Live Commentary Planning and Generation

Chung-Chi Chen,[1] **Huan-Wen Ho,**[2] **Yu-Yu Chang,**[2] **Ming-Hung Wang,**[2]
Ramon Ruiz-Dolz,[3] **Chris Reed**,[3] **Ichiro Kobayashi**,[4] **Yusuke Miyao**,[5] **Hiroya Takamura**[1]
[1]AIST, Japan
[2]National Chung Cheng University, Taiwan
[3]Centre for Argument Technology (ARG-tech), University of Dundee, UK
[4]Ochanomizu University, Japan
[5]University of Tokyo, Japan

Abstract

Live commentary plays a crucial role in helping audiences interpret high-stakes events such as political debates, central bank press conferences, and corporate earnings calls. Unlike generic summarization, professional commentary requires timely decisions about *what* to comment on and *how* to present it, integrating fact-checking, background knowledge, and subjective evaluation. However, little prior work has studied commentary as a structured planning and generation problem. To bridge this gap, we introduce the first multi-domain dataset of **Live Commentary Planning and Generation**, aligning event transcripts with time-synchronized expert analyses and public reactions. Our dataset covers U.S. presidential debates (2016–2024), Federal Open Market Committee press conferences, and corporate earnings calls, enriched with a fine-grained taxonomy of commentary intents (up to 11 categories) and supplemented by Reddit crowd commentary. We define two benchmark tasks: (1) *Commentary Planning*, predicting the type of commentary given a transcript segment, and (2) *Commentary Generation*, producing commentary text conditioned on the segment and a target label. Baseline experiments with large language models show that, despite their fluency, models struggle with expert-level commentary, showing the difficulty of integrating contextual reasoning and external knowledge under real-time constraints.[1]

1 Introduction

Large language models (LLMs) have made it easier than ever to generate fluent text, but true professional-quality commentary demands more than fluency. In high-stakes public discourse, such as political debates, central bank press conferences, or corporate earnings calls, expert commentators provide real-time analysis that contextualizes and critiques what is being said. This live commentary helps transform passive viewership into an engaged, informed experience. Professional commentators translate complex language into accessible insights, fact-check claims in real-time, and offer historical or expert perspective to guide audience understanding. Such commentary must be timely, knowledgeable, and context-rich, going beyond summarization to include opinions, fact-checks, and interpretations. However, simulating this expert ability is challenging: it requires deciding what to comment on (planning) and how to convey it (generation) under time pressure and with domain expertise.

Despite extensive research on these domains individually (e.g. analyzing debate transcripts or summarizing financial reports), little work has aligned transcripts with their simultaneous expert commentary. Existing studies tend to treat the primary content and the reactions separately. For example, focusing on debate speeches or on social media responses in isolation. This leaves a gap in understanding how experts interpret dynamic events in the moment. To address this gap, we introduce a new dataset of Live Commentary Planning and Generation that aligns real-time expert commentary with transcripts across three domains: (1) U.S. presidential debates (2016–2024), (2) Federal Open Market Committee (FOMC) press conferences, and (3) corporate earnings calls. In each setting, multiple expert commentators observed the live event and produced running commentary, which we have collected and aligned with the spoken transcript segments by timestamp. Additionally, for the presidential debates, we incorporate public commentary from Reddit discussion threads to capture non-expert, crowd reactions in real time. The result is a multi-faceted dataset covering both institutional expert analysis and grassroots public reactions.

Crucially, each commentary segment is annotated with a fine-grained category from a new taxonomy we developed for live discourse analysis.

[1]Project Page: http://livecommentary.nlpfin.com/

Proceedings of the 18th International Natural Language Generation Conference: Generation Challenges, pages 37–43
October 29–November 2, 2025. ©2025 Association for Computational Linguistics

For example, debate commentary segments are labeled as Key Summary, Supplementary Explanation, Fact-Check, Personal Opinion, Market Reaction, Public Opinion, or Commentator's Question, with Personal Opinion further broken into subtypes like evaluating performance, analyzing claims, drawing inferences, etc.. This rich labeling (11 distinct labels in total for debates) enables models to learn not just to generate commentary, but to plan what type of comment is appropriate at each moment. In professional settings, the ability to choose an apt perspective, e.g. to fact-check a dubious claim versus to summarize a complex point, is critical. Our dataset supports two complementary tasks: (1) Commentary Planning, i.e. predicting the commentary label given a transcript segment, and (2) Commentary Generation, i.e. producing the content of a commentary given the segment and a target label. By tackling these tasks, models must learn to mimic expert decision-making and contextual writing under real-time constraints.

In summary, our contributions are: (a) a first-of-its-kind dataset aligning transcripts with real-time expert commentary across multiple domains, with fine-grained annotations of commentary intent; (b) benchmark task definitions for commentary planning and generation, to facilitate systematic study of this challenging form of conditional text generation; and (c) initial analyses and baseline results demonstrating the dataset's difficulty and the need for advanced techniques. Even state-of-the-art LLMs like GPT-4 struggle with expert commentary planning and generation (as shown in our pilot studies), underscoring the novelty and challenge of our task. We hope this dataset will spur research at the intersection of content understanding, knowledge integration, and real-time text generation.

2 Dataset

2.1 Dataset Creation

Our dataset encompasses three types of live events: (1) U.S. presidential debates, (2) FOMC press conferences, and (3) corporate earnings calls , along with their live commentary. For U.S. presidential debates, we include all major televised debates from the 2016, 2020, and 2024 election cycles. This totals 10 events (including presidential and vice-presidential debates and a 2023 primary debate), with full transcripts obtained from public sources (e.g. debate commission or media outlets). We collected the real-time expert commentary on these debates from the Bloomberg news service, which had professional journalists providing line-by-line analysis during the live broadcasts. Each commentary piece is timestamped. We align each commentary segment to the corresponding part of the debate transcript by timestamp and content, ensuring the commentator's remark is matched with the specific speaker utterance or segment it addresses. If a comment does not clearly relate to any specific line, it is marked as not applicable to a segment. Using this procedure, we obtained 2,283 commentary-transcript pairs for debates.

	Debates	FOMC	Earnings Call	Reddit
# Pair	2,283	252	1,115	366
# Category	11	5	10	4

Table 1: Dataset statistics.

For FOMC press conferences, we collected transcripts of the Fed Chair's opening statement and the subsequent Q&A with journalists, for multiple meetings, covering 8 FOMC events. We again used Bloomberg's real-time commentary feed, which provides expert economist reactions during these press conferences. After alignment, we have 252 commentary segments paired with FOMC transcript segments. For corporate earnings calls, we focus on earnings calls of S&P 500 companies across various sectors. Earnings calls typically consist of a management presentation and a Q&A session with analysts. We use transcripts and align Bloomberg's live financial commentary on those calls. The dataset includes 1,115 pairs of commentary with earnings call transcript segments. Lastly, for Reddit commentary, we incorporate public reactions from Reddit "mega-threads" created during the 2016 U.S. presidential debates. Using an Intertextual Topic Correspondence (ITC) method (Visser et al., 2018), we matched 366 Reddit comments to relevant debate utterances. These alignments were verified and annotated with simplified labels (described below). The inclusion of Reddit allows us to compare expert vs. crowd commentary directly.

2.2 Label Taxonomy

Table 1 summarizes the size of each domain in our dataset and the label inventory available. In total, the dataset contains over 3,650 expert commentary instances aligned with transcripts (plus 366 Reddit instances), making it the largest resource of its kind to date.

Label	Description
KS (Key Summary)	Summarizing what the speaker said.
SE (Supplementary Explanation)	Providing additional factual context or background (often drawing on external knowledge).
FC (Fact-Checking)	Verifying or refuting the accuracy of a candidate's claim.
PO (Public Opinion)	Noting public sentiment or likely voter reactions (sometimes referencing polls or social media).
MR (Market Reaction)	Commenting on any immediate financial market response or economic implications (included since commentators are financial journalists).
CQ (Commentator's Question)	Posing an open question or something to watch for (e.g., "How will candidate X implement this policy?").
CPO (Commentator's Personal Opinion)	Any subjective analysis or evaluative remark by the commentator. Expanded into five finer labels:
PC (Performance Critique)	Evaluating the debate performance or rhetorical style of the participants.
CS (Claim Analysis)	Opining on specific policy claims or factual statements made.
AC (Analytical Conclusion)	Drawing a conclusion or inference beyond the given facts.
MP (Market/Policy Projection)	Connecting the debate content to economic or policy outcomes (e.g., impact on markets).
O (Other)	Any opinion-based comment that doesn't fit the above (catch-all).

Table 2: Commentary labels in debates

Each domain has a tailored commentary taxonomy reflecting the nature of that discourse, while maintaining some common themes. For the debates, we developed a hierarchical label schema with 7 main categories and several subcategories. In total, as shown in Table 2, the debate commentary taxonomy has 11 fine-grained labels (KS, SE, FC, PO, MR, CQ, and the 5 CPO subtypes), which offer a nuanced view of how commentators respond. Table 2 in Appendix provides frequency statistics of these labels per debate event, confirming that summaries and explanations are most common, but all categories are represented.

The FOMC commentary uses a simpler set of 5 categories reflecting its financial focus. We define labels for: Summary of the Fed's statements; Open Question (similar to CQ, when analysts pose a question or uncertainty); and three sentiment-based Opinion labels – Positive, Neutral, Negative – indicating the tone of the commentator's view on the policy or economic outlook. These sentiment opinions replace the more fine-grained CPO subtypes used in debates, since FOMC commentary often centers on evaluative tone (e.g. optimistic vs pessimistic take on the Fed's message). The earnings call commentary required an even more fine-grained scheme of 10 categories. We include labels for various comparative or contextual analyses that financial journalists provide, such as: comparison with previous company reports, discussion of supply chain details, references to prior quarterly calls, noting market expectations vs actual results, and mentions of competitors' performance. These capture the rich analytical moves typical in earnings analysis. Additionally, earnings commentary labels cover summary of the results, open questions (e.g. uncertainties about guidance), general commentary (uncategorized observations), and sentiment opinions (positive/neutral/negative) about the earnings news. By designing domain-specific labels, we account for differences in commentary style: e.g. debate commentary includes fact-checking political claims, while earnings call commentary often involves comparing numbers to expectations or past quarters.

For the Reddit debate comments, we use a simplified 4-category scheme focusing on how the comment relates to the debate utterance. The labels (drawn from prior work on intertextual links in discussions) are: Agreement, Disagreement, Elab-

oration, or Paraphrase. These indicate whether the Reddit user is agreeing with a candidate's point, disputing it, adding more information or opinion, or simply rephrasing it (often humorously or sarcastically). While not as fine-grained as expert labels, these categories let us study the contrast between expert commentary (which may lean towards factual and analytical responses) and public commentary (which may show more partisanship or humor).

3 Task Design and Evaluation

We consider two primary tasks with our dataset, reflecting the pipeline of a commentary system:

3.1 Commentary Planning

Given a segment of the transcript (e.g. a few sentences of a debate or a turn from the Fed Chair), the model must predict which commentary category an expert would choose for a comment on that segment. This is a multi-class classification task over the label set of the respective domain (e.g. 11-way classification for debates). We evaluate planning performance using standard classification metrics, chiefly accuracy and F1-score. Since the class distribution is imbalanced (certain labels like Key Summary occur more frequently, while others like Commentator's Question are rarer), we report both macro-averaged F1 and micro-F1. The latter emphasizes overall correctness, while macro-F1 highlights performance on less common categories. In our pilot experiments, this task proved very challenging: even powerful LLMs achieved only about 46–49% micro-F1 on debate commentary planning. For example, GPT-4 and Claude 3.5 Sonnet models hovered around 0.5 F1. This indicates that identifying what type of comment to make – essentially, the expert's decision-making – requires deeper understanding of context and likely external knowledge. We expect specialized models or additional context (such as preceding dialogue or world knowledge) to be needed to improve on this task.

3.2 Commentary Generation

Here the goal is to generate the content of a commentary given a transcript segment and a specified commentary label. This reflects producing a particular style of comment (e.g. a fact-check) appropriate to what was said. We treat this as a conditional text generation task. Evaluation of generated commentary is nuanced: we compute automatic metrics like ROUGE (measuring n-gram overlap with the reference expert commentary) and BERTScore (measuring semantic similarity to the reference) to get a quantitative sense of fidelity. However, because commentary is an open-ended task (the model could comment in various valid ways that differ from the single reference), these overlap-based scores tend to be low. Indeed, our pilot tests found ROUGE-1/2 scores in the 0.10 range for even the best LLMs, which underscores that divergent but valid outputs are penalized by reference metrics. We therefore place greater emphasis on human evaluation for generation. We propose to have experts or crowd annotators judge generated commentaries along key dimensions of quality: (a) Importance: does the commentary focus on important or relevant aspects of the segment (as an expert would) rather than trivial details? (b) Expectedness/Novelty: does the commentary provide insight beyond merely restating the transcript (since a good comment should add context or analysis, not just the obvious)? (c) Clarity: is the commentary clearly written and easy to understand? (d) Accuracy: are any factual claims in the commentary correct (this is crucial for fact-checking or explanatory comments). We will use rating scales for these dimensions and also collect an overall preference between different model outputs. Additionally, we plan to utilize LLM-based evaluators for automatic judgment: for example, prompting a strong model to assess a generated commentary for coherence and correctness (drawing on the "news value" criteria from journalism studies). This approach of using LLMs as judges, alongside human evaluation, can help scale the assessment of open-ended generation.

4 Expected Challenges

As a benchmark, we evaluated several cutting-edge LLMs on our tasks. For commentary planning, all models struggled; for instance, Claude's F1 was 0.48, similar to GPT-4, while DeepSeek (a 70B-level open model) was slightly lower, indicating that without fine-tuning, these models often misidentify which strategy to use (e.g. they might summarize when a fact-check was needed, or vice versa). For commentary generation, we experimented with prompting LLMs to produce commentary given segments and target labels. Qualitatively, the models can produce fluent and relevant comments, but often lack the expert precision: e.g. a

fact-check generated by GPT-4 might not actually verify the claim with evidence, or a supposed "market reaction" comment by Claude might be generic since the model doesn't have real financial data. The ROUGE scores around 0.1 for all models reflect that the models' outputs often did not overlap with the reference wording, even if they were topically relevant. This is in stark contrast to, say, news summarization tasks where state-of-the-art models can achieve much higher ROUGE by producing similar summaries. The low scores reaffirm that commentary generation is fundamentally different from summarization: it is a more open-ended, many possible answers problem (especially for opinion and explanation categories). Therefore, we caution against relying solely on reference-based metrics. Instead, our evaluation protocol will use a combination of automatic and human measures, as described, to get a well-rounded picture of performance.

Another challenge is the need for external knowledge. In our dataset, commentators frequently bring in outside information, e.g. citing economic data during a debate or recalling previous statements by the Fed, which a model without retrieval may not know. To encourage research on this, we distinguish between closed-book and open-book commentary generation. A closed-book model must rely only on its internal knowledge and the transcript input, while an open-book model can call a retrieval system or database (for example, retrieve relevant fact-checks or Wikipedia content). We will evaluate both settings. We expect that retrieval-augmented approaches will produce more factual and informative commentary, especially for fact-checking and supplementary explanation categories, at the cost of more complex systems. This setup mirrors real journalists, who often quickly search for data or past news while commenting live.

Overall, our evaluation methodology is designed to capture the multi-dimensional goals of live commentary: factual accuracy, relevance, insight, and timeliness. By providing both the planning labels and the generation task, our dataset allows researchers to decompose the problem.

5 Related Work

Generating live commentary has been explored in limited domains such as sports and games. For example, Ishigaki et al. (2021) generated commentary for racing video games using multimodal inputs, and Marrese-Taylor et al. (2022) proposed open-domain video commentary generation from gameplay. These systems focused on describing visual events, whereas our work deals with discursive events (speeches, discussions) and requires integrating factual knowledge and argumentative context. In the news domain, others have studied generating reader comments or transforming content: e.g., Yang et al. (2019) generated news article comments, and Liu et al. (2024) created *SciNews* to turn scientific papers into lay summaries. Our dataset enables similar grounded generation but in real-time political and financial contexts, which pose unique time-sensitivity and accuracy challenges.

U.S. presidential debates are a rich resource for argument mining and claim analysis. Prior datasets have tackled check-worthy claim detection in debates. The CLEF-2018 CheckThat! lab (Atanasova et al., 2018) introduced tasks to identify which debate statements merit fact-checking. Similarly, *ClaimRank* (Jaradat et al., 2018) prioritized factual claims in debates for fact-checkers. These datasets typically provide binary or priority labels on debate sentences indicating "worth fact-checking." For instance, the Check-Worthy corpus by Patwari et al. (2017) annotated debate sentences with whether they should be checked. However, these resources focus narrowly on factual claims, whereas live commentary covers a broader range of reactions (summaries, opinions, etc.) in real time. Our dataset indeed includes fact-checking commentary labels, but situates them among many other commentary types, providing a more comprehensive view of how debates are analyzed on the fly.

Other work has examined the argumentative structure of debates. The *M-Arg* dataset (Mestre et al., 2021) annotated the 2020 U.S. presidential debates for argument relations (support, attack, neutral) using both text and audio. Goffredo et al. (2023) proposed an argument-based classification of debate content, inspiring parts of our label taxonomy. These efforts treat debates as standalone dialogues to parse or classify, in contrast to our approach of linking debates with external commentary. The CMU *Multivocal* dataset (Jo et al., 2020) integrated social media reactions by categorizing Reddit debate comments into four proposition types. That work illustrated the value of combining debates with crowd commentary, but did not include expert analysis. Our dataset bridges

that gap by including both expert journalist commentary and public Reddit comments for the same debates, enabling direct comparison of institutional versus grassroots discourse. In summary, existing debate datasets each target a slice of the problem (claims, arguments, or crowd opinions), while our dataset provides aligned expert commentary covering fact-checks, summaries, opinions, and more, over multiple election cycles.

Beyond debates, our work draws on NLP research into financial and policy communications. FOMC press conferences (the U.S. Federal Reserve's Q&A sessions after policy meetings) have been studied for their economic impact and rhetoric. For example, prior work analyzed the language of Fed statements to predict market reactions or assess sentiment (e.g., (Zirn et al., 2015; Rohlfs et al., 2016)). Corporate earnings calls are another important domain, with NLP applied to tasks like summarization of call transcripts, extraction of forward-looking statements, and stock movement prediction. Keith and Stent (2019) investigated summarizing earnings calls, and more recent studies (Mukherjee et al., 2022; Huang et al., 2024) use transformer models to analyze financial transcripts. However, these financial NLP works typically operate on monologues or Q&A content alone. Our dataset is novel in that it pairs these financial event transcripts with real-time expert commentary (e.g., from Bloomberg analysts) that interprets and reacts to the content. To our knowledge, this is the first resource to capture how financial experts comment during an unfolding event (press conference or earnings call), adding a layer of analysis akin to real-time summarization plus evaluation.

6 Conclusion

We have presented a new multi-domain dataset for Live Commentary Planning and Generation, covering real-time expert and public commentary on debates, policy press conferences, and earnings calls. This dataset is the first to align transcripts of high-stakes events with time-synchronized expert analyses, annotated with a rich taxonomy of commentary types. By framing both a planning task (deciding what commentary action to take) and a generation task (producing the commentary text), we move toward building systems that not only summarize or classify, but emulate expert commentators in both decision-making and writing. The novelty of our dataset lies in its comprehensive scope and fine granularity: it bridges previously disparate research areas (argument mining, fact-checking, summarization, and discourse analysis) in a unified benchmark. We believe this resource will be highly useful for developing and evaluating the next generation of intelligent assistants capable of providing live analysis. Furthermore, the dataset is extensible: the framework could be applied to other languages (e.g. live translation commentary) or other event types (parliamentary debates, live sports commentary with expert analysts, etc.), enabling cross-cultural and cross-domain studies of real-time commentary.

Looking ahead, we anticipate this dataset will inspire research into planning-enhanced text generation, better integration of external knowledge for live tasks, and evaluation techniques for creative generation. It also offers opportunities for interdisciplinary collaboration with journalism and communication studies, examining how AI can augment or mimic professional commentators. In the era of powerful LLMs, our work highlights that expertise and strategy in generation remain non-trivial to achieve. By providing a challenging benchmark and initial baselines, we set the stage for future innovations in real-time, context-aware text generation. We invite the community to use and build upon our dataset.

References

Pavlin Atanasova, Alberto Barrón-Cedeño, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Stoyan Kyuchukov, Giovanni Da San Martino, and Preslav Nakov. 2018. Overview of the clef-2018 checkthat! lab on automatic identification and verification of political claims. task 1: Check-worthiness. *arXiv:1808.05542*.

Pierpaolo Goffredo, Michele Espinoza, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112. Association for Computational Linguistics.

Yanlong Huang, Wenxin Tai, Fan Zhou, Qiang Gao, Ting Zhong, and Kunpeng Zhang. 2024. Extracting key insights from earnings call transcripts via information-theoretic contrastive learning. *Information Processing & Management*. To appear.

Tatsuya Ishigaki, Goran Topic, Yumi Hamazono, Hiroshi Noji, Ichiro Kobayashi, Yusuke Miyao, and Hiroya Takamura. 2021. Generating racing game commentary from vision, language, and structured data.

In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 103–113, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting check-worthy claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30, New Orleans, Louisiana. Association for Computational Linguistics.

Youngwoo Jo, Elijah Mayfield, Chris Reed, and Eduard Hovy. 2020. Machine-aided annotation for fine-grained proposition types in argumentation. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 1008–1018, Marseille, France. European Language Resources Association.

Katherine A. Keith and Amanda Stent. 2019. Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503, Florence, Italy. Association for Computational Linguistics.

Dan Liu, Yuxi Wang, Jennifer Loy, and Vera Demberg. 2024. SciNews: From scholarly complexities to public narratives – a dataset for scientific news report generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14429–14444, Torino, Italy. European Language Resources Association and International Committee on Computational Linguistics.

Edison Marrese-Taylor, Yumi Hamazono, Tatsuya Ishigaki, Goran Topić, Yusuke Miyao, Ichiro Kobayashi, and Hiroya Takamura. 2022. Open-domain video commentary generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7326–7339, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ricardo Mestre, Roko Milicic, Stuart E. Middleton, Mark Ryan, Jiechi Zhu, and Timothy J. Norman. 2021. M-Arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhijnan Patwari, Dan Goldwasser, and Saurabh Bagchi. 2017. Tathya: A multi-classifier system for detecting check-worthy statements in political debates. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management (CIKM)*, pages 2259–2262.

Christopher Rohlfs, Sunandan Chakraborty, and Lakshminarayanan Subramanian. 2016. The effects of the content of FOMC communications on US treasury rates. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2096–2102, Austin, Texas. Association for Computational Linguistics.

Jacky Visser, Rory Duthie, John Lawrence, and Chris Reed. 2018. Intertextual correspondence for integrating corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ze Yang, Can Xu, Wei Wu, and Zhoujun Li. 2019. Read, attend and comment: A deep architecture for automatic news comment generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5077–5089, Hong Kong, China. Association for Computational Linguistics.

Cäcilia Zirn, Robert Meusel, and Heiner Stuckenschmidt. 2015. Lost in discussion? tracking opinion groups in complex political discussions by the example of the FOMC meeting transcriptions. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2015)*, pages 747–753, Hissar, Bulgaria.