Segmentation of Complex Question Turns for Argument Mining: A Corpus-based Study in the Financial Domain

Giulia D'Agostino, Chris Reed, Daniele Puccinelli
Università della Svizzera Italiana, University of Dundee,
Scuola Universitaria Professionale della Svizzera Italiana
giulia.dagostino@usi.ch, c.a.reed@dundee.ac.uk, daniele.puccinelli@supsi.ch

Abstract

Within the financial communication domain, Earnings Conference Calls (ECCs) play a pivotal role in tracing (a) the presentational strategies and trust-building devices used by company representatives and (b) the relevant hot-topics for stakeholders, from which they form an (e)valuation of the company. Due to their formally regulated nature, ECCs are a favoured domain for the study of argumentation in context and the extraction of Argumentative Discourse Units (ADUs). However, the idiosyncratic structure of dialogical exchanges in Q&A sessions of ECCs, particularly at the level of question formulation, challenges existing models of argument mining, which assume adjacency of related question and answer turns in the dialogue. Maximal Interrogative Units (MIUs) are a novel approach to grouping together topically contiguous argumentative components within a question turn. MIU identification allows application of existing argument mining techniques to a less noisy unit of text, following removal of discourse regulators and splitting into sub-units of thematically related text. Evaluation of an automated method for MIU recognition is also presented with respect to gold-standard manual annotation.

Keywords: text segmentation, Q&A, argument mining, financial communication

1. Introduction

Financial communication data uniquely allows for linguistic and argumentative analysis for objectives such as the prediction of financial results, training for managerial roles, and general insight into the reasoning strategies of participants. It also constitutes an optimal playground for text generation, specifically of questions, to be tested against actual Q&A sessions. Among communicative events in the financial domain, Earnings Conference Calls (ECCs) play a pivotal role in reaching such goals. ECCs are quarterly exchanges between financial analysts and corporate managers, held from nearly all companies listed in the stock market; ECCs are peculiar both for their interactive nature in a public setting and for the active role held by analysts.

Argument analysis turns unstructured text into structured argument data, giving an understanding of the individual points being made and of the relationships between them. Argument mining is the automatic identification and extraction of argument components and structure (Lawrence and Reed, 2019). Argument mining at large scale on ECCs data extends insights and findings from quantitative analysis, making them representative and reliable.

Argument mining on ECCs data currently relies on macro-segmentation (speaker turn) and micro-segmentation (Argumentative Discourse Unit, ADU). Q&A sessions of ECCs, however, collapse multiple questions into one question turn. Such a structure challenges argument mining techniques because it doesn't provide logically sequential ADUs to the mining tools. Here we present a novel mid-level unit of segmentation to address this issue.

Mid-level text segmentation units presently introduced are the Maximal Interrogative Unit (MIU) and the Maximal Answering Unit (MAU) triggered from a MIU.

In the following, MIU and MAU will be presented (section 3), an automated task for MIU recognition will be run with GPT (section 4), and the evaluation of such a task will be conducted in comparison to manually annotated data (section 5).

1.1. ECCs as an Argumentative Activity Type

Structures and strategies used by participants in an argumentative discussion differ according to context in which they are employed (Rigotti, 2006; van Eemeren, 2009). The purpose of the Q&A sessions in ECCs is to seek insight and opinions about past performance and future expectations, not to disclose legally binding information beyond that already made officially available by the company. Typically, analysts only have one turn for their questions, thus leading to an idiosyncratic question-compression strategy. As a consequence, a question turn by an analyst is a collection of sentences, arranged around a number of topics. A question turn triggers one or more answering turns in response. Such an arrangement is different from the assumed structure of a Q&A exchange and is potentially problematic for analysis.

Dialectical exchanges in quarterly ECCs are a resource-rich domain for the study of argumenta-

LREC-COLING 2024, pages 14524–14530

tion in context. Financial analysts become protagonists in soliciting soft information from corporate representatives. Analysis of such data not only provides rich insight for academic work, but can also be used for financial results prediction (Chen et al., 2018) by correlating the type of analyst-manager interaction with stock price movements during the call. In addition, the data can be used for the training of managers, with conclusions drawn from such studies helping managers to learn to convey their speech in the most favourable way for the company.

2. Related Work and Motivation

2.1. Unit Segmentation

ADUs are the minimal units into which a text is segmented for argument analysis. They are the argumentative equivalent of Elementary Discourse Units (EDUs) in linguistics tasks. Argument mining is a composite and complex task (Lawrence and Reed, 2019; Ajjour et al., 2017). The steps in argument mining are: (1) the identification, segmentation, and classification of argumentative discourse units (ADUs), (2) the identification and classification of the relations between ADUs (Peldszus and Stede, 2013), and (3) the identification of argument schemes, namely the implicit and explicit inferential relations within and across ADUs (Macagno and Walton, 2014).

Questioners in ECCs have only one (or exceptionally two) turns in which to pose multiple questions. As questioners condense their questions into a single turn, there is a regular mismatch between the logical and linear structure of the discourse. The logical structure would expect adjacency between simple question-and-answer pairs; the linear structure displayed in such sessions condenses multiple logical question turns into one. Argumentative analysis must handle such interactions and, specifically, slice thematically related sub-units within a question turn.

Approaches to mid-level segmentation such as in Liu et al. (2022) are compatible with the current claims: they “utilise zoning information in the tasks of argumentative component identification and classification” because “relying on zoning information, a model can mine argumentative components more accurately”. Their perspective is consistent with previous literature on argument zoning (Teufel, 1999; Teufel and Moens, 1999; Teufel et al., 1999), which can function as a preparatory step in argument mining via purposeful pre-segmentation of the text. However, the current contribution differs from argument zoning in two respects: (a) it does not categorize segments, but only focuses on boundary identification and (b) it specifically exploits the idiosyncratic information structure offered by ECCs.

2.2. Computer-aided Analyses in the Financial Domain

Chen et al. (2021) show that opinion mining in the financial domain, especially applied to stakeholders’ opinions, helps to determine the link between financial events and market reaction. In contrast, Pazienza et al. (2020) analyze ECCs via an abstract argumentation approach to predict analysts’ recommendations. However, none of these studies addresses the segmentation of question units.

Whilst working with ECCs for argument component identification, Alhamzeh et al. (2022) do not apply directly their methodology to extraction, although this is their eventual aim. Such an additional step is later reached by Alhamzeh (2023), although the author does not deal with text segmentation other than turn and argument component identification.

2.3. Natural Language Generation in the Financial Domain

Natural Language Generation (NLG) is an increasingly popular task in the AI community. ECCs have recently gained attention from NLG as well - particularly for the generation of appropriate questions for a given Q&A session based on the topics of the preceding presentation (Juan et al., 2023). Implementation, however, generates single interrogative sentences only and therefore is not consistent with the structure of real question turns.

Such studies are still very preliminary. The integration of such a notion in the framework, however, would be beneficial for the improvement of results in further research, making them more congruent with naturally occurring turns.

3. Maximal Interrogative Units (MIUs)

Mid-level segmentation of question turns would lower the processing cost of each of the steps of argument mining, by reducing the noise deriving from contiguous unrelated sections. A Maximal Interrogative Unit (MIU) is a series of one or more discursive moves that may maximally cover a question turn and minimally cover a single interrogative sentence. A MIU is characterized by the following attributes:

- It is a macro-unit which groups discursive moves within the same question turn and comprising no less than one question, each discursive move being the length of (at least) one sentence;
- All the discursive moves in a MIU prepare, rephrase or modulate the same objective;

- All discursive moves within a MIU can be satisfied by a single corresponding Maximal Answering Unit.

Example 1 (DASH Q1 2021, analyst Youssef Squali) illustrates an instance of segmentation of a turn into two MIUs:

(1) a. **First,** [can you just speak to the recent trends that you've seen so far in May?]$_{question}$ [I think your guide speaks to it,]$_{preface}$ but [anything to highlight in terms of just the competitive intensity and how you guys feel about the – particularly the growth in the nonfood delivery business in the quarter and contribution to it?]$_{question}$

b. **And second,** [as you look at the diversification that you're embarking on into nonfood, convenience, grocery, etc.,]$_{preface}$ [I was wondering if you can just speak to the broader – well, first, how big do you think that business could become over time?]$_{question}$ [Is this a situation where you could see a scenario where half of your business is coming from these new initiatives, say, over the next, I don't know, three to five years?]$_{question}$ But probably also, just [how do you see that impacting the take rate over time?]$_{question}$

The first MIU presents two questions, accompanied by a preface hinting to a need for the answer to go beyond what was already disclosed. The second MIU presents three questions, accompanied by a preface shifting the topic from the first MIU.

A Maximal Answering Unit (MAU) is triggered by a MIU: it is a series of sentences within one answering turn, maximally covering the entire turn, in reaction to a MIU.

The dataset is fully annotated according to both MIUs and MAUs. This paper, however, will focus on MIUs.

3.1. Added Value of MIU Segmentation

An inference is a support relation between a premise and a conclusion. Inference chains in this context are either intra-unit, i.e., within either a MIU or a MAU, or cross-unit, i.e., between a MAU and its triggering MIU. Both MIU-MAU pairs and individual units thus have an argumentative role.

Within a MIU, the argumentative structure is constituted by discursive moves "preface" and "question". A preface is an assertive statement that can either precede, follow or be contained in a question, providing arguments supporting the relevance of the speech act of the question (Lucchini et al., 2022). Example 2 (ABNB Q1 2021, analyst Justin Post) is an instance of intra-unit argumentation in a MIU:

(2) ([I think in the letter, it said post listings were stable with Q4,]$_{preface}$ but [it seems like you're really encouraged by what you're seeing.]$_{preface}$)$^{\textbf{premises}}$ $\rightarrow$ ([**So** maybe you could dive in there and tell us, you know, what is encouraging about what you're seeing with hosts]$_{question}$ and [whether you see – expect a lot of new listings to hit the market over the next year?]$_{question}$)$^{\textbf{conclusions}}$

Here the two prefaces constitute the argumentative premises in support of the implicit conclusion that the questions are relevant and deserving of an answer. The two questions are thus the explicit counterpart of the conclusion of the inference.

4. Data

ECCs constitute a remarkably extensive dataset, currently quantifiable in the billions of words freely accessible from past transcriptions and with new sets being added quarterly. They represent a unique corpus of publicly available data, rich in argumentative exchanges. Automated analysis of such texts on a large scale, overcoming the limitations of manual annotation, would result in a singularly rich collection of reasoning instances.

MIUs are not specific to this domain and can be retrieved in other comparable Q&A interaction schemes as well; narrowing the scope to ECCs, however, has the twofold benefit of exceptional quantity, in a particularly structured and identifiable appearance.

4.1. Dataset

The dataset comprises 24 ECC Q&A sessions of companies ABNB, CS, DASH, HAS, RDS, and Z from fiscal year 2021,[1] each manually and independently annotated by at least two trained annotators according to the coding standard developed by the research team. The guidelines for annotation are publicly available (Lucchini and D'Agostino, 2023). The annotations were later curated by at least one member of the research team. Such labelling of texts both provides the gold standard against which to test the automatic segmentation and supplies the exemplary cases with which to instruct the automatic segmentation tool. The resulting, manually identified MIUs are 522; the dataset is further described in Table 1.

[1] The transcripts of ECCs are publicly available and can be retrieved from specialized websites such as The Motley Fool and Seeking Alpha

	total
documents	24
words	207,013
question turns	341
MIUs	522

Table 1: Dataset description

The selected software for manual annotation process is the INCEpTION platform (Klie et al., 2018). MIUs are identified with linear labelling of text spans.

The inter-annotator reliability applied to the unitizing of textual continua (Krippendorff, 1995; Krippendorff et al., 2016) is calculated in the form of alpha values for each annotated document, which proved to be the most appropriate measure for the current task according to Artstein and Poesio (2008). The median U-alpha coefficient value is $_U\alpha$ = .933 across all manually annotated texts.

4.2. Automatic Segmentation via Prompting

The automatic segmentation task presented here is performed by GPT. At the time of the study, GPT-3.5 Turbo model was not available for fine-tuning; therefore, it was accessed via OpenAI API. The API was called for each document in the dataset and fed with a prompt that comprised: (a) the prompting text, describing the task; (b) an example of a turn and its segmentation into MIUs; (c) all question turns of the document, arranged sequentially in the same string of text. The model was thus prompted with a one-shot learning task at each iteration. The prompt was selected as the one obtaining best outputs after three cycles of tuning.

This is the formulation of the final prompting text:

> "Given an input document, we need to break it down into spans. Each span should represent a coherent semantic unit (typically above the sentence level), and the spans should be stored in a list as strings, as in the example below. The example breaks down one turn into spans; note that a document will contain many turns, and that each turn will contain at least one span."

5. Results

Evaluation of GPT performance in the segmentation task against the baseline – the latter represented by manual annotation – is calculated as a Krippendorff's alpha value.[2] The evaluation was performed twice:

- The first round considered boundaries of text spans only. The median value of the measure is $_U\alpha$ = .377 across documents, unitizing the textual continuum for all identified segments.

- The second round also included IOB-tags (at the sentence level) for each identified MIU. The median value results in $_{cu}\alpha$ = .170.

	alpha value
inter-annotator reliability	$_U\alpha$ = .933
GPT vs baseline (boundaries)	$_U\alpha$ = .377
GPT vs baseline (IOB)	$_{cu}\alpha$ = .170

Table 2: Results summary

6. Discussion

Evaluation of inter-annotator agreement on manually annotated data shows that the MIU is an unmistakable unit for human assessment. Such a claim is further reinforced by the heterogeneity in background and expertise of the annotators. Results also show that automatic segmentation performed by GPT - modulo the limitations of the current study - is not appropriate. Krippendorff's alpha values for both evaluation approaches demonstrate that the output is unreliable. A lower agreement rate is moreover exhibited by IOB-tagged segmentation due to inconsistency in tagging values.

7. Conclusion and next steps

The ECC activity type displays an idiosyncratic question turn structure; their analysis would benefit from a tailored intermediate segmentation. Such segmentation has the role to improve the performance of argument mining efforts in such a domain. The original notion of Maximal Interrogative Unit is introduced and presented as relevant for both argumentation theory and argument mining. Excellent degrees of reliability of manual annotation of MIUs demonstrates that such an unit is evidently identifiable.

The automation of the segmentation task with one-shot instructed GPT currently under-performs; this is supported by Krippendorff's alpha evaluation of GPT segmentation. Next steps include:

- Fine-tuning of ML models for MIU identification; particularly BiLSTMs such as BERT (Devlin et al., 2019)).

- Extraction of MAUs as textual units necessarily related and linked to a MIU.

- Mining of argumentative components within and between units.

[2] The code for text processing is available on GitHub.

Concerning future applications of MIU-MAU text segmentation, it will benefit research on argument mining, since it narrows the scope for the identification of related argumentative components. Generation tasks will equally profit from it because MIU-MAU segmentation provides a genuine description of the shape of turns, allowing for their correct replication.

8. Limitations

The tool employed for automatic text segmentation lacks transparency (design limitation) and displays high latency (infrastructural limitation); both are unavoidable from the user's end. Better performance can be expected to be achieved with fine-tuning of the model, rather than one-shot prompting; such an option was not available at the time of development of the current study.

Regarding research design, we fed the API entire question turns, comprising irrelevant discourse regulators. The results therefore contain unpolished data to this respect. False positives, however, were left in the boundaries-driven evaluation process because their omission invariably led to a worsening of the alpha value. On the other hand, IOB-driven evaluation slightly improved with the omission of those discourse regulators that were entirely recognized as an independent unit. In both cases, the evaluation was purposefully tweaked to return the highest value.

Overall, results show that the instrument chosen for the empirical testing limited the potential of the theoretical concept, the validity of which remains however intact.

9. Acknowledgements

The work in this paper was supported by the Swiss National Science Foundation under the project “Mining argumentative patterns in context. A large scale corpus study of Earnings Conference Calls of listed companies" (grant n. 200857).

10. Bibliographical References

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit Segmentation of Argumentative Texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128, Copenhagen, Denmark. Association for Computational Linguistics.

Alaa Alhamzeh. 2023. *Language Reasoning by Means of Argument Mining and Argument Quality*. Ph.D. thesis, Universität Passau.

Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022. It's Time to Reason: Annotating Argumentation Structures in Financial Earnings Calls: The FinArg Dataset. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 163–169, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. *From Opinion Mining to Financial Argument Mining*. SpringerBriefs in Computer Science. Springer Singapore, Singapore.

Jason V. Chen, Venky Nagar, and Jordan Schoenfeld. 2018. Manager-analyst conversations in earnings conference calls. *Review of Accounting Studies*, 23(4):1315–1354.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, pages 4171–4186. ArXiv: 1810.04805 version: 2.

Yining Juan, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Generating Multiple Questions from Presentation Transcripts: A Pilot Study on Earnings Conference Calls. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 449–454, Prague, Czechia. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Klaus Krippendorff. 1995. On the Reliability of Unitizing Continuous Data. *Sociological Methodology*, 25:47.

Klaus Krippendorff, Yann Mathet, Stéphane Bouvry, and Antoine Widlöcher. 2016. On the reliability of unitizing textual continua: Further developments. *Quality & Quantity*, 50(6):2347–2364.

John Lawrence and Chris Reed. 2019. Argument Mining: A Survey. *Computational Linguistics*, 45(4):765–818.

Boyang Liu, Viktor Schlegel, Riza Batista-Navarro, and Sophia Ananiadou. 2022. Incorporating Zoning Information into Argument Mining from Biomedical Literature. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6162–6169, Marseille, France. European Language Resources Association.

Costanza Lucchini and Giulia D'Agostino. 2023. Good answers, better questions. Building an annotation scheme for financial dialogues. Technical report. Ark:/12658/srd1326777.

Costanza Lucchini, Andrea Rocci, and Giulia D'Agostino. 2022. Annotating argumentation within questions. Prefaced questions as genre specific argumentative pattern in earnings conference calls. In *Proceedings of the 22nd Edition of the Workshop on Computational Models of Natural Argument (CMNA 22)*, volume vol. 3205, pages 61–66, Cardiff. CEUR.

Fabrizio Macagno and Douglas Walton. 2014. Argumentation schemes and topical relations. In Giovanni Gobber and Andrea Rocci, editors, *Language, reason and education*, pages 185–216. Peter Lang, Bern.

Andrea Pazienza, Davide Grossi, Floriana Grasso, Rudi Palmieri, Michele Zito, and Stefano Ferilli. 2020. An abstract argumentation approach for the prediction of analysts' recommendations following earnings conference calls. *Intelligenza Artificiale*, 13(2):173–188.

Andreas Peldszus and Manfred Stede. 2013. From Argument Diagrams to Argumentation Mining in Texts. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31.

Eddo Rigotti. 2006. Relevance of Context-bound loci to Topical Potential in the Argumentation Stage. *Argumentation*, 20(4):519–540.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Articles*. Ph.D. thesis, University of Edinburgh.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway. Association for Computational Linguistics.

Simone Teufel and Marc Moens. 1999. Discourse-level argumentation in scientific articles: human and automatic annotation. In *Towards Standards and Tools for Discourse Tagging*.

Frans H van Eemeren. 2009. *Examining Argumentation in Context: Fifteen Studies on Strategic Maneuvering*. John Benjamins Publishing Co, Amsterdam/Philadelphia.

A. Further examples of annotation and segmentation

1. [Thanks a lot. It's Kevin from Cowen.]$_{\textit{discourse regulator 1}}$
   { [Can you give us a sense of the booking trends in the fourth quarter, quarter to date?]$_{\textit{question 1}}$ [You mentioned acceleration in the shareholder letter,]$_{\textit{preface 1}}$ [are you seeing that back to Q2 levels yet in terms of growth as compared to the same quarter in 2019,]$_{\textit{question 2}}$ [just given kind of the Delta slowdown?]$_{\textit{preface 2}}$ }$_{\textit{MIU}}$
   [That would be helpful.]$_{\textit{discourse regulator 2}}$

2. [So yes, I'll ask two questions as well.]$_{\textit{discourse regulator 1}}$
   { [The first one,]$_{\textit{discourse regulator 2}}$ [just trying to get a sense, I appreciate you don't prejudge the outcome, but – so the strategic review,]$_{\textit{discourse regulator 3}}$ [just if we can get a bit more color in terms how the process works, how that's being conducted, how decisions will be made and the kind of trade-offs and the processes involved in that,]$_{\textit{question 1}}$ }$_{\textit{MIU 1}}$
   [that would be helpful.]$_{\textit{discourse regulator 4}}$
   { [And secondly,]$_{\textit{discourse regulator 5}}$ [just coming back to the Slide 10 in terms of the, I guess, employee hiring and attrition.] $_{\textit{discourse regulator 6}}$ [I'm just curious, if you were to cut back instead of total employees but rather just looking at, for example, MDs or material risk takers, does it give the same picture? Or is it then slightly different?]$_{\textit{question 2}}$ }$_{\textit{MIU 2}}$

3. [Morning.]$_{\textit{discourse regulator 1}}$ [Thank you very much.]$_{\textit{discourse regulator 2}}$ [Apologies for taking on the painful bits, but I still think there's more clarification that we need.]$_{\textit{discourse regulator 3}}$ [I wanted to just ask two things.]$_{\textit{discourse regulator 4}}$
   { [One is on Greensill.]$_{\textit{discourse regulator 5}}$ [You've

got about CHF5 billion cash, but also about CHF5 billion remaining exposure in those funds.]$_{\textit{preface 1}}$ [And I just wondered if you could put a number on how much of that CHF5 billion remaining exposure is to doubtful borrowers, including, obviously, Gupta, but also some of the other doubtful borrowers who seem reluctant to pay.]$_{\textit{question 1}}$ [So, that's my first question.]$_{\textit{discourse regulator 6}}$}$_{\textit{MIU 1}}$

{ [And my second question is on the other painful, like I said, I'm afraid, on the Archegos situation.]$_{\textit{discourse regulator 7}}$ [Could you walk us through the mechanics of how that loss came about in terms of what the outstanding gross exposure was at the moment of problem?]$_{\textit{question 2}}$ [How much margin you had and the sequence of events in terms of, were you slow to sell down or how do you assess what happened?]$_{\textit{question 3}}$}$_{\textit{MIU 2}}$

[Those are my two questions please.]$_{\textit{discourse regulator 8}}$