily Theophilou, Lidia Scifo, Rene Alejandro Lobo, Francesco Lomonaco, Sabrina Eimler, H. Ulrich Hoppe, and Nils Malzahn. 2021. Challenging Social Media Threats using Collective Well-being Aware Recommendation Algorithms and an Educational Virtual Companion. arXiv:2102.04211 [cs.CY]

[18] Sree Hari Krishnan Parthasarathi, Mathew Magimai.-Doss, Daniel Gatica-Perez, and Hervé Bourlard. 2009. Speaker Change Detection with Privacy-Preserving Audio Cues. In *Proceedings of the 2009 International Conference on Multimodal Interfaces* (Cambridge, Massachusetts, USA) *(ICMI-MLMI '09)*. ACM, New York, NY, USA, 343–346. https://doi.org/10.1145/1647314.1647385

[19] Lance Ramshaw and Mitch Marcus. 1995. Text Chunking using Transformation-Based Learning. In *Third Workshop on Very Large Corpora*. Association for Computational Linguistics, Cambridge, MA, USA, 82–94. https://www.aclweb.org/anthology/W95-0107

[20] George Sanchez. 2019. Sentence Boundary Detection in Legal Text. In *Proceedings of the Natural Legal Language Processing Workshop 2019*. Association for Computational Linguistics, Minneapolis, Minnesota, 31–38. https://doi.org/10.18653/v1/W19-2204

[21] Iulian V. Serban and Joelle Pineau. 2015. Text-Based Speaker Identification For Multi-Participant Open-Domain Dialogue Systems. In *Machine Learning for Spoken Language Understanding and Interaction, NIPS 2015 Workshop*. Montreal, Canada. http://slunips2015.wixsite.com/slunips2015/accepted-papers

[22] Hye Jeong Song, Hong Ki Kim, Jong Dae Kim, Chan Young Park, and Yu Seop Kim. 2019. Inter-sentence segmentation of YouTube subtitles using Long-Short Term Memory (LSTM). *Applied Sciences (Switzerland)* 9, 7 (2019). https://doi.org/10.3390/APP9071504

[23] Mark Stevenson and Robert Gaizauskas. 2000. Experiments on sentence boundary detection. In *Proceedings of the sixth conference on Applied natural language processing -*. Association for Computational Linguistics, Morristown, NJ, USA, 84–89. https://doi.org/10.3115/974147.974159

[24] Ottokar Tilk and Tanel Alumäe. 2015. LSTM for punctuation restoration in speech transcripts. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association*. ISCA, Dresden, Germany, 683–687. http://www.isca-speech.org/archive/interspeech_2015/i15_0683.html

[25] Ottokar Tilk and Tanel Alumäe. 2016. Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*, Nelson Morgan (Ed.). ISCA, San Francisco, CA, USA, 3047–3051. https://doi.org/10.21437/Interspeech.2016-1517

[26] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 -*, Vol. 4. Association for Computational Linguistics, Morristown, NJ, USA, 142–147. https://doi.org/10.3115/1119176.1119195

[27] S.E. Tranter and D.A. Reynolds. 2006. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech and Language Processing* 14, 5 (sep 2006), 1557–1565. https://doi.org/10.1109/TASL.2006.878256

[28] K. Xu, L. Xie, and K. Yao. 2016. Investigating LSTM for punctuation prediction. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, Tianjin, China, 1–5.