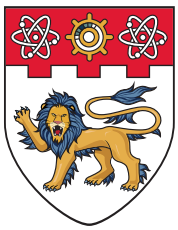
Coreference Resolution in Natural Language Processing



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Liu Ruicheng

College of Computing and Data Science

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy of Computer Science

2025

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

Liu Ruicheng

06/07/2024

. . . . . .
Date

. . . . . .
Liu Ruicheng

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

06/07/2024

. . . . . .
Date

. . . . . .
Erik Cambria

Authorship Attribution Statement

Please select one of the following; *delete as appropriate:

~~*(A) This thesis **does not** contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.~~

*(B) This thesis contains material from 2 paper(s) published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Chapter 2 is published as Liu, R., Mao, R., Luu, A.T. et al. A brief survey on recent advances in coreference resolution. Artif Intell Rev 56, 14439–14481 (2023). https://doi.org/10.1007/s10462-023-10506-319

The contributions of the co-authors are as follows:

- Prof Erik Cambria provided the initial project direction and edited the manuscript drafts.
- I prepared the manuscript drafts. The manuscript was revised by Mao Rui and Asst/Prof Luu An Tuan.
- I designed the overall structure of the survey and collected all the data regarding databases, evaluation metrics and methods. I also analyzed the data.
- All the relevant papers are summarized and organized in a clean structure by me. The way to present the development trends of Coreference Resolution is designed by me.

Chapter 3 is published as Liu, R., Chen, G., Mao, R., \& Cambria, E. (2023). A Multi-task Learning Model for Gold-two-mention Co-reference Resolution. 2023 International Joint Conference on Neural Networks (IJCNN), 1-8.

The contributions of the co-authors are as follows:

- Prof Erik Cambria provided the general research direction and performed final review of the manuscript drafts.
- The initial drafts of the manuscript were composed by myself, with subsequent revisions being collaboratively undertaken alongside Dr. Mao Rui and Dr. Chen Guanyi.
- I was responsible for the collection of all raw data, which involved transforming the data into a standardized format compatible with our model.
- The execution of experiments across the three collected datasets, as well as the assessment of their outcomes, were conducted under my direction.
- I was instrumental in the development of the Dynamic Weight Balancing feature within our model and contributed to enhancements in the multi-task learning framework. Furthermore, I spearheaded the ablation study segment of our research, including the analysis of its results.
- The task of error analysis was jointly carried out by Dr. Chen Guanyi and myself, ensuring a thorough examination of discrepancies and inaccuracies within our findings.

06/07/2024

Liu Ruicheng

. . . . . .
Date

. . . . . .
Liu Ruicheng

Acknowledgments

It is with deep respect and appreciation that I acknowledge the opportunity to undertake doctoral studies. Foremost, my heartfelt gratitude is extended to my primary advisor, Professor Erik Cambria, for his unparalleled mentorship throughout my doctoral studies. Our intellectual engagements have been indispensable in shaping my research focus, particularly in the realms of Natural Language Processing and Coreference Resolution. As an astute and objective critic of my work, Professor Cambria has been pivotal in deepening my comprehension of the research landscape through his invaluable insights and constant encouragement.

Equally deserving of acknowledgement are my co-authors, whose collaborative efforts have further enriched my understanding and added depth to my research. Their perspectives and expertise have provided me with a broader outlook on my research area, augmenting the quality of my work.

I also wish to express my indebtedness to Continental Singapore for their generous financial support, especially during the initial biennium of this research. Further, my sincere thanks go to Nanyang Technological University and the School of Computer Science and Engineering for providing an academically stimulating environment that has been integral to my research.

Concluding, I would like to extend my deepest appreciation to my parents for their unwavering support throughout my doctoral journey. The emotional sustenance they have provided, in the form of love, care, and motivation, has been a linchpin in maintaining my mental well-being during this intellectually demanding period.

Contents

List of Figures

List of Tables

List of Publications

1. Liu, Ruicheng, Rui Mao, Anh Tuan Luu, and Erik Cambria. "A brief survey on recent advances in coreference resolution." Artificial Intelligence Review (2023): 1-43.

2. Liu, Ruicheng, Guanyi Chen, Rui Mao, and Erik Cambria. "A Multi-task Learning Model for Gold-two-mention Co-reference Resolution." International Joint Conference on Neural Networks (IJCNN) 2023

Abstract

Coreference Resolution (CR), a key task in Natural Language Processing, involves identifying expressions that refer to the same entity or event. This thesis presents a detailed survey of CR models from the past decade, introduces a novel Multi-task Learning (MTL) model for the Gold-two-mention CR task, and explores CR's application in financial sentiment analysis. The models surveyed are classified into feature-based, neural networks (including multilayer perceptrons and recurrent neural networks), knowledge-based, and transformer-based frameworks. The new MTL model simultaneously addresses mention identification and linking in gold-two-mention style CR, using a dynamic weight balancing mechanism to optimize task-specific weights during training. This model achieves state-of-the-art results on three benchmark datasets. Furthermore, the thesis integrates advanced CR techniques with domain-specific models such as FinBERT to improve sentiment analysis in financial texts. Rigorous empirical methods, including Propensity Score Matching and multivariate analysis, assess the impact of sentiments on financial outcomes, validating the framework's predictive capabilities in financial sentiment analysis. This research not only improves CR methodologies but also demonstrates their practical relevance in financial analytics, encouraging further interdisciplinary studies.

Chapter 1

Introduction

1.1 Background

In the realm of linguistics and computational linguistics, discourse is understood as a collection of statements that, when combined, exhibit a logical structure and maintain a consistent meaning. Discourse encompasses various forms, such as monologues, which feature a single or implicit speaker, and dialogues, involving interactions between two or more speakers.

The attainment of coherence within a discourse necessitates an astute understanding of its argumentation structure and the flow of information. Within this framework, coreference resolution plays a pivotal role as a sophisticated parsing endeavor, with anaphora resolution being a critical subset of this process. Anaphora resolution involves identifying the antecedents of referring expressions, thereby clarifying the referential relationships within the text. Coreference resolution (CR), in a broader sense, pertains to the task of identifying all text spans within a context that refer to the same physical entity or event. This process can be either anaphoric, linking backwards to previous discourse elements, or cataphoric, linking forwards to subsequent elements in the discourse [Mitkov, 1999].

The significance of coreference resolution extends far beyond its immediate linguistic applications. It is a crucial component in various downstream natural language processing (NLP) tasks. These tasks include entity linking [Kundu et al., 2018], which involves associating distinct entities mentioned in the text with unique identifiers; named entity recognition [Dai

et al., 2019], a process of classifying text segments into predefined categories like names of persons, organizations, or locations; question answering systems [Bhattacharjee et al., 2020], where understanding referential relations can drastically improve the accuracy and relevance of responses; and sentiment analysis [Krishna et al., 2017, Mao and Li, 2021], where resolving coreferences can lead to a more accurate understanding of the sentiment conveyed in the text. Additionally, coreference resolution finds its utility in enhancing the performance of chatbots [Zhu et al., 2018], making their interactions more coherent and contextually relevant. Moreover, it has profound implications in the field of referring expression generation [Li et al., 2018, Chen et al., 2018], facilitating the creation of more precise and contextually appropriate referential language.

1.2 A Survey on Development for Coreference Resolution In the Past Decade

Previous surveys, such as Mitkov [1999], laid the groundwork in anaphora resolution and algorithm analysis. Ng [2010] explored the initial 15 years of Machine Learning's role in coreference resolution, while Lu and Ng [2018] offered a comprehensive review of event coreference resolution from 1997 to 2017, spanning supervised to unsupervised approaches, but not extensively covering entity coreference or neural network methodologies. Sukthanker et al. [2020] provided insights into more recent advancements, including deep learning techniques. However, these reviews do not encompass the latest developments, particularly post the advent of Transformers [Vaswani et al., 2017] and their significant impact on NLP.

Our survey addresses this gap, tracing the evolution of coreference resolution from feature-based and classical machine learning methods to contemporary deep learning approaches, including neural-contextual, neural-knowledge, and transformer-based models. We offer detailed insights into each technical trend, associated datasets, and evaluation metrics. Our survey surpasses previous surveys in depth, presenting comprehensive sum-

maries of coreference annotation tools, application-oriented datasets (Table 2.1), methodologies (Tables 2.2 and 2.3), and feature analyses (Table 2.4).

1.3 Multi-task Learning in Coreference Resolution

In the domain of coreference resolution (CR), a prominent query category is the "Gold-Two-Mention" style. This approach involves a scenario where, within a specified context, a pronoun and two potential referent mentions are presented. The fundamental challenge lies in accurately associating the pronoun with the appropriate referent mention. This chapter presents the development of a multitask learning model specifically tailored for datasets adhering to the Gold-Two-Mention style. Our model innovatively integrates the optimization of both mention identification and coreference linkage tasks within a dual-tower architecture, underpinned by a RoBERTa-based contextualization framework.

Our exploration encompasses two distinct methodologies for weight balancing: fixed weightage and dynamic weight balancing. The experimental results demonstrate that our multitask learning model, employing dynamic weight balancing, attains state-of-the-art performance on two Gold-Two-Mention style datasets. Additionally, it yields competitive outcomes in another dataset. Notably, these achievements are realized despite a significantly more cost-effective fine-tuning process. This underscores the efficiency and effectiveness of our proposed model in the context of Gold-Two-Mention style coreference resolution.

1.4 Coreference Resolution's Application in Finance

We presented a comprehensive investigation into the application of Coreference Resolution techniques within the financial domain. Our research focused on integrating state-of-the-art coreference resolution algorithms as

a crucial pre-processing step in financial sentiment analysis. By employing advanced language models, such as FinBERT, and machine learning techniques, our objective was to enhance the accuracy and granularity of sentiment attribution in financial texts. Our methodology incorporated a refined entity recognition process, sentiment extraction, and a robust empirical framework that combined Propensity Score Matching (PSM) and Difference-in-Differences (DID) analysis, allowing us to isolate the causal impact of sentiments on financial outcomes while accounting for potential biases and confounding factors.

Our study introduced novel analytical techniques, such as mediation effect regression, to uncover the underlying mechanisms through which sentiment influenced financial performance. We also conducted rigorous robustness checks, including postfinancial sentiment analysis DID regression and multivariate regression, to validate the stability and reliability of our findings. Furthermore, we performed a quantitative analysis of financial texts post-coreference resolution, utilizing advanced models to investigate the interaction of textual factors and their impact on sentiment identifiability. Finally, we evaluated the performance of financial sentiment prediction models, particularly the Random Forest model, across original and algorithmically-processed texts, highlighting the enhanced predictive capabilities of our proposed methodological framework and the critical role of coreference resolution in improving the accuracy and depth of financial sentiment analysis.

1.5 Major Contributions

The major contributions of this doctoral thesis are as follows:

1. A comprehensive survey of the development trends in Coreference Resolution over the past decade was conducted, categorizing models into feature-based, neural network-based, knowledge-based, and transformer-based approaches. This survey provides a valuable resource for researchers and practitioners in the field, offering insights

into the evolution of Coreference Resolution techniques and the current state-of-the-art methods.

2. The introduction of a novel Multi-task Learning model for Gold-two-mention Coreference Resolution that outperforms existing methods on multiple datasets by jointly learning mention identification and linking tasks. This innovative approach addresses the limitations of traditional single-task learning methods and demonstrates the benefits of leveraging the interdependencies between mention identification and linking tasks to improve overall performance.

3. The implementation of a dynamic weight balancing mechanism in the co-reference resolver, allowing for adaptive balancing between mention identification and linking tasks during training. This mechanism ensures that the model effectively learns from both tasks, optimizing its performance and generalizability across different datasets and domains.

4. An in-depth exploration of the application of Coreference Resolution techniques in the financial domain, particularly in the context of financial sentiment analysis. This study showcases the effectiveness of integrating advanced coreference resolution algorithms with state-of-the-art language models and machine learning techniques to enhance the precision and granularity of sentiment attribution in financial texts. The findings contribute to a deeper understanding of the role of natural language processing in financial analysis and offer novel perspectives on using linguistic techniques to make more accurate and comprehensive market assessments.

These contributions advance the field of correlation resolution and demonstrate its practical applications in the financial domain, paving the way for future research and development in these areas.

1.6 Thesis Organization

The remainder of this doctoral thesis is structured as follows:

Chapter 2 presents a comprehensive survey of the development trends in Coreference Resolution over the past decade. This chapter provides a thorough analysis of the evolution of Coreference Resolution techniques, categorizing models into feature-based, neural network-based, knowledge-based, and transformer-based approaches. The survey offers valuable information on current state-of-the-art methods and identifies potential avenues for future research.

Chapter 3 introduces a novel Multi-task Learning model for Gold-two-mention Coreference Resolution, incorporating a dynamic weight balancing mechanism. This chapter details the architecture of the proposed model, its training methodology, and the evaluation of its performance on multiple datasets. The results demonstrate the superiority of the multitask learning approach over traditional single-task learning methods and highlight the benefits of the dynamic weight balancing mechanism in optimizing model performance.

Chapter 4 explores the application of correlation resolution techniques in the financial domain, focusing on their integration with financial sentiment analysis. This chapter presents a comprehensive methodology that combines state-of-the-art coreference resolution algorithms, advanced language models, and robust empirical frameworks to enhance the accuracy and depth of sentiment attribution in financial texts. The findings underscore the critical role of Coreference Resolution in improving the performance of financial sentiment prediction models and offer novel insights into the complex interplay between linguistic techniques and financial market dynamics.

Chapter 5 concludes the thesis by summarizing the key findings and contributions of the research. This chapter also discusses the implications of the results for the fields of Coreference Resolution and financial sentiment analysis, and outlines potential future research directions, including the extension of the proposed Multi-task Learning model to other domains and the

further exploration of the synergies between natural language processing and financial analysis.

This structure is designed to reflect a comprehensive and deliberate intellectual progression through the field of Coreference Resolution, demonstrating a journey from broad knowledge acquisition to focused innovation and finally to pragmatic, real-world application. Chapter 2 establishes foundational mastery through a comprehensive survey of the field. Chapter 3 showcases focused innovation by developing a novel Multi-Task Learning model to advance the methodology for a specific, challenging sub-task. Finally, Chapter 4 demonstrates mature research judgment by applying the most appropriate state-of-the-art tool—a sequence-to-sequence model—to a complex, open-ended problem in the financial domain. This methodological diversity is therefore a deliberate feature, not an inconsistency, highlighting the ability to select the optimal tool for each distinct research problem.

Chapter 2

Coreference Resolution: A Survey

2.1 Introduction

The objective of this scholarly review is to furnish an exhaustive analysis of the latest developments in tackling issues related to coreference resolution. The models scrutinized in this survey are systematically classified into four primary frameworks: Feature-Based Approaches, Neural Networks utilizing Multilayer Perceptrons or Recurrent Neural Networks, Knowledge-Based Approaches, and Transformer-Based Approaches. Feature-Based Approaches are predicated on the exploitation of lexical, grammatical, and semantic facets. Neural Networks employing Multilayer Perceptrons or Recurrent Neural Networks serve as end-to-end systems that capitalize on contextual information in mentions, yet they abstain from the explicit incorporation of external knowledge. On the other hand, Knowledge-Based Neural Approaches are explicitly designed to assimilate external knowledge and are typically built upon existing neural network frameworks. Transformer-Based Approaches have recently elicited considerable interest in the field of Natural Language Processing [Devlin et al., 2019, Liu et al., 2019b, Yang et al., 2019, Lan et al., 2019]. Notably, architectures constructed upon BERT and SpanBERT have yielded unparalleled results in coreference resolution tasks. These pre-trained language models can be construed as intricate neu-

[0]The work in this chapter has been published in Liu, R., Mao, R., Luu, A.T. et al. A brief survey on recent advances in coreference resolution. Artif Intell Rev 56, 14439–14481 (2023). https://doi.org/10.1007/s10462-023-10506-3

ral network architectures that implicitly embed commonsense reasoning and contextual variables through elaborate embeddings.

2.2 Typologies and Approaches in Coreference Resolution

2.2.1 Resolution of Entity Coreferences

In the realm of natural language processing, entities are often represented by textual spans, or word sequences, that symbolize entities existing in the real world. The task of Entity Coreference Resolution (ECR) involves clustering these textual spans into groups where each group's elements signify the same real-world entity. The term "antecedent" refers to the specific real-world entity that is linked to anaphoric expressions within the text. When multiple textual spans point to the same antecedent, they are said to be coreferential. Conversely, a "singleton" is a term used to describe a textual span that appears only once within the context of a document.

Consider the following instance to elucidate entity coreference resolution:

> *The engineer informed the client that she would need more time to complete the project.*

In the given sentence, five textual spans can be identified as potential representations of real-world entities: *The engineer*, *the client*, *she*, *time*, and *the project*. Among these, *The engineer* and *she* reference the same entity, thus they are coreferential. The pronoun *she* is anaphoric as it refers back to *The engineer*, the antecedent. The remaining three spans are identified as singletons.

2.2.2 Resolution of Event Coreferences

Event mentions, in contrast to entity mentions, are composed of various textual segments, which include an event trigger (such as a verb, gerund, or noun) and its associated arguments. The process of Event Coreference Resolution is aimed at grouping event mentions that pertain to the same

real-world occurrence. These event mentions can either be confined to a single document (intra-document) or span across multiple documents (inter-document). This process is pivotal for tasks like information consolidation and augments several downstream natural language processing tasks, including contradiction detection [de Marneffe et al., 2008], text summarization [Ferracane et al., 2016], and reading comprehension [Khashabi et al., 2018, Welbl et al., 2018].

To illustrate, consider this example:

> *Yesterday the Delhi Police {slapped}$_{ev1}$ a protester while she was {demonstrating}$_{ev2}$ outside a hospital. Simultaneously, a woman in her 60s experienced {beating}$_{ev3}$ by police officers during another {protest}$_{ev4}$ in Uttar Pradesh, northern India. Recently, the Delhi Police suspended the officer who {assaulted}$_{ev5}$ the female protester.*

In this excerpt, five event mentions ($ev1 - ev5$) are identified: "*slapped*", "*demonstrating*", "*beating*", "*protest*", and "*assaulted*". Each event mention, or trigger word, may be associated with arguments such as subjects and objects. For instance, in the event "the Delhi Police slapped a protester", "the Delhi Police" and "a protester" are arguments of the trigger "*slapped*". While $ev1, ev3$, and $ev5$ belong to the ATTACK subtype, only $ev1$ and $ev5$ share coreference, as $ev3$ is linked to a separate protest event. Furthermore, $ev2$ and $ev4$ are not coreferential as they relate to distinct PROTEST occurrences.

2.2.3 Processes in Coreference Resolution

The tasks of entity coreference resolution and event coreference resolution are executed via a structured methodology that encompasses both the detection of mentions (pertaining to either entities or events) and the subsequent task of mention linkage.

The detection of mentions is a fundamental aspect in the context of entity coreference resolution. Studies have shown that the effectiveness of mention detection can significantly influence the overall performance of the

coreference resolution system [Poesio et al., 2016]. Typically, the assessment of mention detection is conducted as an independent task. This approach allows for a more accurate comparison across various models [Lu and Ng, 2020].

The task of mention linking involves clustering the identified mentions in alignment with the standard references. Different models approach coreference resolution with varying strategies. Some integrate the tasks of mention detection and mention linking, executing them integratedly [Lee et al., 2018, Joshi et al., 2020]. Others opt to separate these tasks, focusing exclusively on the linking of pre-identified mentions [Khosla and Rose, 2020, Caciularu et al., 2021b].

2.3 Metrics

In many commonly used datasets, such as GAP [Webster et al., 2018], DPR [Rahman and Ng, 2012], WSC [Levesque et al., 2012], Winogender [Rudinger et al., 2018] and PDP [Davis et al., 2017], coreference resolution problems can be treated as word-level binary classification problems. These datasets are prepared in a gold-two-mention style, containing paired sentences, the first of which has two or more mentions, and the second of which contains an ambiguous pronoun. A model should link the ambiguous pronoun to the correct mention. In this case, precision, recall, and F1 score (Section 2.3.1) are widely used for measuring both mention detection and mention linking evaluation.

However, coreference resolution problem can go beyond gold-two-mention problems because it usually includes clustering mentions into multiple coreference clusters and each cluster could contain multiple mentions (on the contrary, gold-two-mention problems only have one coreference link, which is the pronoun and its linked mention from the two candidate mentions). When mention linking is evaluated in general coreference resolution tasks, specialized metrics such as MUC (Section 2.3.2), B-Cubed (Section 2.3.3), CEAF (Section 2.3.4), BLANC (Section 2.3.5), and LEA (Section 2.3.6) are

employed. In this section, we also present evaluation metrics with special purposes (Section 2.3.7), e.g., metrics for measuring the gender bias of coreference resolution systems. Finally, we present the typical combinations of individual measures (Section 2.3.8) in evaluating coreference resolution.

2.3.1 F1 Score

It is common to use precision, recall and F1 score to evaluate the mention detection [Yu et al., 2020a, Peng et al., 2015] and binary selection-based mention linking tasks [Kocijan et al., 2019, Attree, 2019]. They are the most widely used performance measures in a binary classification task. Precision is defined as number of true positive predictions divided by number of all positive predictions:

$$precision = \frac{|true\ positives|}{|true\ positives| + |false\ positives|}, \tag{2.1}$$

where $|\cdot|$ denotes the number of items. Number of true positive predictions divided by number of actual positive items is referred to as recall:

$$recall = \frac{|true\ positives|}{|true\ positives| + |false\ negatives|}. \tag{2.2}$$

F1 score relates to the harmonic mean of precision and recall, which is given by

$$F_1 = \frac{2}{recall^{-1} + precision^{-1}}. \tag{2.3}$$

Sometimes, accuracy is also reported for measuring performance [Rudinger et al., 2018, Zhao et al., 2018, Kocijan et al., 2019]. It refers to the ratio of the number of properly anticipated items (the sum of true positive and true negative predictions) to the total number of items

$$accuracy = \frac{|true\ positives| + |true\ negatives|}{|total\ items|}. \tag{2.4}$$

2.3.2 MUC

MUC is the earliest coreference evaluation measure that was introduced by Vilain et al. [1995]. MUC is a measure that is based on links. Links are

coreferential relations between mentions. If two mentions corefer, there is a link between them. We define K as the key set which is a set of mentions that is clustered in the correct way. Each cluster within K is denoted as $k \in K$. All the mentions within the same cluster k are co-referential according to the hard truth (gold standard). R denotes the response set which is the set of mentions clustered by an evaluated model. r denotes one of the cluster within response set R. Then, the MUC Precision value is computed as below:

$$MUCPrecision(K, R) = \sum_{r \in R} \frac{|r| - |partition(r, K)|}{|r| - 1}, \tag{2.5}$$

where $|r|$ denotes the total number of mentions within cluster r, $|partition(r, K)|$ denotes the number of segments induced in the response cluster r in relation to the key clusters in K. It is formed by intersecting r with each key cluster $k \in K$ that overlaps with r. For example, if the mentions within a response cluster r belongs to 5 different key clusters $k \in K$, then $|partition(r, K)| = 5$, which means this response cluster r can be partitioned by K into 5 segments. We refer readers to Vilain et al. [1995] for more details regarding this calculation.

Similar to MUC Precision, the MUC Recall value is computed as below:

$$MUCRecall(K, R) = \sum_{k \in K} \frac{|k| - |partition(k, R)|}{|k| - 1}, \tag{2.6}$$

where $|k|$ denotes the total number of mentions within cluster k, $|partition(k, R)|$ denotes the count of segments of key cluster k relative to response set R. Each partition is formed by intersecting k and those response set $r \in R$ that overlaps with k [Vilain et al., 1995].

2.3.3 B^3(B-Cubed)

The B^3 score was introduced by Bagga and Baldwin [1998]. B^3 is a mention-based measure, i.e., the overall recall or precision is calculated by using the recall or precision of individual mentions. For each mention m_i, B^3 recall examines the proportion of overlapped mentions in both the key cluster (K_i) containing mention m_i and the response cluster (R_i) containing mention m_i

above the number of mentions in the key cluster (K_i) containing mention m_i. B^3 recall for mention m_i is computed as follows:

$$Recall_i = \frac{|K_i \cap R_i|}{|K_i|} \tag{2.7}$$

Similarly, B^3 precision for mention i is computed by changing the key clusters to response clusters in the denominator:

$$Precision_i = \frac{|K_i \cap R_i|}{|R_i|} \tag{2.8}$$

The final B^3 precision and recall are the weighted sum of individual entity scores:

$$\begin{aligned} Precision &= \sum_{i=1}^{N} w_i * Precision_i \\ Recall &= \sum_{i=1}^{N} w_i * Recall_i. \end{aligned} \tag{2.9}$$

Usually the weights (w_i) are assigned with $1/N$, where N represents the total number of mentions to be considered.

2.3.4 CEAF

The Constrained Entity Alignment F-measure (CEAF) proposed by Luo [2005] is used for entity or mention-based similarity detection. CEAF first creates a one-to-one mapping between response clusters and key clusters based on similarity. It then calculates accuracy and recall using this mapping. Luo [2005] provided four distinct forms of the similarity assessments:

$$\phi_1(K, R) = \begin{cases} 1 & \text{if } R = K \\ 0 & \text{otherwise} \end{cases} \tag{2.10}$$

$$\phi_2(K, R) = \begin{cases} 1, & \text{if } R \cap K \neq \emptyset \\ 0, & \text{otherwise} \end{cases} \tag{2.11}$$

$$\phi_3(K, R) = |R \cap K| \tag{2.12}$$

$$\phi_4(K, R) = 2 \cdot \frac{|R \cap K|}{|R + K|} \tag{2.13}$$

The key entities are represented by K, while the response entities are represented by R. $\phi_1(K, R)$ asserts that two entities are the same only if they share all the mentions, whereas $\phi_2(K, R)$ asserts that two entities are the same as long as they share at least a mention. $\phi_3(\mathrm{K}, \mathrm{R})$ is the number of shared mentions between key clusters and response clusters, whereas $\phi_4(\mathrm{K}, \mathrm{R})$ represents the number of shared mentions relative to the size of key clusters and response clusters.

CEAF comes in two flavors: mention-based and entity-based. The function $m(k)$ maps each key cluster k to a response cluster r using the Kuhn-Munkres algorithm [Kuhn, 1955]. The precision and recall of mention-based $CEAF_m$ are specified as follows:

$$CEAF_m Precision(K, R) = \frac{\sum_{k_i \in K} \phi\left(k_i, m\left(k_i\right)\right)}{\sum_{r_i \in R^*} |r_i|}, \tag{2.14}$$

$$CEAF_m Recall(K, R) = \frac{\sum_{k_i \in K} \phi\left(k_i, m\left(k_i\right)\right)}{\sum_{k_i \in K} |k_i|}, \tag{2.15}$$

where ϕ could be any function from ϕ_1 (Eq. 2.10) to ϕ_4 (Eq. 2.13), whereas ϕ_3 (Eq. 2.12) and ϕ_4 are most commonly used [Luo, 2005]. $|r_i|$ represents the total number of mentions within cluster r_i. $|k_i|$ represents the total number of mentions within cluster k_i. R^* represents the subset of response entities that can be mapped to K.

The precision and recall of entity-based $CEAF_e$ are computed as:

$$CEAF_e Precision(K, R) = \frac{\sum_{k_i \in K} \phi\left(k_i, m\left(k_i\right)\right)}{N_r}, \tag{2.16}$$

$$CEAF_e Recall(K, R) = \frac{\sum_{k_i \in K} \phi\left(k_i, m\left(k_i\right)\right)}{N_k}, \tag{2.17}$$

where N_r represents the total number of response entities and N_k represents the total number of key entities.

2.3.5 BLANC

BLANC [Recasens and Hovy, 2011] is a link-based measure that is based on rand indices. It looks at coreference links and non-coreference links separately. Recall and precision of coreference links are computed as:

$$R_c = \frac{|C_k \cap C_r|}{|C_k|}, \quad P_c = \frac{|C_k \cap C_r|}{|C_r|}, \tag{2.18}$$

where C_k represents the coreference links in the key clusters and C_r represents the coreference links in the response clusters.

Recall and precision of non-coreference links are computed as:

$$R_n = \frac{|N_k \cap N_r|}{|N_k|}, \quad P_n = \frac{|N_k \cap N_r|}{|N_r|},$$

where, N_k represents the non-coreference links in the key clusters and N_r represents the non-coreference links in the response clusters.

Final BLANC recall and precision are the average scores by coreference and non-coreference links

$$Recall = \frac{R_c + R_n}{2},$$

$$Precision = \frac{P_c + P_n}{2}.$$

2.3.6 LEA

In LEA [Moosavi and Strube, 2016], the proportion of successfully resolved connections between mentions is used to compute recall. The amount of mentions for each entity is weighted in the results, such that successfully resolving an entity with more mentions contributes more to the total score. Precision is calculated by inverting the key and response cluster roles.

$$\text{Recall} = \frac{\sum_{k_i \in K} \left[|k_i| \times \sum_{r_j \in R} \frac{\text{link}(k_i \cap r_j)}{\text{link}(k_i)} \right]}{\sum_{k_i \in K} |k_i|}$$

$$\text{Precision} = \frac{\sum_{r_i \in R} \left[|r_i| * \sum_{k_j \in K} \frac{\text{link}(r_i \cap k_j)}{\text{link}(r_i)} \right]}{\sum_{r_z \in R} |r_z|}$$

where for any cluster S, link(S) denotes the total number of edges of a complete graph with each node representing a mention from the same cluster. $\text{link}(S) = |S| \times (|S| - 1)/2$

2.3.7 Special aspects of analysis on coreference systems

There are other miscellaneous metrics which focus on certain specialized aspects, such as gender bias in coreference resolution. Zhao et al. [2018] created a new benchmark dataset called WinoBias, who measured the difference between pro-stereotyped and anti-stereotyped scenarios (e.g. in a woman dominated profession, linking a female pronoun with the job name is considered as 'pro-stereotypical', and linking a male pronoun with that job is considered as 'anti-stereotypical'). A robust coreference resolver should be able to handle both scenarios well. Besides that, the performance difference under the two scenarios should not be significant.

Similarly, Emami et al. [2019] proposed a corpus that switches candidate antecedents with different gender and number cues in order to mislead coreference resolvers. An outstanding system which relies on knowledge and contextual information should not be misled by such kind of lexical changes. The *consistency* score is thus defined as the proportion of correct predictions with the modified sentences in the corpus.

Varkel and Globerson [2020] and Wu et al. [2020] also used a bias factor. The bias factor is defined as F_1^f / F_1^m. It is the ration of F1 on feminine examples (f) and F1 on masculine examples (m).

If neither of the mentions in the gold-two-mention task is taken into account, the task is formalized as a three-class classification problem. Abzaliev [2019] used logarithmic loss to assess the model performance. The loss is given by

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log\left(p_{i,j}\right),$$

where N denotes the number of test set samples. M denotes number of classes ($M = 3$). $y_{i,j}$ is an indicator function. It takes value 1, if observation i belongs to class j. Otherwise, it takes value 0. $p_{i,j}$ denotes the predicted probability that i belongs to class j.

2.3.8 Combination of Measures

There is no single metric that is universally suitable in the coreference resolution domain, due to the complexity of the task. It is common to incorporate

several evaluation metrics together in the CR research. B-cubed, MUC and CEAF are the three most commonly used evaluation metrics in both entity and event coreference resolution tasks. Each of them can be formalized in the form of precision, recall and F1 score measures, respectively, where F1 scores are most commonly chosen as the overall measure. For example, Joshi et al. [2020] compared SpanBERT-based models, BERT-based models and end-to-end models with regard to F1 scores of MUC, B-cubed and $CEAF_{\phi_4}$ respectively. The final performance is measured by the averaged F1 score. This average score is also called the CoNLL score, as it was used by the CoNLL 2012 shared task [Pradhan et al., 2012, Huang et al., 2019]. Lu and Ng [2021a] further took BLANC into account when computing the average of F1 scores over event coreference tasks. The average of the four scores is also known as AVG-F. Additionally, for a more comprehensive overview of the performance of a coreference resolution system, detection precision, detection recall and detection F-measure could be also employed in coreference resolution performance evaluation [Zhang et al., 2018].

2.3.9 Summary of evaluation metrics

For gold-two-mention style coreference resolution tasks, F1 and accuracy are the standard evaluation metrics as they are quite intuitive. For higher order coreference resolutions, concerns have been expressed by researchers regarding the assessment measures that are used. Despite recent model advances, the CoNLL score remains the key evaluation measure employed by state-of-the-art models in recent years, which uses the F1 average of MUC, B^3 and $CEAF$. However, these three measures all have their pitfalls. MUC is considered to have the weakest ability to differentiate good and bad coreference resolution results [Recasens and Hovy, 2011]. It also prefers coreference result that is over-merged [Luo, 2005]. B^3 may lead to counter-intuitive results under some edge cases [Luo, 2005]. It cannot handle repeated mentions very well [Luo and Pradhan, 2016]. $CEAF_e$ treat mention clusters equally irrespective of their sizes [Stoyanov et al., 2009]. Additional measurements have been created in recent years to overcome the limitations of these three

traditional metrics. Agarwal et al. [2019] established new measures for evaluating name entity coreference (NEC) after determining that existing metrics did not meet the criteria of the NEC task. Moosavi and Strube [2016] introduced the LEA measure to account for the importance of entities with greater mentions. In addition to traditional measurements, researchers are advised to consider using these new metrics as well.

2.4 Datasets and Learning Methods

2.4.1 Datasets

CoNLL 2012 shared task [Pradhan et al., 2012] proposed three coreference resolution datasets in English, Chinese and Arabic. The datasets were built upon OntoNotes v5.0 [Hovy et al., 2006], containing texts from different sources, e.g., broadcast, magazine, newswire, weblogs and newsgroups. The texts may involve multiple speakers, e.g., in broadcast and telephone conversations. Alternatively, the texts are monologues. The English dataset of CoNLL-2012 shared task includes 2802 training documents, 343 development documents and 348 testing documents. The Chinese version dataset consists of 1810 training documents, 252 validation documents and 218 test documents. The Arabic version dataset consists of 359 training documents, 44 validation documents and 44 test documents. To the best of our knowledge, the current state-of-the-art model is Wang et al. [2021] with CoNLL score of 87.5%.

The GAP dataset [Webster et al., 2018] was sourced from Wikipedia snippets. Each snippet is annotated with one gender-ambiguous pronoun, two names, and two flags. A model has to decide which name the gender-ambiguous pronoun refers to. The model is then evaluated, based on the coreference connections between the two names and the pronoun. F1 scores on masculine, feminine and overall examples are commonly used metrics on the GAP dataset. And the ratio F_1^f/F_1^m (F1 score for feminine examples over F1 score for masculine examples) is also calculated to evaluate the gender bias. GAP dataset contains 8908 pairs of ambiguous pronouns and candidate

mentions. The training, validation and testing snippets have 4000, 908, and 4000 samples, respectively. The current state-of-the-art on GAP test dataset is the ProBERT [Attree, 2019] with F1 score of 92.5% and 0.97 gender bias.

The TAC KBP Event Track dataset [Mitamura et al., 2017] is used to resolve event coreference in the TAC KB 2017 shared task. The goal of the TAC KBP Event track is to extract information about events such that the information could be used as inputs into a knowledge base. Event Nugget (EN) Detection, Coreference, and Sequencing tasks, as well as Event Argument and Linking (EAL) tasks in the shared task are evaluated at the document level. Except for event sequencing, all other event tasks are in three languages, namely English, Chinese, and Spanish. The KBP 2017 shared task provided a standard measure, AVG-F. AVG-F is the average of four widely used metrics: MUC, B^3, $CEAF_e$ and BLANC (see Section 2.3). KBP 2017 dataset contains 167 documents. The state-of-the-art performance is introduced in the work of Yu et al. [2020b] with an AVG-F of 57.12%.

The Linguistic Data Consortium (LDC) created the ACE (Automatic Content Extraction) 2005 Multilingual Training Corpus, which comprises about 1,800 files of mixed genre texts with annotations in entities, relations, and events in English, Arabic, and Chinese. ACE2005 is the whole collection of the training data with various languages in the 2005 ACE technology evaluation. The genres cover the texts from newswire, broadcast news, broadcast conversation, weblog, discussion forums, and conversational telephone voice. LDC annotated the data with assistance from the ACE Program and additional assistance from LDC. ACE2005-English contains 599 files, ACE2005-Chinese contains 633 files and ACE2005-Arabic contains 403 files. The state-of-the-art performance on ACE2005-English is introduced by Lai et al. [2021] with a CoNLL score of 87.90% and AVG-F score of 88.30%.

LitBank is a new dataset of literary text coreferences [Bamman et al., 2020]. The collection contains 100 literary texts with an average length of about 2100 words. Singletons are recognized and evaluated. The original evaluation was based on 10-fold cross validation with 80%, 10%, and 10% data splits for training, validation and testing. It restricts the mentions to

six entity categories (location, organizations, people, vehicles, geo-political entities, facilities) with the bulk of mentions (83.1%) pointing to entities belonging to the people category. Khosla and Rose [2020] introduced the state-of-the-art performance with CoNLL score of 80.26% on this dataset.

The Winograd Schema Challenge (WSC) [Levesque, 2011] is a hard pronoun resolution challenge based on Winograd's [Winograd, 1972a] examples.

A Winograd Schema example reflects the situation where a single word modification in a sentence changes the referent of the pronoun, making the resolution difficult. The goal is to determine which entity the pronoun or possessive adjective refers to in a context. The context includes two entities. The text contains a "special word". The statement remains technically valid when the "alternative word" is used for substitution, whereas the referent of the pronoun changes. Consider the example below:

> *William could only climb beginner walls while Jason climbed advanced ones because he was very [weak/strong].*

In this sentence, *weak* is a special word while *strong* is an alternative word. When *weak* is used in this sentence, pronoun *he* will refer to *William*. If *strong* is used to substitute *weak*, then the pronoun *he* will refer to *Jason* instead.

The Winograd Schema Challenge consists of challenging cases that need commonsense to answer. These cases could not be solved simply using statistical analysis of co-occurrences and associations. The SuperGLUE [Wang et al., 2019] version of WSC dataset contains 554 training examples, 104 validation examples and 146 test examples. The state-of-the-art model is claimed to be ERNIE 3.0 [Sun et al., 2021] with accuracy of 97.3%.

The Definite Pronoun Resolution (DPR) corpus [Rahman and Ng, 2012] is a modified version of Winograd Schema Challenge-style issues. These sentence pairs span a wide range of themes, from real occurrences to cinematic events to entirely fictitious circumstances, primarily representing pop culture as experienced by American children born in the early 1990s. DPR includes cases that do not need commonsense reasoning, as well as situations where

the "special word" is a phrase. DPR contains 1322 training examples and 564 test examples. Totally, there are 1886 example sentences. The state-of-the-art model on DPR dataset is the BERT_WIKICREM_ALL [Kocijan et al., 2019] with an accuracy of 84.8%.

The Pronoun Disambiguation Problem (PDP) [Davis et al., 2017] is a modest set of 60 questions that served as the first round of the 2016 Winograd Schema Challenge. Unlike WSC, the cases do not involve a "special word" in PDP. However, they still need commonsense thinking to understand the texts. The samples were hand-picked from literature. The state-of-the-art model for PDP is BERT_WIKICREM_ALL [Kocijan et al., 2019] with an accuracy of 86.7%.

Winogender [Rudinger et al., 2018] is a dataset for testing the gender biases in coreference resolution, using the WSC format. Each sentence has an occupational noun and a referring pronoun. The pronoun could be represented as "he", "she" or "they", respectively. The occupational nouns are usually gender-oriented. E.g., women are likely to be employed as secretaries. Given "the secretary asked the visitor to sign in so that he could update the guest log" [Rudinger et al., 2018], a coreference resolution classifier may fail in connecting "he" to "secretary", if the classifier is gender-biased. This dataset means to examine how altering the gender of the pronoun impacts the accuracy of a model. Winogender contains 720 sentences in total. The state-of-the-art model on this dataset is BERT_WIKICREM_DPR [Kocijan et al., 2019] with accuracy of 82.1%.

WinoBias [Zhao et al., 2018] is also a WSC-inspired dataset that measures gender biases in coreference resolution algorithms. Similar to Winogender, WinoBias contains examples of occupations with a high gender imbalance. It contains 3160 Winograd Schemas examples, equally divided into training and test sets. The test set examples are divided into two types, where Type 1 examples are prototypical WSC phrases. Coreference judgments must utilize world knowledge based on the given conditions. Such instances are difficult to understand because they lack syntactic clues. Type 2 examples utilize syntactic knowledge and a pronoun comprehension. Since

both semantic and syntactic clues aid in disambiguation, resolvers are likely perform better in Type 2 instances. The gender of the pronominal reference is immaterial for the co-reference judgment in both types. To pass the test, systems must be able to produce valid linkage predictions in both pro- and anti-stereotypical circumstances. The stereotyped jobs were chosen using data from the US Department of Labor. The best performance is introduced by BERT_DPR [Kocijan et al., 2019] on Type 1 subset (with accuracy of 78.0%-78.2%) and BERT_WIKICREM_ALL [Kocijan et al., 2019] on Type 2 subset (with accuracy of 98.7%-99.0%).

Emami et al. [2019] introduced KnowRef, a coreference resolution corpus that particularly tests the capacity of a system to reason about a scenario stated in the context.

KnowRef is a human-labeled corpus with 8,724 Winograd-like text samples, the resolution of which necessitates considerable commonsense and domain knowledge. Each instance consists of a brief text with a target pronoun that must be appropriately resolved to one of two potential antecedents. The KnowRef dataset was created by collecting text samples from a vast collection of documents, including 2018 English Wikipedia, OpenSubtitles, and Reddit comments. KnowRef contains 7455 training sentences and 1269 testing sentences. The state-of-the-art model on KnowRef is BERT(KnowRef) [Emami et al., 2019].

WikiCoref [Ghaddar and Langlais, 2016] includes annotated Wikipedia documents. Documents were carefully chosen to span a variety of stylistic articles. Each mention is annotated with entity type and coreference properties, as well as the Freebase subject to which it belongs. The annotation scheme of WikiCoref is the extension of the OntoNotes scheme. WikiCoref consists of 30 documents with an average document size of 2000 tokens. Khosla and Rose [2020] held the best performance with a CoNLL score of 71.35%.

Extension to Event Coreference Bank (ECB+) [Cybulska and Vossen, 2014] consists of within- and cross-document coreference annotations for entities and events. The identification of groupings of related texts that

describe the same foundational event is a key stage in the construction of the ECB+ corpus, enabling for the annotation of coreferential event references across documents. Different topics from Google News archives were chosen in order to contain intentionally selected keywords. ECB+ contains 976 documents in total which are divided into 574 documents for training, 196 documents for validation and 206 documents for testing. The current state-of-the-art model for ECB+ is Cross Document Language Model (CDLM) [Caciularu et al., 2021a] with a CoNLL score of 85.6%.

Richer Event Description (RED) corpus [O'Gorman et al., 2016] annotates entities, events, and times, as well as their coreference connections and the temporal, causal, and subevent linkages between the events. It contains 8731 events, 1127 temporal expressions, and 10320 entities in 95 documents (totaling 54287 tokens), sampled from both news data and casual discussion forum interactions. It includes 2390 identity chains, 1863 bridging relations, and 4969 event-event relations that include temporal, causal, and subevent relationships, as well as 8731 DocTimeRel temporal annotations that connect these events to the document time.

Georgetown University Multilayer corpus (GUM) [Zeldes, 2017] was collected in the context of classroom teaching. It includes rich annotated texts of twelve genres from various sources including Wikinews, Wikivoyage, Wikihow, Reddit. Main annotations in this corpus include multiple Part-of-Speech (POS) tags, document structure in TEI XML (paragraphs, headings, figures, etc.), constituent and dependency syntax, entity and coreference annotation, discourse dependencies. It includes 168 documents with 150824 tokens.

The Wikipedia Event Coreference [Eirew et al., 2021] is a data set for a cross-document event coreference task. Data annotation is boosted by leveraging available information in Wikipedia while the coreferences are not restricted by predefined topics. The information is gathered by grouping together the anchor texts of (internal) Wikipedia links pointing to the same Wikipedia concept. This is typically justified because all of these links are about the same real-world subject. As a result, the WEC dataset is made

up of mentions, each of which contains the mention span corresponding to the link anchor text, the surrounding context, and the mention cluster ID. Since Wikipedia was not divided into predefined topics, mentions can have coreference links across the entire corpus. WEC training set consists of 40529 event mentions in 7042 clusters. The validation set consists of 1250 event mentions in 233 clusters. The test set consists of 1893 event mentions in 322 clusters. The state-of-the-art model is introduced in Eirew et al. [2021] with a CoNLL score of 62.3%.

EmailCoref [Dakle et al., 2020] includes 46 email threads and 245 email messages. This is the first dataset to address the problem of entity resolution in email threads. It has set two rules for choosing email threads: The thread must have at least three email messages, with at least half of the email messages including text content. EmailCoref contains 36 training email threads and 10 testing email threads. Khosla and Rose [2020] introduced the best performance with a CoNLL score of 76.17%.

Levy et al. [2021] presented BUG, a large scale gender bias dataset that has similar challenging style as Winogender [Rudinger et al., 2018] and WinoBias [Zhao et al., 2018]. BUG was semi-automatically collected with help of SPIKE [Shlain et al., 2020] from three different sources: Wikipedia, PubMed and Covid19 research papers. BUG has 108K sentences and the state-of-the-art model is the SpanBERT fine tuned on anti-stereotypical part of BUG with an accuracy of 64.1%.

Table 2.1 provides an overview of datasets used in coreference resolution research, detailing their size (training, validation, test set, and total examples) and task focus (entity or event coreference resolution). Datasets like KBP 2017, ACE 2005, ECB+, RED, and WEC are primarily used for event coreference resolution, with ECB+ also applicable to entity coreference. Other datasets mainly target entity coreference resolution.

Table 2.2 outlines the application scenarios of these datasets and lists state-of-the-art (SOTA) models for each, with further model details discussed in subsequent sections. It's noted that not all models are evaluated on the same datasets, reflecting their varied design focuses, making direct comparison challenging.

Table 2.1: The statistics and features for different coreference resolution datasets

	Size				Tasks	
Dataset	Train	Val.	Test	Total	Entity	Event
CoNLL 2012[Pradhan et al., 2012]	2802	343	348	3493	✓	
GAP[Webster et al., 2018]	4000	908	4000	8908	✓	
KBP 2017[Mitamura et al., 2017][1]	-	-	167	167		✓
ACE 2005	529	28	40	599		✓
LitBank[Bamman et al., 2020]	-	-	-	100	✓	
WSC[Levesque, 2011]	554	104	146	804	✓	
DPR[Rahman and Ng, 2012]	1322	-	564	1886	✓	
PDP[Davis et al., 2017]	-	-	-	60	✓	
Winogender[Rudinger et al., 2018]	-	-	-	720	✓	
WinoBias[Zhao et al., 2018]	1580	-	1580	3160	✓	
KnowRef[Emami et al., 2019]	7455	-	1269	8724	✓	
WikiCoref[Ghaddar and Langlais, 2016]	-	-	-	30	✓	
ECB+[Cybulska and Vossen, 2014]	574	196	206	976	✓	✓
RED[O'Gorman et al., 2016]	-	-	-	95		✓
GUM[Zeldes, 2017]	-	-	-	168	✓	
WEC[Eirew et al., 2021]	40529	1250	1893	43672		✓
EmailCoref[Dakle et al., 2020]	36	-	10	46	✓	
BUG [Levy et al., 2021]	-	-	108K	108K	✓	

2.4.2 Tasks and Learning Methods

Table 2.3 summarizes the target tasks and learning methods of models from Section 2.5 to Section 2.8. This table indicates whether models focus on entity or event coreference resolution and their reliance on rule-based, traditional machine learning, or deep learning methods. Earlier models (Section 2.5) often employ traditional machine learning, whereas later models (Sections 2.6-2.8) predominantly use deep learning techniques. Deep learning-based models vary in structure, with recurrent neural networks as a common foundation, but later replaced by Transformers in pre-trained language model-based methods. Supplementary techniques like gated mechanisms, attention mechanisms, and reinforcement learning are also utilized in some deep learning models.

2.4.3 Features

Table 2.4 outlines the features used in models from Section 2.5 to Section 2.8, including semantic and syntactic features, word embeddings, and pre-trained language models. Earlier models, particularly feature-based and early neural network models, primarily use semantic and syntactic features. Post-Lee et al. [2017], word embeddings become the norm for mention representation. The use of pre-trained language models for mentions begins with Joshi et al. [2019]. While these models seldom explicitly incorporate semantic and syntactic information, exceptions exist, such as Khosla and Rose [2020] using NER tags and Wang et al. [2021] considering word position.

2.5 Feature-based Approaches

Feature-based models include linguistic information such as part-of-speech (POS) and named-entity recognition (NER) tags, as well as surface-level semantic information, such as opinion words. The models themselves are also not based on deep learning, but rather on standard Machine Learning approaches (e.g. decision trees, memory-based learning and conditional random fields). These models represent the early stages of coreference resolution research. They are typically fairly intuitive, whereas they are incapable of capturing deep level contextual information and comparatively lack generalization capacity.

The objects and attributes of coreference resolution issues, proposed by Ding and Liu [2010], are the challenge of detecting whether references of objects and attributes correspond to the same entities. To tackle the problem, they employed a supervised learning approach. The major contribution of the article is the creation and testing of two unique opinion-related characteristics for learning. The first feature was based on non-comparative sentence sentiment analysis, comparative sentence sentiment analysis, and the idea of sentiment consistency. The second feature took into account which objects and attributes were modified by which opinion words. Opinion words, such as *good*, *best*, *bad*, and *poor*, are often used to convey positive

or negative feelings. Their model workflow included preprocessing, feature vector construction, classifier construction, and testing. The model first preprocessed the corpus by running a POS tagger and a Noun Phrase finder. They then generated the object-noun phrase (O-NP) set, which includes potential objects, attributes, and other noun phrases. Then, for each pair in the O-NP set, they created a feature vector. Since their study focused on products and attributes, they left out personal pronouns, gender agreement features, and appositive features. Training data were created in the classifier construction step with each pair containing at least one object or attribute. To fit the training data, a decision tree was built. Several novel features in the opinion mining context were proposed in this work, including sentiment consistency, object/attribute, and opinion word associations.

Atkinson et al. [2015a] combined features-based coreferencing and memory-based learning which improves opinion retrieval in social media. The working model was built on top of three main tasks: massage retrieval, message preprocessing and reference analysis, and opinion retrieval. Message retrieval collected and stored the hierarchies of various tweets in a local database. Tokenization, POS tagging and named-entity identification were used to extract essential underlying linguistic information from the collected tweets. In the message referencing analysis and opinion retrieval stage, a memory-based learning approach (MBL) was utilized. A Machine Learning approach that searches for the training data item that is most similar to the test data item and make predictions based on the similarity is referred to as an MBL. As major generalization approaches, memory-based learning systems employed nearest-neighbor search, space decomposition techniques, and clustering. This feature-based referencing classification model was tested using formal and informal text corpora. The results showed that the accuracy for extracting referential links on the formal texts improved more compared with the informal texts, due to the linguistic features of informal messages.

A joint model of three essential activities for the entity analysis stack was provided by Durrett and Klein [2014a]: coreference resolution, the identification of entities and the entity linking. The joint model took unary, binary

and ternary factors into account when solving these three problems. Unary factors were features employed when solving each task in isolation. Binary and ternary factors were introduced to capture cross-task interactions. For example, the restriction of coreferential references having the same semantic kind. Based on Durrett and Klein [2014a]'s original argument for jointly modeling, namely that the three tasks have possible impacts on each other, they showed that making use of the interactions between the modules resulted in higher performance overall. As a result, any pipelined system would inevitably underperform a combined model.

The mention-ranking technique to coreference was used in Durrett and Klein [2014a]. Their feature set focused on the surface features of mentions, such as starting and ending word, mention length, and the syntactic role of each mention. Coreference features incorporated multiple features between mention pairs as well as aspects of the mention pair itself, such as distance between mentions and whether their heads matched. Anaphora features explored each of these qualities in turn.

Raghunathan et al. [2010a] applied a multi-level sieve structure that applied one sieve at each level, the sieve with the higher precision will always come before the lower precision sieve. This design aims to avoid the phenomenon of lower precision feature prevailing over the higher precision feature.

In summary, common features that are widely used in feature-based approaches are opinion words [Ding and Liu, 2010], POS tags [Ding and Liu, 2010, Atkinson et al., 2015a], text chunking [Ding and Liu, 2010], NER tags [Atkinson et al., 2015a, Raghunathan et al., 2010a], semantic information [Durrett and Klein, 2014a], syntactic roles [Durrett and Klein, 2014a], word positions and head words [Durrett and Klein, 2014a, Raghunathan et al., 2010a]. Feature-based approaches mainly represent the early stage of coreference resolution studies. Their performance has been exceeded by the latest deep learning-based approaches. Since feature-based approaches are not the focus of coreference resolution research community in the past decade, we only list a few approaches, published after 2010. If readers are interested in very early studies, one can refer to the survey of Ng [2010].

2.6 Multilayer Perceptron/Recurrent Neural Network Approaches

Neural-based models in coreference resolution, primarily from the pre-BERT era, leverage neural networks for contextual understanding and abstracting features into high-dimensional vectors, typically without external knowledge beyond the training dataset.

Clark and Manning [2015] demonstrated how scores from mention pair models can be aggregated to create entity-level insights for building coreference chains. These scores were used in an entity-centric system trained to progressively form coreference clusters, using prior decisions to inform future ones.

Wiseman et al. [2016] emphasized the need for global context in coreference resolution and introduced a Recurrent Neural Network (RNN) for sequential mention cluster representation training. Their model, integrating these representations into a mention-ranking coreference system, was trained end-to-end without explicit clustering features.

Lee et al. [2017] introduced the first end-to-end coreference resolution (CR) model, outperforming previous models without a syntactic parser or mention detector. It considered all text spans as potential mentions, learning a distribution over antecedents. The model integrated contextual boundary representations and a head-finding attention mechanism for span embeddings, trained to maximize the likelihood of gold antecedent spans.

This model employed bidirectional LSTMs for vector representations, including endpoint BiLSTM hidden states and an attention vector over span tokens. It implicitly learned to generate mention candidates and developed a task-specific head-finding attention mechanism, excelling on the OntoNotes benchmark without additional preprocessing.

Lee et al. [2018] enhanced this model with higher-order inference, using attention mechanisms to iteratively refine span representations. This approach allowed for multi-hop evaluations within predicted clusters, addressing the limitations of first-order models that might produce locally consistent

but globally inconsistent clusters. They introduced a coarse-to-fine strategy for computational efficiency, initially using a simple bilinear scoring function for antecedent pruning before applying a more detailed scoring method.

However, Xu and Choi [2020] later found that higher-order inference did not significantly improve performance when using advanced encoders like SpanBERT Joshi et al. [2020].

2.7 Knowledge-based Models

Recent knowledge-based models, like neural-based contextual models, are trained using neural networks but also integrate external knowledge (e.g., commonsense or domain-specific knowledge bases) typically stored in triplets.

Aralikatte et al. [2019] combined Wikipedia and Wikidata knowledge with reinforcement learning models for coreference resolution, using world knowledge. They enhanced coreference resolver performance by evaluating predictions against an OpenRE system and knowledge bases. To extend beyond the knowledge base limits, a Universal Schema model Riedel et al. [2013] was implemented, using its confidence as a reward function. This model, fine-tuned using policy-gradient, aligned predictions with world knowledge, converting documents into subject-relation-object triples for resolver updates Angeli et al. [2015].

Multiple Universal Schema models Riedel et al. [2013], Verga and McCallum [2016] were developed to align with world knowledge, resulting in three distinct reward functions based on Wikidata connections. These functions, focusing on different entity relationship aspects, informed the RE-Distill model, which incorporated Coref-KG, Coref-Text, and Coref-Joint strategies through multi-task reinforcement learning, following DisTraL Teh et al. [2017] principles. The combined strategy, fine-tuned with a unified reward function, surpassed the Lee et al. [2018] model in mention identification and linkage Aralikatte et al. [2019].

Emami et al. [2018] created a system adept at the Winograd Schema Challenge and COPA tasks, requiring complex reasoning and knowledge.

Their approach used a knowledge-hunting module to form web search queries from input issues, categorize returned information, and make decisions. This framework processed Winograd phrases through semantic mapping, query generation, and search engine utilization, using the Stanford CoreNLP coreference resolver Raghunathan et al. [2010b] and Python Selenium for web scraping. Search results, restricted to relevant sentence snippets, informed final coreference determinations.

Zhang et al. [2019] explored enhancing coreference resolution using external knowledge and contextual information, directly incorporating triplets Lin et al. [2023], a common format in modern knowledge graphs. Their knowledge attention module, which adapts to various contexts, demonstrated superior performance on CoNLL and i2b2 datasets, outdoing baselines and showing better cross-domain adaptability compared to Lee et al. [2018]. The model's integration of external knowledge sources like commonsense knowledge graphs and medical concepts proved effective.

2.8 Transformer-based Pre-trained Models

Prior to the introduction of transformer architecture, the most widely used sequence conversion models were built on top of advanced CNN or RNN models that included both encoding and decoding processes. For boosting performance, the best versions included attention mechanisms to link the encoder and decoder together. Vaswani et al. [2017] proposed Transformer, a new basic network designed based on attention mechanism, without recurrent or convolutional structures. Transformer is trained considerably quicker than recurrent- or convolutional-based encoders for seq2seq tasks, such as language translation.

BERT (Bidirectional Encoder Representations from Transformers) [Devlin et al., 2019] is a Transformer-based pre-trained language model. BERT is designed to pre-train representations from unlabeled texts by conditioning all layers on the entire sentence. As a consequence, BERT can be fine-tuned with just one additional layer to offer cutting-edge models for a broad variety

of tasks, such as question answering and language inference, without needing substantial architecture modifications. In this section, we introduce coreference resolution models that incorporate contextualized representations from BERT and its variants.

Joshi et al. [2019] used BERT to resolve coreferences. Their model is based on top of a coarse to fine coreference model, termed c2f-coref for short from Lee et al. [2018].

In the work of Joshi et al. [2019], BERT fully replaced the LSTM-based encoder in c2f-coref. The span representations were the concatenation of beginning word piece, ending word piece, as well as the attended form of the whole span. Joshi et al. [2019] divided documents into parts before applying BERT in two ways: independent method and overlap method. The independent variation employed independent segments without overlapping. Each segment served as an independent input of BERT. By moving a sliding window T/2 steps each time, the overlap variations split the document into T-sized overlapping sections. The BERT encoder was then fed each segment separately, and the final representation was created via element-wise interpolation of the overlapping segment embeddings. The BERT-based models were tested against two datasets: the paragraph-level GAP dataset [Webster et al., 2018], and the document level CoNLL 2012 dataset [Pradhan et al., 2012]. According to Joshi et al. [2019], both BERT-base and BERT-large outperformed the ELMo-based c2f-coref, with BERT-large beating the original c2f-coref by a larger margin.

Joshi et al. [2020] proposed SpanBERT pre-trained language model, developing a coreference resolution task-specific model by combining SpanBERT and the work of Lee et al. [2018]. The main difference between Joshi et al. [2020] and Joshi et al. [2019] is that Joshi et al. [2020] used SpanBERT as the encoder rather than BERT. SpanBERT has the same architecture as BERT [Devlin et al., 2019], whereas it was trained with a different masking method in which spans were masked. The outer boundaries of spans were trained to predict all tokens inside the masked spans, which was called span-boundary objective (SBO). It is useful for coreference resolution, since

entity mentions are often spans of tokens. Span ranking models benefit from better span representations.

SpanBERT retained the regular masked language model (MLM) objective in vanilla BERT but substituted SBO for next sentence prediction (NSP), because Joshi et al. [2020] discovered that single sequence training outperformed bi-sequence training on downstream tasks.

In another research later, Xia et al. [2020] reduced the memory usage of the original SpanBERT [Joshi et al., 2020] with an incremental algorithm. It kept the track of clusters, each of which had its own representation. The model suggested a possible set of spans for a particular phrase or segment. A scorer compared each span representation to all of the clusters, determining the the best fit cluster. Following the addition of the new span, the representation of the chosen cluster was likewise changed. The model periodically evicted less important entities and wrote them to disk. Each clustering choice made by this method was permanent.

Lai et al. [2022] incorporated the SpanBERT encoder into the e2e-coref model of Lee et al. [2017] but introduced a few simplifications to the original e2e-coref structure. Lai et al. [2022] excluded span length information when generating span representation and excluded feature information such as genre and distance when doing mention linking. It also reduced the number of candidate mentions when doing mention extraction. Despite those simplifications, Lai et al. [2022] still achieved comparative results with Joshi et al. [2020].

Unlike most previously discussed methods that have no chance of recovering a missed mention, Wu et al. [2020] permitted the mention linking module to find mentions that were missed during the mention proposal phase. The proposed model CorefQA defined coreference resolution as a span prediction issue under a question answering setting. It first generated a query for each mention before extracting the relevant mentions depending on the query. As long as at least one of the candidates in the associated coreference cluster was utilized in the query, other candidates in the cluster could be recovered during the mention linking phase.

CorefQA was made up of two modules: mention proposal and mention linking. The mention score was calculated using a FFNN taking into account the SpanBERT representation of the first and last constituent token of spans. Only spans with a mention score greater than a predefined threshold were retained in this module.

For the mention linking module, the query and the context were combined into a single sequence, and BIO (Beginning, Inside and Outside) tags were assigned to tokens that constituted candidate mentions. The probability of assigning one of BIO tags to a certain token was calculated via feed forward neural network. Thus, the probability of span j being coreferent to i depended on the probability that BIO tags were assigned correctly.

Furthermore, Wu et al. [2020] augmented data with Quoref dataset [Dasigi et al., 2019] and the SQuAD dataset [Rajpurkar et al., 2016], as well as using the speaker modeling strategy which directly combined the speaker names with the utterance, rather than converting the speaker information into binary features.

Khosla and Rose [2020] incorporated semantic knowledge into the model of Bamman et al. [2020]. It reduced errors that were caused by type mismatches in coreference resolution. For each token, the model of Khosla and Rose [2020] first passed the BERT embeddings through a bi-directional LSTM in order to get the corresponding representation. The representation of a mention was given by a concatenation of token representations and different features including entity type. Coreference score of two mentions was given by a feedforward neural network whose input is the concatenation of mention representations, their element wise product and different mention-pair features including whether they have identical entity types. Empirical result showed that models incorporating type information outperformed baseline models without type information on four coreference resolution datasets. Thus Khosla and Rose [2020] argued that explicitly incorporating external knowledge would further benefit contextualized embedding-based models, e.g., BERT-based models.

Wang et al. [2021] introduced a reinforcement learning-based resolver capable of handling problems, caused by the same mentions appearing in

different document contexts. They utilized mention-level training examples, rather than merely sentence- or document-level samples. This algorithm has the advantage of mitigating the detrimental effect of noisy sentence-level information while retaining enough contextual information. The distance between two mentions was also taken into account in the work of Wang et al. [2021], since co-reference is sensitive to the mention distance. The span representation of Wang et al. [2021] was a combination of BERT embedding and a head-finding attention mechanism. The representations of two spans to be judged were then passed on to a actor-critic-based reinforcement learning model with two neural networks representing actor and critic separately. The states were the concatenation of the two mention spans. Action was defined as whether to create and store the links between the two spans and then move on to the next pair of spans. Reward was a biaffine attention mechanism to model the probability for the two spans to be coreferential. It also considered the distance between the two spans as there was usually a inverse relation between the distance and the coreference probability.

Kocijan et al. [2019] fine-tuned BERT with WikiCREM. When the model was trained, sentences containing one masked personal name and two candidate mentions were given and the goal was to choose the more suitable candidate from the two. The objective function was a combination of the negative log-likelihood of the correct candidate as well as the max-margin loss term of the two candidates. It was observed that this combination of losses consistently outperformed single loss terms alone on various tasks.

$$\begin{aligned}\mathcal{L} = &- \log \mathbb{P}(\mathbf{a} \mid S) + \\ &+ \alpha \cdot \max(0, \log \mathbb{P}(\mathbf{b} \mid S) - \log \mathbb{P}(\mathbf{a} \mid S) + \beta)\end{aligned}$$

where **a** represents the correct candidate. **b** represents the incorrect candidate. S denotes the sentence that contains the masked personal name. α and β are hyperparameters that control the influences of the loss components.

Attree [2019] presented a evidence-based deep learning model for the GAP shared task. It includes two main components: Pronoun BERT module and Evidence Pooling module. Pronoun BERT module extracted the

last layer embedding for the pronoun from the BERT model. Evidence Pooling module combined the clustering information from four other coreference resolution models: AllenNLP [Gardner et al., 2017], NeuralCoref[3], Parallelism+URL [Webster et al., 2018] and e2e-coref [Lee et al., 2017]. The Evidence Pooling would encode the information from all these models via self-attention mechanism and generate an evidence vector. The readers is referred to Attree [2019] for details about how this evidence vector is generated. Finally, the evidence vector is concatenated with the BERT embedding of pronoun to go through the linear and softmax layer to get the classification result.

CorefBERT[Ye et al., 2020] employed two training tasks: mention reference prediction (MRP) and MLM. For the input tokens $X = (x_1, x_2, \ldots, x_n)$, each token was first represented by aggregating the embeddings of token and positional information, and then the input representations were fed into the bidirectional Transformer to obtain hidden states $\boldsymbol{H} = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n)$, which were then used to compute the loss. The final loss function of CorefBERT was the sum of MRP loss and MLM loss, among which, the MRP loss was defined as a function to jointly maximize all the probability of choosing a word in the sequence to recover the masked word.

Yu et al. [2020b] presented pairwise representation learning (PAIRWISERL) which was used for both entity coreference resolution and event coreference resolution. It treated entity coreference as a simplified version of event coreference resolution because event coreference resolution also includes arguments besides trigger itself. When processing event coreference resolution, PAIRWISERL concatenated two sentences containing the two events and passed it through RoBERTa [Liu et al., 2019a] to get the representation for event triggers and four arguments: subject, object, time, and location. For each argument, PAIRWISERL concatenated representation from both sentences as well as their element wise product, then passing the concatenated vector through feedforward neural networks in order to get the compatibility scores for that argument. The final binary classification result

[3]https://github.com/huggingface/neuralcoref

was given by a multilayer perceptron where the inputs is the concatenation of RoBERTA representation of two trigger words, their element wise product and the compatibility scores for the four arguments.

Lai et al. [2021] proposed a gating mechanism to selectively extract information from predicted features. The predicted features of event mentions included type, polarity, modality, genericity, tense and realis [Mitamura et al., 2016]. For each event mention, its own representation was obtained using SpanBERT encoder [Joshi et al., 2020]. The K symbolic features were converted into K vectors, using trainable matrices. Lai et al. [2021] then proposed a context-dependent gated module to filter information for each feature. In addition, Lai et al. [2021] introduced the noisy training method for regularization by randomly replacing some predicted feature values with some noise before feeding the input data into the model. By doing this, it could force the model to identify reliable features.

Caciularu et al. [2021a] presented Cross Document Language Modeling (CDLM). All the related documents were concatenated and fed to the Longformer encoder [Beltagy et al., 2020] during pre-training. Caciularu et al. [2021a] masked 15% of the tokens in each training example and forced the model to predict the masked token, based on the whole set of documents, rather than the individual document. Caciularu et al. [2021a] employed the Multi-News dataset [Fabbri et al., 2019] which contains 44972 document clusters for pre-training. For each pre-training example, documents within the same cluster were randomly picked in order to make sure that the documents were related. During the fine-tuning of coreference resolution, relevant documents were concatenated into a single sequence with document separator tokens ([CLS]) at the beginning of the sequence. The pair-wise vector representation $m_t(i, j)$ between mention i and mention j within the t-th example was the concatenation of CDLM representations of the [CLS] token, mention i and j, and their element wise product. $m_t(i, j)$ was then passed through a multi-layer perceptron to get the binary classification result (coreferent or not).

In order to reduce the large memory footprint faced by many previous models, Kirstain et al. [2021] presented start-to-end (s2e) model which only

use information on the start and end tokens of the span in order to calculate the mention score and antecedent score. By doing this, it reduced the memory footprint significantly compared with Joshi et al. [2020]. s2e model utilized bilinear functions between pairs of endpoints tokens to calculate mention score f_m and antecedent score f_a without relying on the span level representation.

Similar to Kirstain et al. [2021], Thirukovalluru et al. [2021] also aimed at reducing the memory and time cost of coreference resolution systems. They presented an approximation to the end-to-end model Lee et al. [2017] which can scale to long documents. Beside using token level bilinear inference to calculate scores, it also proposed other tricks such as token k-nearest neighbour approximation, an approximation to the token similarity matrix and also a probing approach to drop less important tokens.

Cattan et al. [2021] presented an end-to-end model that focus on cross document (CD) coreference resolution. It first pre-trained the mention scorer $s_m(\cdot)$ on the gold mention spans of ECB+ dataset. During the training phase, the pairwise scorer $s_a(i, j)$ compared the mention with all the spans across all the documents and optimized the cross entropy loss of mention-pair scores.

In order to identify paraphrase relations between event mentions and avoid the propagation errors, Zeng et al. [2020] proposed Event-specific Paraphrases and Argument-aware Semantic Embeddings (EPASE). EPASE improved generalization ability in two aspects: recognizing event paraphrases under more situations and incorporating the argument roles into the event mention embedding.

Yadav et al. [2021] proposed a way to solve event and entity coreference resolution jointly under the cross-document coreference resolution. It took the uncertainty of coreference decision into consideration when defining the cost function. The joint coreference model built cluster trees to represent the uncertainty with mentions as its leaves and trained a joint cost function. The core idea of the joint cost function relied on two parts: pairwise mention scorer and relational similarity. Pairwise mention score was calculated via

the model proposed by Cattan et al. [2020]. The mention pair's RoBERTa encoded representation and their element wise product were concatenated and passed through an MLP to get the score. As for relational similarity, it was a weighted average of the similarity score of different arguments of the event mentions. The similarity score of the arguments was calculated based on different properties of the structure of the cluster tree.

Another example of joint learning in coreference resolution is Lu and Ng [2021a] in which the models jointly learned six related tasks: trigger detection, entity coreference, anaphoricity detection, realis detection, argument extraction, and event coreference. The model also used consistency constraints to guide this multi-task learning process. Lu and Ng [2021b] further did an empirical analysis of this model and draw a few interesting findings such as event CR performance could be enhanced by improving mention boundary detection, anaphoricity detection, and subtype detection.

Dobrovolskii [2021] proposed a word-level coreference resolution model wl-coref that focused on individual words rather than spans in order to reduce the complexity of model. It first constructed each word's representation by combining the constituent tokens' contextualized representation. Wl-coref used a bilinear function to get the most possible antecedents for each token. Then for each candidate antecedent, its coreference score was calculated by a feed-forward neural network taking into consideration token embeddings as well as feature information such as distance and speaker. Finally, a feature extraction module was employed to to determine the boundaries of spans based on word-level coreference links.

Beyond the models previously discussed, recent work explores Large Language Models (LLMs) for coreference resolution. Specialized seq2seq approaches using models like T5 (e.g., the transition-based system by Bohnet et al. [2023]) have achieved new state-of-the-art results on standard benchmarks. Concurrently, direct evaluations of general LLMs like GPT and LLAMA (e.g., Gan et al. [2024]) show strong conceptual understanding but face challenges with traditional metrics and fine-grained analysis. Effectively harnessing LLMs for coreference resolution, potentially requiring adapted evaluation methods, remains an active research direction.

2.9 Conclusion

This chapter reviewed the evolution of coreference resolution, from early feature-based models to advanced transformer-based methods. Feature-based approaches, grounded in linguistic features, laid the groundwork for the field. Neural network models, particularly those using MLPs and RNNs, improved contextual understanding and mention clustering. Knowledge-based models further enhanced performance by incorporating external commonsense and domain-specific data. However, the emergence of transformer-based models like BERT and SpanBERT marked a major leap, offering powerful contextual embeddings and achieving state-of-the-art results. Key takeaways from this Transformer era include the significant performance gains realized through sophisticated, context-sensitive span representations and specialized architectural choices like span boundary objectives. These advances have significantly improved accuracy in handling complex referential structures, providing a strong foundation for further research. However, as discussed at the end of Section 2.8, the recent advent of large language models (LLMs) introduces new challenges concerning interpretability, reliable evaluation, and effective task-specific adaptation. Addressing these challenges represents crucial directions for future research, explored further in Chapter 5. For a more extensive discussion on the future trajectory of coreference resolution in the era of Large Language Models, including the potential for hybrid architectures and new evaluation paradigms, see Section 5.2.6.

Table 2.2: The aim, state of the art (SOTA) models and their performances for different coreference resolution datasets. PCR denotes pronoun coreference resolution.

Dataset	Aim	SOTA model	Metrics and result for SOTA
CoNLL 2012 [Pradhan et al., 2012]	Shared task corpus	Wang et al. [2021]	CoNLL score: 87.5%
GAP[Webster et al., 2018]	Gender bias in PCR	Attree [2019]	F1: 92.5%, Gender bias:0.97
KBP 2017[Mitamura et al., 2017]	Within-document Event Coreference	Yu et al. [2020b]	AVG-F: 57.12%
ACE 2005	Within-document Event Coreference	[Lai et al., 2021]	CoNLL score: 87.9%, AVG-F:88.30%
LitBank[Bamman et al., 2020]	Long-distance within-document coreference	Khosla and Rose [2020]	CoNLL score: 80.26%
WSC[Levesque, 2011]	Commonsense knowledge in PCR	Sun et al. [2021]	Accuracy: 97.3%
DPR[Rahman and Ng, 2012]	Complex cases of definite pronouns	Kocijan et al. [2019]	Accuracy: 84.8%
PDP[Davis et al., 2017]	Commonsense knowledge in PCR	Kocijan et al. [2019]	Accuracy: 86.7%
Winogender[Rudinger et al., 2018]	Gender bias in PCR	Kocijan et al. [2019]	Accuracy: 82.1%
WinoBias[Zhao et al., 2018]	Occupational gender bias in PCR	Kocijan et al. [2019]	Subset1 accuracy: 78.0%-78.2%, Subset2 accuracy: 98.7%-99.0%
KnowRef[Emami et al., 2019]	Challenging cases in PCR	Emami et al. [2019]	Accuracy: 71%
WikiCoref[Ghaddar and Langlais, 2016]	Coreference on Wikipedia	Khosla and Rose [2020]	CoNLL score: 71.35%
ECB+[Cybulska and Vossen, 2014]	Cross-document Coreference	Caciularu et al. [2021b]	CoNLL score: 85.6%
RED[O'Gorman et al., 2016]	Within-document Event Coreference	-	-
GUM[Zeldes, 2017]	Shared task corpus	-	-
WEC[Eirew et al., 2021]	Cross-document Event Coreference	Eirew et al. [2021]	CoNLL score: 62.3%
EmailCoref[Dakle et al., 2020]	Entity coreference in email threads	Khosla and Rose [2020]	CoNLL score: 76.17%
BUG[Levy et al., 2021]	Gender bias in PCR	Levy et al. [2021]	Accuracy: 64.1%

Table 2.3: The tasks and learning techniques of different models. DT denotes Decision Tree-based method. MBL denotes memory-based learning. CRF denotes conditional random field. GM denotes gated mechanism. AM denotes attention mechanism. RL denotes reinforcement Learning

	Target Tasks		Rule-based	Traditional Machine Learning			Deep Learning			
Models	Entity	Event	Method	DT	MBL	CRF	GM	RNN	AM	RL
Ding and Liu [2010]	✓			✓						
Atkinson et al. [2015a]	✓				✓					
Durrett and Klein [2014a]	✓					✓				✓
Clark and Manning [2015]	✓									
Wiseman et al. [2016]	✓							✓		
Lee et al. [2017]	✓							✓	✓	
Lee et al. [2018]	✓						✓	✓	✓	
Aralikatte et al. [2019]	✓							✓		✓
Emami et al. [2018]	✓		✓							
Zhang et al. [2019]	✓							✓	✓	
Joshi et al. [2019]	✓								✓	
Joshi et al. [2020]	✓								✓	
Kocijan et al. [2019]	✓									
Ye et al. [2020]	✓									
Wu et al. [2020]	✓								✓	
Khosla and Rose [2020]	✓							✓	✓	
Wang et al. [2021]	✓								✓	✓
Attree [2019]	✓								✓	
Yu et al. [2020b]	✓	✓								
Lai et al. [2021]		✓					✓			
Caciularu et al. [2021b]	✓	✓								
Kirstain et al. [2021]	✓									
Thirukovalluru et al. [2021]	✓									
Cattan et al. [2021]	✓	✓								
Zeng et al. [2020]		✓								
Yadav et al. [2021]	✓	✓								
Dobrovolskii [2021]	✓									

Table 2.4: The features employed by different models. SC denotes semantic consistency, OW denotes opinion words. TC denotes text chunking, WP denotes word position. Pre-trained LM denotes pre-trained language models.

	Semantic Features			Syntactic Features			Word	Pre-trained
Models	SC	OW	NER tags	POS tags	TC	WP	Embedding[5]	LM[6]
Ding and Liu [2010]	✓	✓		✓	✓			
Atkinson et al. [2015a]			✓	✓				
Durrett and Klein [2014a]						✓		
Clark and Manning [2015]			✓	✓		✓		
Wiseman et al. [2016]			✓	✓		✓		
Lee et al. [2017]							G,T	
Lee et al. [2018]							G,E,T	
Aralikatte et al. [2019]							G,E,T	
Emami et al. [2018]								
Zhang et al. [2019]							G,E	
Joshi et al. [2019]								B
Joshi et al. [2020]								S
Kocijan et al. [2019]								B
Ye et al. [2020]								C
Wu et al. [2020]								S
Khosla and Rose [2020]			✓					B
Wang et al. [2021]						✓		B
Attree [2019]								B
Yu et al. [2020b]								R
Lai et al. [2021]								S
Caciularu et al. [2021b]								L
Kirstain et al. [2021]								L
Thirukovalluru et al. [2021]								S
Cattan et al. [2021]								R
Zeng et al. [2020]								B
Yadav et al. [2021]								R
Dobrovolskii [2021]								R

[e]In the word embedding column, G denotes GloVe embedding, E denotes ELMo embedding, T denotes Turian embedding

[f]In the pre-trained language model (LM) column, B denotes BERT, S denotes SpanBERT, C denotes CorefBERT, R denotes RoBERTa, L denotes LongFormer.

Chapter 3

Multi-task Learning in Gold-two-mention style Coreference Resolution

3.1 Introduction

In this chapter, we introduce a novel multi-task learning (MTL) framework tailored for a special type of coreference resolution (CR). Building upon the survey in Chapter 2, this work focuses on jointly optimizing mention identification and linking using a dual-tower transformer-based model.

A recent trend in research has concentrated on challenging instances of CR, where resolution typically necessitates either inferencing from the discourse context or the incorporation of external lexical and commonsense knowledge. This research trajectory was initiated by Levesque et al. [2012], with a particular focus on Terry Winograd's concept of minimal pairs[Winograd, 1972b]. Consider the following examples:

(1) a. My cat only eats canned food because **it** is very selective.
    b. My cat only eats canned food because **it** is highly appetizing.

In these instances, the pronoun it' is linked to different antecedents, despite the presence of only a single varying word in each sentence (i.e., selective' and 'appetizing'). Expanding upon this, Levesque et al. introduced the

[0]The work in this chapter has been published in Liu, R., Chen, G., Mao, R., & Cambria, E. (2023). A Multi-task Learning Model for Gold-two-mention Coreference Resolution. 2023 International Joint Conference on Neural Networks (IJCNN), 1-8.

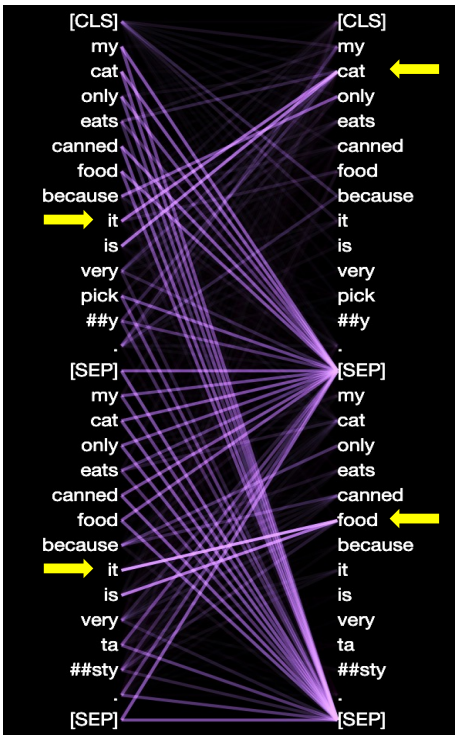


Figure 3.1: Illustration of Transformer attention weight distribution. Higher intensity indicates greater attention weights.

Winograd Schema Challenge (WSC), comprising numerous such minimal pair instances, aimed at evaluating the proficiency of CR models in correctly linking references in these scenarios. Further developments in this field have included expanding the dataset [Rahman and Ng, 2012] and exploring other complex cases related to gender bias [Webster et al., 2018, Rudinger et al., 2018]. CR tasks akin to the WSC, where the precise positions of each reference are provided and models are required to select between two potential antecedents, are often referred to as Gold-two-mention CR (hereinafter, GTM-CR).

To enhance Coreference Resolution (CR) models with extensive world knowledge, GTM-CR frameworks frequently incorporate Pre-trained Language Models (PLMs). This approach is exemplified by Joshi et al. [2019] and Attree [2019] who integrated BERT [Devlin et al., 2019] into their models. Additionally, Kocijan et al. [Kocijan et al., 2019] modified BERT by retraining it on the expansive Wikicrem dataset, a large corpus from Wikipedia with masked mentions.

In this discussion, we contend that while GTM-CR models may not require direct identification of mentions, the capability to differentiate mentions from their surrounding context is beneficial for neural network-based models. This is because such differentiation could facilitate the acquisition of additional positional data pertinent to CR. Such data can encompass aspects like the grammatical function, syntactic parallelism (indicating if the referent and its antecedent occupy analogous syntactic positions), among other factors, which have been identified as crucial for effective CR [Stevenson et al., 1994, Chambers and Smyth, 1998]. In fact, the combined approach of modeling both identification and linking has been applied in traditional CR tasks. For instance, Daumé III and Marcu [2005] approached CR as a search problem by integrating these two tasks within a unified search space. Similarly, Lee et al. [2017] considered mention identification and linking as a singular, end-to-end process, demonstrating the advantages of this joint approach. However, both the search-based approach[Daumé III and Marcu, 2005] and the single-task learning (STL) paradigm [Lee et al., 2017] fail to leverage valuable dependency information acquired from distinct tasks and training objectives.

In this study, we adopt a novel approach within the GTM-CR framework, utilizing Multi-Task Learning (MTL) to concurrently perform mention identification and linking. Specifically, we have developed an MTL-based GTM-CR model that incorporates a shared encoder and distinct task-specific towers built on top of this encoder. The shared encoder is designed to understand general dependency relationships within an input sequence, while the task-specific towers are dedicated to learning the nuances of mention identification and linking. Both the shared encoder and the task-specific towers employ a Transformer [Vaswani et al., 2017] architecture.

The Transformer's architecture, featuring a multi-head attention mechanism and a feed-forward layer, is adept at capturing varied dependency features across different tasks. For instance, as illustrated in Figure 3.1, the multi-head attention in our model shows that learning mention identification enables the attention head corresponding to a referent (like 'it')

to focus on its relevant antecedent ('cat') in Example (1-a), and on 'food' in Example (1-b). This observation underscores the importance of mention identification in enhancing the performance of mention linking, such as linking 'it' with 'cat' in Example (1-a), and 'it' with 'food' in Example (1-b).

To improve the training efficiency of the dual tasks in our MTL model, we introduce a dynamic weight-balancing mechanism. This mechanism adjusts the weight assigned to each task based on the ratio of their respective reduced losses during the training process.

Our model was evaluated on three GTM-CR datasets: GAP [Webster et al., 2018], DPR [Rahman and Ng, 2012], and Winogender [Rudinger et al., 2018]. The results show that our model surpasses state-of-the-art (SOTA) baselines for GAP in terms of both F1 score and Bias, outperforms the SOTA for Winogender, and delivers comparable results to the SOTA for DPR, despite using a significantly smaller fine-tuning dataset. Moreover, when compared to its Single-Task Learning (STL) counterpart, our MTL model demonstrates superior performance across all three datasets. An error analysis further reveals that our MTL model significantly reduces the likelihood of incorrect predictions for feminine pronouns.

The contributions of this chapter are threefold:

- Introduction of an MTL learning paradigm for GTM-CR that outperforms existing methods on multiple datasets;
- Implementation of a dynamic weight balancing mechanism in our coreference resolver, allowing for adaptive balancing between mention identification and linking tasks;
- Comprehensive analyses that explore the impacts of dynamic weight balancing and MTL in GTM-CR, as well as an examination of the types of errors our model more effectively mitigates.

3.2 Prior Studies

3.2.1 Advancements in Coreference Resolution via Pre-trained Language Models

In the realm of Coreference Resolution (CR), research is mainly featured on four methodologies as delineated by Liu et al. [2023]. These include feature-centric approaches[Ding and Liu, 2010, Durrett and Klein, 2014b, Atkinson et al., 2015b], RNN frameworks [Clark and Manning, 2015, Wiseman et al., 2016], knowledge-driven strategies [Emami et al., 2018, Aralikatte et al., 2019], and methods based on Transformer architectures [Joshi et al., 2019, 2020]. The progression in this field typically evolved from feature-based strategies towards deep learning models, including multilayer perceptrons and RNNs, culminating in the recent surge of Transformer-based PLMs. Notably, Liu et al. [2023] emphasize the pre-eminence of Transformer-based models in 18 distinct CR datasets, underscoring their superiority.

A landmark contribution by Joshi et al. [2019] in CR involved the pioneering use of PLMs in CR frameworks. This groundbreaking technique, improving upon Lee et al.'s c2f-coref modelLee et al. [2018], substituted the LSTM-based encoder with BERT Joshi et al. [2019]. The methodology synthesized BERT representations of starting and ending word pieces and the encompassing mention, forging a comprehensive representation for CR, consistent with the c2f-coref architecture. Further advancement was achieved with the introduction of SpanBERT Joshi et al. [2020], which refined the representation of mentions beyond simple token concatenations.

3.2.2 GTM-CR Model Developments

Recent advancements in GTM-CR datasets such as GAP Webster et al. [2018], DPR Rahman and Ng [2012], and Winogender Rudinger et al. [2018] (elaborated in Section 3.4.1) have paved the way for specialized GTM-CR models.

Attree et al.'s leading GTM-CR model for the GAP dataset Attree [2019] incorporates a pronoun-specific BERT module and an evidence pool-

ing component. The former extracts the pronoun's ultimate layer embedding from BERT, while the latter amalgamates clustering inputs from four distinct CR models: AllenNLP Gardner et al. [2017], NeuralCoref, Parallelism+URL Webster et al. [2018], and e2e-coref Lee et al. [2017]. Employing a self-attention mechanism, this module compiles data from these sources into an evidence vector, which, in conjunction with the pronoun's BERT embedding, is utilized for classification via linear and softmax layers.

For the DPR and Winogender datasets, Kocijan et al. [2019] introduced a model that retrained BERT using the WikiCREM dataset, an extensive unsupervised corpus derived from English Wikipedia. This model assigns BERT the task of pronoun resolution, feeding it a sentence with a masked antecedent or pronoun alongside two candidate selections, to predict the more apt choice. Two model variants were developed: *BERT_WIKICREM_DPR* for the DPR dataset and *BERT_WIKICREM_ALL* for both GAP and DPR. The *BERT_WIKICREM_ALL* variant achieved a 84.8% accuracy on DPR, while *BERT_WIKICREM_DPR* attained a 82.1% accuracy on Winogender.

3.2.3 Motivation of the Current Research

While traditional state-of-the-art GTM-CR models Daumé III and Marcu [2005], Lee et al. [2017] have predominantly concentrated on single-task learning, recent perspectives in multi-task learning Mao and Li [2021] suggest that engaging in related yet distinct tasks could enhance overall task efficacy. Our investigation indicates that expertise in mention identification can yield pivotal ancillary data for mention linking, as depicted in Fig. 3.1. Driven by these observations, our research aspires to amalgamate multi-task learning into the GTM-CR arena, contrasting this innovative approach with the conventional single-task learning methodology to assess their respective merits and drawbacks.

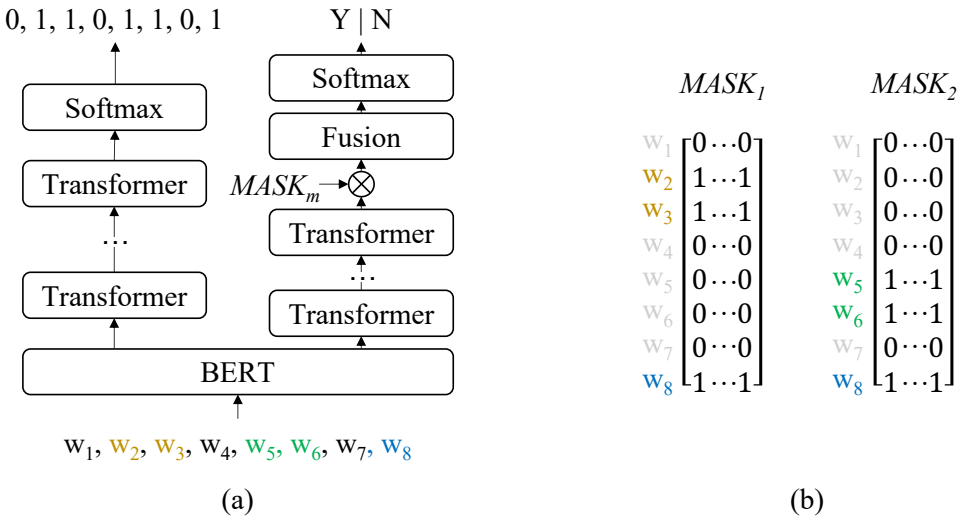


Figure 3.2: (a) The architecture of the multi-task coreference resolution model (Coref-MTL). Colored input tokens indicate mentions and pronouns. $Mask_m$ represents the mask used for deriving antecedent-pronoun pair representations. $\otimes$ symbolizes element-wise multiplication. (b) The configuration of masks for extracting representations for pronoun w_8 and antecedents w_2, w_3 and w_5, w_6.

3.3 Model

The Coref-MTL model adapts a standard multi-task learning paradigm involving a shared encoder and task-specific towers. The use of a BERT encoder and Transformer layers are established techniques. Our novel contributions within this framework are twofold: (1) the specific formulation of the GTM-CR problem into two concurrent tasks for MTL—mention identification and mention linking—which has not been previously explored in this context; and (2) the development of the Dynamic Weight Balancing algorithm (Algorithm 1, Section 3.3.2), which adaptively tunes the focus between these two tasks during training. The equations describing the standard Transformer forward pass (Eq. 3.1) and softmax outputs (Eq. 3.2, 3.7) are standard, while the overall loss formulation (Eq. 3.9) and the logic of our dynamic weighting scheme are integral to our contribution.

3.3.1 GTM-CR Model for both Mention Identification and Linking

Figure 3.2 displays the structure of our proposed Coref-MTL model that concurrently optimizes mention identification and linking tasks. Initially, the input sentence is tokenized and processed through a Pre-trained Language Model (PLM), after which the resulting outputs are fed into two separate task-specific modules. The following subsections detail each module.

3.3.1.1 Mention Identification Module

In this module, the PLM-encoded input representation is used to determine if each token is part of a mention, which may be an antecedent or referent and can span multiple tokens. The Mention Identification (MI) task-specific module processes the PLM outputs through l Transformer encoder layers. Assuming $X_0 \in \mathbb{R}^{s \times e}$ as the contextualized input sentence representation, where s and e stand for sequence length and embedding size respectively, the representation after the ith Transformer encoder, TransEnc$i^{MI}(\cdot)$ ($i \in 1, ..., l$), is expressed as:

$$Xi^{MI} = \text{ TransEnc } i^{MI}\left(X_{i-1}^{MI}\right). \tag{3.1}$$

Then, the hidden states from the final Transformer layer (X_l^{MI}) undergo a linear transformation and are processed through a softmax layer to produce a probability distribution for each token's inclusion in a mention:

$$P^{MI} = \text{softmax}\left(W_1^T X_l^{MI} + b_1\right), \tag{3.2}$$

where $P^{MI} \in \mathbb{R}^{2 \times s}$ represents the probability distribution for all input tokens in the mention identification task. The learnable parameters include $W_1 \in \mathbb{R}^{e \times 2}$ and $b_1 \in \mathbb{R}^{s \times 2}$. The Cross Entropy loss L_{MI} for this task is calculated as:

$$L_{MI} = CrossEntropy\left(\hat{Y}^{MI}, Y^{MI}\right), \tag{3.3}$$

with $\hat{Y}^{MI}$ and Y^{MI} being the predicted and actual labels for mention identification, respectively.

3.3.1.2 Mention Linking

In a procedure analogous to mention identification, the Mention Linking (ML) task commences with the intricate processing of a tokenized sentence through a sequence of Transformer layers, as depicted in Fig. 3.2. This initial step is crucial, as it converts the input into a refined vectorized format specifically tailored for the ML task, thus leveraging the advanced computational capabilities of the Transformer architecture. Central to the ML task's architecture is the deployment of k Transformer encoders, strategically arranged in a sequential, stacked configuration. This arrangement is instrumental in enhancing the depth and quality of sentence representation. The representation yielded from the jth Transformer encoder, denoted as TransEnc$j^{ML}(\cdot)$, where j is a member of the set $1, ..., k$, is symbolized as Xj^{ML} within the space $\mathbb{R}^{s \times e}$. The transformation process is elegantly encapsulated by the following equation:

$$X_j^{ML} = \text{TransEnc } j^{ML}\left(X_{j-1}^{ML}\right). \tag{3.4}$$

A novel aspect of the ML task is the derivation of vector representations for mentions. This is achieved through a highly specialized masking mechanism applied to the output of the final Transformer layer. The mask, as illustrated in Fig. 3.2 (b), is a meticulously crafted binary sequence, correlating with each token in the input sentence. This mask enables precise element-wise multiplication with token vectors, thereby enabling focused extraction and retention of relevant information. The representations of non-mention tokens are effectively masked out, ensuring the preservation and emphasis of vectors pertinent to mentions and pronouns.

In practical applications, a sentence may comprise multiple mention spans, each requiring a distinct mask. These masks are instrumental in retaining the vector representations of specific antecedent-pronoun pairs, underscoring the adaptability of the ML process. The representation post-masking, denoted as V in the space $\mathbb{R}^{s \times e}$, is computed through a methodical element-wise multiplication as shown in the following equation:

$$V = Mask_m \otimes X_l^{ML}, \tag{3.5}$$

where $\otimes$ represents the element-wise multiplication operation.

Subsequent to the masking process, attention is directed towards the representations of candidate mentions. These representations are derived by averaging over all unmasked constituent tokens, a methodological decision that caters to antecedents composed of multi-word expressions. The representation vectors for the candidate antecedent and the pronoun, denoted as $\mathbf{v}_c$ and $\mathbf{v}_p$ respectively and both residing in $\mathbb{R}^e$, undergo a fusion operation, culminating in $\mathbf{v}_f$ within the same dimensional space. This fusion process, after an extensive exploration of various techniques, was optimized using element-wise products, as detailed in the following equation:

$$\mathbf{v}_f = \mathbf{v}_c \otimes \mathbf{v}_p \in \mathbb{R}^e. \tag{3.6}$$

The fused vector $\mathbf{v}_f$ then undergoes processing through a softmax layer, a critical step in translating the vector into a probabilistic distribution over two classes. This distribution serves as an indicator of the presence or absence of a coreference link:

$$p^{ML} = \text{softmax}\left(W_2^T \mathbf{v}_f + b_2\right) \in \mathbb{R}^2, \tag{3.7}$$

where p^{ML} encapsulates the probability distribution, and W_2 and b_2 are the adjustable parameters. The training of the model incorporates the use of cross-entropy loss:

$$LML = CrossEntropy(\hat{y}^{ML}, y^{ML}), \tag{3.8}$$

where $\hat{y}^{ML}$ represents the predicted label, and y^{ML} signifies the actual label.

The ultimate loss for coreference resolution, L_{Coref}, is calculated as a weighted amalgamation of the losses from the mention identification (L_{MI}) and mention linking (L_{ML}) tasks:

$$L_{Coref} = w_{MI} L_{MI} + w_{ML} L_{ML}, \tag{3.9}$$

with w_{MI} and w_{ML} as the respective loss weights for these subtasks.

3.3.2 Dynamic Weight Balancing

In the advanced computational framework of GTM-CR, which cohesively amalgamates the tasks of mention identification and mention linking, the judicious allocation of task-specific weights emerges as a pivotal aspect of the training regimen. This nuanced approach is necessitated by the inherent disparities in the complexity and learning trajectories associated with each task. To adeptly navigate these disparities, we have instituted a refined dynamic weight-balancing scheme. This scheme is anchored in the principle of loss adaptation, a concept that is both innovative and critical for optimizing task performance.

The operational essence of this methodology lies in its iterative and responsive nature. At the conclusion of each training epoch, a detailed recording of the task-specific losses is undertaken. These recorded losses are then meticulously compared with the initial losses observed after the first epoch, serving as a benchmark for assessing progress and adjusting weights. The innovative aspect of our method is the strategic adjustment of weights in the subsequent epochs, which is inversely proportional to the degree of loss reduction observed in each task. Concretely, tasks that demonstrate a more pronounced reduction in loss are assigned proportionally lower weights, while those with lesser loss reduction are accorded higher weights in the next phase of training.

This weight adjustment is executed through a calculated formulation, employing the square of the relative loss percentage of each task as its basis. This calculation is then complemented by a softmax normalization process. The softmax function plays a crucial role in this context, as it ensures that the cumulative weight distribution across the different tasks adheres to the constraint of summing to unity. This approach not only imbues the training process with a high degree of adaptability but also aligns the resource allocation with the evolving needs and challenges of each task.

A comprehensive and systematic presentation of this dynamic weight-balancing algorithm is delineated in Algorithm 1, where the algorithmic steps are elaborated in a structured manner. This algorithm stands as a

testament to the meticulous and adaptive approach employed in the GTM-CR framework, ensuring that each task receives a tailored and dynamic weight allocation, reflective of its individual learning curve and complexity.

Algorithm 1 Dynamic Weight Balancing Algorithm based on Loss Metrics

Establish the count of distinct tasks, denoted as T.
Initial task weights are set uniformly at $1/T$.
for each epoch e **do**
  for each batch b within the epoch **do**
    Aggregate the batch loss to compute total epoch loss
    Formulate the epoch's weighted loss $\ell_{(e,b)} = \sum_{i=1}^{T} \ell_{(e,b,t)} \times w_t$
    Perform weight optimization for W with respect to $\ell_{(e,b)}$.
  end for
  Calculate the epoch's cumulative loss for each task, expressed as $\ell_e \in \mathbb{R}^T$.
  Reference the loss from the initial epoch, denoted as $\ell_0 \in \mathbb{R}^T$.
  for each task t **do**
    Adjust task t's weight, defined as $w_t = \left(\frac{\ell_{(e,t)}}{\ell_{(0,t)}}\right)^2$.
  end for
  Normalize task weights to achieve a total sum of one:

$$w_t = \frac{e^{w_t}}{\sum_{i=1}^{T} e^{w_i}}$$

end for

3.4 Experiments

3.4.1 Datasets and Evaluation Protocols

For our investigation, we meticulously selected three datasets pertinent to GTM-CR: GAP, DPR, and Winogender. The comprehensive statistics for each dataset are systematically presented in Table 3.1.

GAP The GAP dataset Webster et al. [2018], sourced from Wikipedia, is crafted to echo the real-world intricacies of pronoun coreference resolution. It encompasses 8,908 pronoun and candidate mention pairs, divided into test (4000 pairs), development (4000 pairs), and validation (908 pairs) subsets.

Each subset maintains an equilibrium of instances with masculine pronouns (e.g., him, his) and feminine pronouns (e.g., she, her).

For the GAP dataset, our evaluation metrics include F1 scores and a Bias score. F1 scores are computed for three categories: instances with masculine pronouns, those with feminine pronouns, and an aggregated set including all pronoun types. The Bias score is calculated as the ratio of feminine F1 to masculine F1 scores.

DPR The Definite Pronoun Resolution (DPR) corpus Rahman and Ng [2012], an extension of the WSC minimal pairs (referenced in Section 3.1), spans a wide array of themes, encompassing real-world events, cinematic narratives, and entirely fictitious scenarios, predominantly reflecting the pop culture familiar to American children of the early 1990s. DPR includes instances both requiring and independent of commonsense reasoning and incorporates cases where the key element is a phrase. This dataset contains 1,322 training examples and 564 test examples, amounting to a total of 1,886 sentences.

Winogender The Winogender dataset Rudinger et al. [2018] serves as an instrumental tool for examining gender bias within GTM-CR, employing the framework of the Winograd Schema Challenge (WSC). It consists of sentences each featuring an occupational noun linked to a pronoun, which can be "he," "she," or "they." These occupational nouns often carry implicit

Table 3.1: Dataset Statistics Overview. 'Val.' denotes the validation set, 'P-M' represents the count of pronoun-mention pairs, and 'Avglen' indicates the average length of examples in words. Notably, each example includes two mentions and one pronoun, thus doubling the count of P-M pairs in relation to the total number of examples.

Dataset	Train	Val.	Test	Total	P-M	Avglen
GAP	4000	908	4000	8908	17816	71.57
DPR	1322	-	564	1886	3772	14.27
Winogender	-	-	720	720	1440	14.49

gender connotations. For instance, the sentence "the secretary asked the visitor to sign in so that he could update the guest log" Rudinger et al. [2018] could pose a challenge to a coreference resolution classifier that harbors gender biases, potentially misaligning "he" with "secretary." The dataset's primary objective is to scrutinize the influence of pronoun gender alteration on the accuracy of a model. Winogender, comprising a total of 720 sentences, is exclusively utilized for testing purposes.

3.4.2 Baseline Model Selection

For the foundational comparison in our study, we strategically selected an array of state-of-the-art models in GTM-CR as baselines, ensuring a rigorous and comprehensive evaluation framework. These models are at the forefront of their respective domains, demonstrating cutting-edge performance and robustness in GTM-CR tasks.

Specifically, our selection included ProBERT Attree [2019] as the benchmark model for the GAP dataset. ProBERT is renowned for its tailored optimization and superior performance in handling gendered pronoun resolution challenges within the GAP context.

For the DPR and Winogender datasets, we employed variants of BERT fine-tuned on the WikiCREM dataset Kocijan et al. [2019], referred to as $BERT\_WIKICREM$. Recognizing the nuanced differences in these datasets, we explored two distinct versions of this fine-tuned model: $BERT\_WIKICREM\_ALL$ and $BERT\_WIKICREM\_DPR$. Each variant represents a unique adaptation of the BERT architecture, optimized to address the specific characteristics and challenges inherent in the DPR and Winogender datasets.

A detailed exposition of the architectural intricacies, optimization strategies, and performance metrics of each baseline model can be found in Section 3.2. This comprehensive analysis offers a deeper understanding of the models' capabilities and their alignment with the objectives of our study, thereby establishing a solid foundation for subsequent comparative analysis and discussion.

3.4.3 Experimental Setup and Configuration

For the training of all models in this study, we utilized the NVIDIA RTX3080Ti GPU, equipped with a 16GB memory capacity, ensuring efficient and robust computational performance. The language model components were meticulously developed using the renowned *Transformers* library from Hugging Face Wolf et al. [2020], with initial configurations established through pre-trained checkpoints to leverage existing advancements in the field.

The optimization of the models was conducted using the Adam optimizer Kingma and Ba [2014], a choice motivated by its effectiveness in handling large-scale data and complex model architectures. The optimizer was configured with hyperparameters $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e^{-8}$, accompanied by a warm-up learning rate of $2e^{-6}$, to facilitate a gradual and controlled learning process.

In terms of architecture, the mention identification specific tower was designed with 2 Transformer encoder layers ($l = 2$), while the mention linking specific tower incorporated 4 Transformer encoder layers ($k = 4$). This architectural distinction reflects the differing complexities and requirements of the two tasks. Regularization was uniformly applied across all layers, employing a fixed dropout Srivastava et al. [2014] rate of 0.2 to mitigate overfitting.

Each Transformer encoder was configured with an embedding size e of 1024 and 16 heads, parameters selected to optimize the balance between model complexity and computational efficiency. The batch size was set to 32 during training, a size determined to be optimal for balancing computational load and learning dynamics.

Performance evaluation of each model followed a rigorous protocol. Models were trained on the DPR training set for evaluations involving DPR and Winogender, and on the GAP training set for assessments related to GAP. Model checkpoint selection was guided by performance on the validation dataset, where available, or otherwise based on the test dataset. Training was designed to cease if the targeted metrics (F1 or accuracy) did not exhibit

improvement over a span of 60 consecutive epochs, ensuring an efficient and objective-oriented training process.

To ensure the robustness of our results and to account for the stochastic nature of Transformer-based models, all experiments for our proposed models, baselines, and ablation variations presented in Section 3.5 were conducted 10 times, each with a different random seed. The results reported in the main comparison (Table 3.3) and the ablation study (Table 3.4) represent the mean (μ) and standard deviation (σ) of these 10 runs. To validate our claims of improvement, we perform a one-tailed paired-samples t-test on the distributions of results, with the standard significance threshold of $p < 0.05$.

3.5 Results

This section is dedicated to a meticulous comparative analysis of our fully developed model against the prevailing state-of-the-art (SOTA) frameworks, specifically within the contexts of GAP, DPR, and Winogender datasets. This comparison is followed by an in-depth exploration of the model's hyperparameters. Subsequently, we have integrated an ablation study aimed at discerning the performance dynamics of our model in the absence of the Dynamic Weight Balancing (DWB) and Multi-Task Learning (MTL) mechanisms. Additionally, an extensive error analysis is undertaken to explicate the mechanisms through which our multi-task learning model enhances its efficacy on GTM-CR datasets. It should be noted that while formal statistical significance tests were not performed, the consistency of the results across multiple datasets and evaluation metrics provides strong evidence for the effectiveness of our proposed model.

3.5.1 Overall Outcomes and Performance Metrics

Results on GAP Table 3.2 presents a detailed tabulation of our model's performance in contrast to various baseline models within the GAP dataset framework. Remarkably, our Coref-MTL model exhibits a pronounced superiority over all compared baseline models, particularly in terms of overall F1

Table 3.2: Evaluation Results on the GAP Test Dataset, Evaluating F1 Scores and Bias.

Model	Overall	Masculine	Feminine	Bias
ProBERTAttree [2019]	89.70	90.80	88.60	98.00
GREP Attree [2019]	92.50	**94.00**	91.10	97.00
Coref-MTL (Ours)	**92.72**	92.65	**92.45**	**99.76**

scores and Bias metrics. This notable enhancement is primarily ascribable to the model's enhanced capability in accurately linking feminine pronouns to their respective referents. A increment of 1.3% in the feminine F1 score relative to the GREP model is observed, which substantially contributes to the reduction of overall bias in pronoun resolution.

Results on DPR and Winogender Table 3.3 presents the outcomes of our Coref-MTL model and baseline models on the DPR test set and the Winogender dataset. Our Coref-MTL model slightly lags behind the state-of-the-art BERT WIKICREM ALL model in DPR performance but surpasses the BERT_WIKICREM_DPR model. This slight shortfall is attributed to BERT_WIKICREM_ALL's re-training on diverse corpora, including WIKICREM, GAP, and DPR, specifically tailored for Coreference Resolution (CR). The voluminous WIKICREM corpus, encompassing approximately 2.4 million samples, provides a substantial advantage. In contrast, our Coref-MTL model is fine-tuned exclusively on the DPR dataset, which represents a mere 0.05% of the WIKICREM sample size.

Nevertheless, the efficacy of our model is underscored by the new statistical analysis. As shown in Table 3.3, our Coref-MTL model's mean accuracy of 84.61% (±0.18) on the DPR dataset, while not statistically superior to the BERT_WIKICREM_ALL baseline ($p = 0.092$), is a **statistically significant** improvement ($p < 0.001$) over the BERT_WIKICREM_DPR baseline, which was trained on comparable data.

Crucially, on the Winogender dataset, our model's mean accuracy of 83.09% (±0.17) represents a **statistically significant** improvement over

the SOTA baseline (BERT_WIKICREM_DPR, 82.10% ± 0.20, $p = 0.021$). This confirms the consistency of our model's advance.

An interesting observation from Table 3.3 is the performance difference between the two baseline models on the Winogender dataset. The BERT_WIKICREM_ALL model, despite being trained on a larger variety of data including the GAP dataset, performs considerably worse on Winogender (76.70% accuracy) than the BERT_WIKICREM_DPR model (82.10% accuracy). This suggests that exposure to the specific gender bias patterns in the GAP dataset may have caused the _ALL model to develop certain biases that did not generalize well to the occupational gender stereotypes present in Winogender. This highlights the sensitivity of models to their training distributions and underscores the challenge of creating truly generalizable debiasing methods.

Table 3.3: Performance on the DPR and Winogender datasets. Results are the mean (μ) ± standard deviation (σ) over 10 runs. P-values compare Coref-MTL to the strongest baseline for each dataset.

Model	**DPR acc.** ($\mu \pm \sigma$)	**Winogender acc.** ($\mu \pm \sigma$)
BERT_WIKICREM_ALL	84.80 ± 0.15	76.70 ± 0.24
BERT_WIKICREM_DPR	80.00 ± 0.22	82.10 ± 0.20
Coref-MTL (Ours)	**84.61 ± 0.18**	**83.09 ± 0.17**
p-value (vs. ALL)	$p = 0.092$	$p < 0.001$
p-value (vs. DPR)	$p < 0.001$	$p = 0.021$

Corroborating the GAP results, our model's exceptional performance on the Winogender dataset, which also targets gender bias in CR, is noteworthy. The success on both GAP and Winogender datasets is a testament to the efficacy of our approach in better modeling antecedent locations through joint mention identification, effectively eliminating spurious correlations in coreference resolver training. Furthermore, the comparison between the two baseline models highlights, once again, the advantages of Multi-Task Learning (MTL) in terms of time and energy efficiency.

Table 3.4: Ablation study demonstrating the impact of removing dynamic weight balancing (DWB) and multi-task learning (MTL). Results are the mean (μ) ± standard deviation (σ) over 10 runs. P-values from one-tailed paired t-tests compare the full model to each ablated version.

Model	**GAP F1** ($\mu \pm \sigma$)	**GAP Bias** ($\mu \pm \sigma$)	**DPR acc.** ($\mu \pm \sigma$)	**Winogender acc.** ($\mu \pm \sigma$)
Coref-MTL	**92.74 ± 0.15**	**99.75 ± 0.03**	**84.61 ± 0.18**	**83.09 ± 0.17**
-DWB	92.38 ± 0.14	99.73 ± 0.04	84.33 ± 0.20	80.37 ± 0.21
p-value	$p = 0.002$	$p = 0.045$	$p = 0.042$	$p < 0.001$
-MTL (Coref-STL)	92.24 ± 0.16	99.22 ± 0.09	83.62 ± 0.19	80.23 ± 0.20
p-value	$p < 0.001$	$p < 0.001$	$p = 0.015$	$p < 0.001$

3.5.2 Ablation Study

This section presents a detailed ablation study to elucidate the individual contributions of each component in our proposed model. The study involves systematically omitting certain components, namely Dynamic Weight Balancing (DWB) and Multi-Task Learning (MTL), and observing the resultant impact on model performance. Specifically, the removal of DWB alone, as well as the combined removal of both DWB and MTL, resulting in a single-task learning model (hereafter referred to as Coref-STL), are investigated.

It is evident from the results in Table 3.4 that the exclusion of any component detrimentally affects the model's efficacy. For instance, on the GAP dataset, the removal of DWB leads to a statistically significant decrease in the mean F1 score (from 92.74% to 92.38%, $p = 0.002$). This trend is consistently observed across the other datasets as well, underscoring the critical and statistically validated roles played by both MTL and an effective task balancing strategy in Coreference Resolution (CR).

A particular focus on the results from the Winogender dataset reveals a greater benefit from the incorporation of MTL and DWB compared to the other datasets. The mean performance drops from 83.09% to 80.37% without DWB and to 80.23% without MTL. Both of these performance drops are statistically significant ($p < 0.001$). This substantial and consistent performance degradation upon removing our proposed components provides statistical validation for the hypothesis that the MTL framework enhances the model's generalizability.

3.5.3 Analyzing the Synergistic Impact of Multi-Task Learning and Dynamic Weight Balancing

This section delves into the intricate dynamics between Multi-Task Learning (MTL) and Dynamic Weight Balancing (DWB), and their collective impact on model training and performance. Our analysis is anchored on the observation of task-specific loss evolution and weight adjustments over the course of training, as depicted in Figure 3.3.

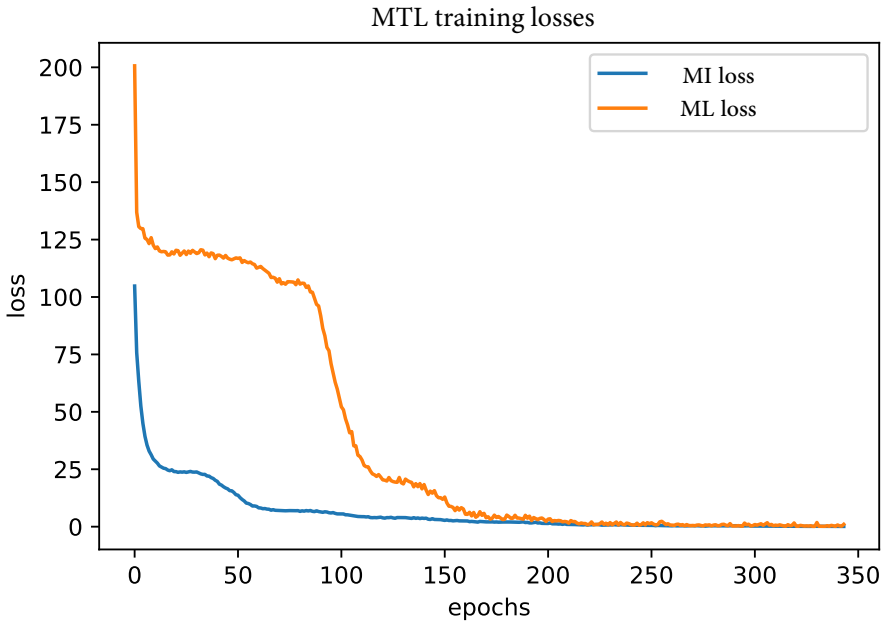
At the onset of training, a notable rapid decline in the loss of both tasks is observed. Interestingly, the loss associated with mention identification begins at a relatively lower magnitude, which leads to its swift approach towards negligible values. This initial phase is characterized by a subsequent decrease in the weight allocated to mention identification and a corresponding increase in the weight for mention linking. This adaptive weighting mechanism effectively shifts the training emphasis towards the mention linking task.

As training progresses, an intriguing interplay between the two tasks emerges. The proficient performance in mention identification fosters enhanced and expedited learning in mention linking. This synergy is evident as the mention linking loss decreases, while there is a subtle fluctuation in the mention identification loss, reflecting the dynamic adjustments in task prioritization.

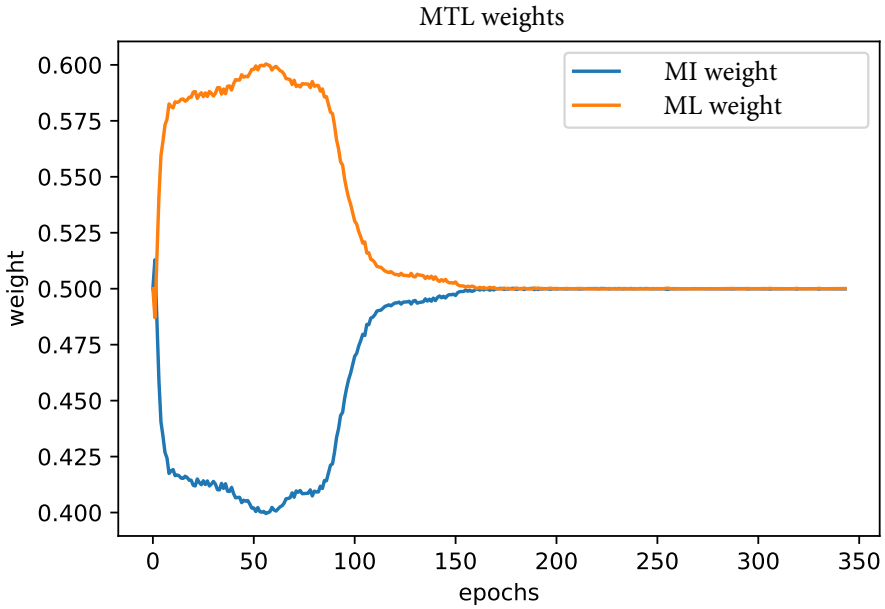
The culmination of this process is observed when the weights for both tasks reach a balanced and stable state. In this equilibrium, the tasks are trained in a collaborative manner, leading to a mutual enhancement of performance, particularly in mention linking. This balanced training approach underscores the efficacy of the integrated MTL and DWB framework in achieving optimal learning outcomes.

3.5.4 Error Analysis

This section aims to provide a nuanced understanding of the performance characteristics of our Coref-MTL model, particularly in comparison to the



(a)



(b)

Figure 3.3: (a) Graphically represents the changes in training losses for the Coref-MTL model on the DPR training set. (b) Depicts the evolution of task weights during the training of Coref-MTL, where MI signifies the mention identification task and ML denotes the mention linking task.

Coref-STL model, through a detailed error analysis conducted on the Winogender dataset. The ensuing discussion outlines key observations derived from this analysis.

First, we embarked on quantifying errors associated with pronouns of different gender categories: masculine (e.g., 'he', 'his', 'him'), feminine (e.g., 'she', 'her'), and gender-neutral (e.g., 'they', 'them', 'their'). The frequency of these errors for each model is systematically tabulated in Table **??**. Corroborating the findings from our previous evaluations (refer to Section 3.5.1) and the ablation study (refer to Section 3.5.2), it is evident that, in contrast to Coref-STL, the Coref-MTL model reduces errors involving feminine pronouns. Notably, errors in Coref-MTL are almost evenly distributed across the three pronoun categories.

Furthermore, the issue of agreement mismatch, particularly when using gender-neutral pronouns for reducing gender bias, has been a point of discussion in the field Poesio et al. [2023]. Consider the following example from the Winogender dataset:

(2) The clerk provided someone with paperwork to return to **them** upon completion.

Here, the plural pronoun 'them' refers to a singular noun phrase. Despite this apparent syntactic mismatch, both Coref-MTL and Coref-STL models demonstrate competence in these cases. This suggests that neural coreference resolution (CR) models may not exclusively rely on traditional semantic constraints for resolving coreferences.

The GTM-CR dataset presents a unique scenario wherein each text undergoes dual evaluations by the models: once for the connection between the pronoun and the first candidate antecedent, and again for a link with an alternative antecedent. Upon scrutinizing the distribution of incorrect predictions by the two models, we observed a propensity for Coref-STL to more frequently make incorrect predictions in both evaluations, compared to Coref-MTL. Moreover, Coref-STL tends to erroneously predict the absence of a coreference relation more often than Coref-MTL, which exhibits

a more uniform distribution of incorrect predictions. This pattern may indicate that, in certain instances, Coref-STL resorts to random prediction strategies, perhaps indicative of underlying limitations in its processing capabilities.

The marked improvement on feminine pronouns may stem from how the two models learn. The Coref-STL model, which only learns the linking task, may be more susceptible to learning spurious correlations from the training data, where masculine pronouns might be statistically more common or linked with certain syntactic roles. In contrast, the Coref-MTL model is also tasked with mention identification. This forces the model to develop a more robust, context-aware representation of what constitutes an entity, independent of its eventual role as an antecedent. This foundational understanding likely makes the model less reliant on superficial gender-based statistical patterns, leading to more balanced performance.

To provide more granular, qualitative evidence of these performance differences, we examine how the models handle challenging, anti-stereotypical examples from the Winogender dataset. These examples are specifically constructed to test a model's reliance on societal gender biases, where an occupation stereotypically associated with one gender is paired with a pronoun of the opposite gender.

A systematic pattern of failure emerges in the Coref-STL model when faced with these scenarios. Consider the first example, which pairs a stereotypically masculine occupation with a feminine pronoun:

(3) The mechanic told the customer that **she** had to wait for the part to arrive.

Here, the Coref-STL model makes a stereotype-driven error, incorrectly resolving **she** to 'the customer'. This suggests the model relies on a shallow statistical association between "mechanic" and masculinity, thus avoiding the correct link. In contrast, the Coref-MTL model correctly resolves **she** to 'The mechanic', demonstrating its ability to prioritize the immediate discourse structure over learned occupational stereotypes.

This pattern is further confirmed in cases involving stereotypically feminine occupations, as shown in this second example from the dataset:

(4) The secretary asked the visitor to sign in so that **he** could update the guest log.

Here again, the Coref-STL model makes a stereotype-driven error. It incorrectly links **he** to ‘the visitor‘, avoiding the stereotypically feminine occupation "secretary" despite the clear functional logic of the sentence. The Coref-MTL model, however, correctly identifies ‘the secretary‘ as the antecedent because its multi-task training has given it a more robust understanding of the entity’s role. This consistent success provides qualitative support for the hypothesis that the multi-task learning framework fosters more robust and abstract linguistic representations. By being simultaneously trained to identify mentions as distinct entities, the Coref-MTL model is compelled to focus more on the grammatical and structural roles of those entities. This foundational understanding appears to override the superficial, stereotype-based statistical patterns learned by the single-task model, leading to more accurate and less biased resolutions.

3.6 Conclusion

This chapter has endeavored to explore and enhance the domain of coreference resolution, a pivotal component in the realms of natural language understanding and semantic cognition. Specifically, we have focused on the Gold-two-mention Coreference Resolution (GTM-CR), a sophisticated variant of coreference resolution. While prevailing state-of-the-art models predominantly concentrate on the mention-linking aspect of GTM-CR, leveraging pre-identified mentions, our research presents a novel perspective. We assert that the task of mention identification plays a vital role in the construction of an effective GTM coreference resolver.

In pursuit of this, we have introduced a novel multi-task learning model designed to concurrently train on both mention identification and mention linking tasks. This approach is predicated on the premise that mentions

are initially unknown during the mention identification phase, thereby compelling the model to discern them autonomously. Such an integrated framework allows for the leveraging of potentially complementary and interdependent information inherent in these two related tasks.

A distinctive feature of our model is the dynamic adjustment of task weights throughout the training process, a mechanism that finely tunes the focus between mention identification and linking as the model evolves. This innovative approach has culminated in the model achieving unprecedented state-of-the-art performance on two GTM-style datasets, namely GAP and Winogender. Additionally, it has demonstrated competitive results on another dataset, the DPR, without the necessity for extensive fine-tuning on large-scale external corpora.

The implications of these findings are substantial, suggesting that a more holistic approach to coreference resolution, which encompasses both mention identification and linking, can enhance the model's understanding and processing of natural language. This advancement opens new avenues for research and application in semantic cognition and computational linguistics, potentially leading to more nuanced and accurate natural language processing systems.

Chapter 4

Financial Sentiment Analysis with Coreference Resolution

4.1 Introduction

In the rapidly advancing financial industry, the deluge of data ranging from intricate transactional records to the vast sprawl of news articles and social media discourse has introduced significant challenges and opportunities. The critical task of extracting actionable insights from such voluminous and complex data sets has become increasingly indispensable for stakeholders in the financial ecosystem [Gupta et al., 2020, Xing et al., 2018]. Sentiment analysis, particularly, stands out as a vital analytical tool in decoding the multifaceted narratives within financial markets, enabling a deeper comprehension of market sentiments and investor behaviors. Using diverse data sources, including financial reports, news feeds, and social media platforms, to gauge the prevailing market mood, thereby assisting investors, financial institutions, and regulators in making well-informed decisions. This analytical approach not only facilitates the prediction of market trends, but also aids in monitoring potential market manipulations or fraudulent activities, underpinning a more transparent and efficient financial environment.

The field of financial sentiment analysis has seen a remarkable transformation with the introduction of advanced NLP and deep learning techniques. Initially grounded in lexicon-based methods [Loughran and McDonald, 2011], the advent of machine learning models marked the beginning of a

new era [Kearney and Liu, 2014]. The development of pre-trained language models, such as BERT [Devlin et al., 2019] and its subsequent adaptations, has significantly improved the accuracy of sentiment analysis [Yang et al., 2019]. These models, trained on vast amounts of text, have the ability to understand the context and nuances of language, offering insights that were previously unattainable with traditional methods [Gupta et al., 2020]. Integration with neural network architectures such as LSTM to enhance predictive capabilities, allowing for more precise forecasts of market trends was also explored [Kohsasih et al., 2022]. This synergy between domain-specific adaptations [Liu et al., 2020] and advanced neural networks underscores a significant leap forward, providing a deeper and more nuanced understanding of financial texts, which is crucial for making informed decisions in the financial sector [Sahu et al., 2023].

The primary application of financial sentiment analysis, and the central problem setting for this chapter, lies in its use within quantitative financial modeling. Financial analysts and investment firms seek to identify predictive "signals" from diverse data sources to forecast market movements, such as changes in stock prices or trading volumes. A time series of sentiment scores derived from public news about a specific company represents a powerful potential signal. For example, a consistent stream of positive news about a company might precede a rise in its stock price. This process typically involves (1) Signal Generation, where textual data is converted into a numerical sentiment score for each company over a given period (e.g., daily), and (2) Model Integration, where this sentiment score is used as an independent variable in an econometric model to test its statistical relationship with a dependent financial variable (e.g., next-day stock performance).

However, a fundamental challenge in this pipeline is the "noise" and ambiguity inherent in financial news. A single news article often discusses multiple entities (e.g., companies, executives, products), and the sentiment expressed towards each can vary significantly. Consider this hypothetical news excerpt:

"Ford reported a 5% rise in Q2 sales, beating analyst expectations and showcasing the strength of its F-Series lineup. This performance stood in

stark contrast to its rival, General Motors, which announced an unexpected 2% sales dip. The company cited ongoing tariff challenges as a primary reason for the poor results."

This problem of correct attribution is where Coreference Resolution (CR) becomes critical. Without CR, a sentiment analysis tool might struggle to determine which entity "The company" refers to. It could incorrectly associate the negative sentiment from "poor results" with Ford, the main subject of the first sentence. CR is the mechanism that correctly links "The company" to its immediate antecedent, "General Motors." This ensures the positive sentiment from the first sentence is attributed solely to Ford, and the negative sentiment from the second is correctly assigned to GM, thus creating two distinct and accurate sentiment signals from a single piece of text. Therefore, CR is not just an incremental NLP enhancement; it is a foundational step for improving the purity and predictive power of sentiment-based financial signals.

This chapter explores the application of coreference resolution (CR) in financial sentiment analysis. We integrate CR as a preprocessing step to improve entity-level sentiment attribution in financial texts such as earnings reports and news headlines. By resolving pronouns and ambiguous references before sentiment scoring, we aim to enhance the accuracy and reliability of financial signal extraction.

A key methodological distinction of this chapter is the adoption of a state-of-the-art sequence-to-sequence (seq2seq) model for the coreference resolution task, departing from the encoder-based Multi-Task Learning (MTL) framework developed in Chapter 3. This choice is necessitated by the different nature of the CR problem addressed here. The GTM-CR task in Chapter 3 is a discriminative task focused on selecting between two given candidates. In contrast, the analysis of financial texts requires a more comprehensive, generative approach to first identify all potential entity mentions within a document and then link them into coreference chains. The seq2seq transition-based system is the state-of-the-art for this open-ended, end-to-end CR task and is thus the appropriate choice for this application.

Furthermore, our downstream analysis employs robust econometric and machine learning models, such as PSM-DID and Random Forest, which are standard in empirical finance for their ability to facilitate causal inference and provide interpretable results.

The main contributions of this work are as follows:

1. **Coreference-enhanced sentiment attribution:** We used coreference resolution as a preprocessing step to improve sentiment attribution by accurately linking pronouns and mentions to specific financial entities within complex texts.

2. **Innovative application of PSM-DID in sentiment analysis:** Our research extends the conventional application of PSM-DID methodology to financial sentiment analysis, incorporating baseline regression and fixed effects models to examine the causal impact of CR-informed financial sentiment on financial metrics, thereby highlighting the versatility and potency of this method in uncovering nuanced insights from financial narratives.

3. **Mediation analysis for sentiment pathways:** We introduced a regression-based mediation framework to explore how sentiment affects financial performance through intermediary variables, deepening our understanding of causal pathways.

4. **Robustness verification across methods and datasets:** We conducted rigorous robustness checks—including DID regressions and multivariate analyses—to validate the consistency and reliability of our results.

5. **Evaluation of sentiment prediction models:** Finally, our study assessed the accuracy of financial sentiment prediction models, particularly the Random Forest model, across original and algorithmic texts, demonstrating the enhanced predictive capabilities of our methodological framework.

To provide a concrete illustration of this entire multi-stage pipeline, a detailed methodological walkthrough using a real-world example from our dataset is presented in Section 4.5.1.

4.2 Related works

4.2.1 Coreference Resolution

The intersection of coreference resolution and financial sentiment analysis has been a burgeoning area of interest within the computational linguistics and financial analytics communities. Early strides in coreference resolution, as established by Clark and Manning [2016], laid the groundwork for subsequent innovations in natural language processing (NLP) technologies. Their approach, further integrated into practical NLP tools like the spaCy library by Press and Wolf [2017], provided a foundational model for subsequent research endeavors.

In the realm of financial text analysis, the specificity and ambiguity of financial lexicon present unique challenges, as explored by Hovy et al. [2006] through the OntoNotes project. This comprehensive dataset, encompassing a wide range of text genres, has served as a critical benchmark for testing the efficacy of coreference resolution models. However, the financial domain demands a nuanced approach to coreference resolution, given the sector's specialized vocabulary and the critical importance of accurately identifying and linking entity references within texts.

Recent advancements in machine learning models for NLP, particularly those leveraging deep learning architectures like ELMo [Peters et al., 2018] and BERT [Devlin et al., 2019], have significantly pushed the boundaries of coreference resolution accuracy. The works of Lee et al. [2018] and Joshi et al. [2019] exemplify the potential of these models to achieve state-of-the-art results on benchmarks such as OntoNotes, demonstrating considerable improvements over traditional methods. Building upon these successes, Joshi et al. [2020] introduced SpanBERT, a pre-training approach focused on span representation, which further enhanced coreference resolution performance.

More recently, Wu et al. [2020] proposed CorefQA, a novel approach that formulates coreference resolution as a question answering task, leveraging the power of pre-trained language models to achieve impressive results. Ye et al. [2020] introduced a coreferential reasoning approach that incorporates entity-level information to improve the accuracy of coreference resolution in dialogues.Furthermore, Dobrovolskii [2021] introduced a word-level coreference resolution model that achieves competitive performance while being more computationally efficient than span-based models. A significant breakthrough in coreference resolution was recently achieved by Bohnet et al. [2023], who introduced a text-to-text (seq2seq) transition-based system using the multilingual T5 language model. This approach jointly predicts mentions and links, achieving state-of-the-art accuracy on the CoNLL-2012 datasets for English, Arabic, and Chinese. The system also demonstrated impressive performance in zero-shot, few-shot, and supervised settings on the SemEval-2010 datasets, substantially outperforming previous approaches.

These recent advancements highlight the ongoing progress in the field of coreference resolution, with researchers continually pushing the boundaries of what is possible with deep learning and large-scale language models. As these technologies continue to evolve, we can expect further improvements in the accuracy and efficiency of coreference resolution systems, enabling more sophisticated applications in various domains, including financial sentiment analysis.

4.2.2 Financial Sentiment Analysis

In the evolving field of financial sentiment analysis, recent advancements underscore the integration of sophisticated NLP techniques to address the nuanced demands of financial text analysis. The development of FinBERT [Liu et al., 2020], tailored for financial text mining, exemplifies this trend, leveraging the power of deep learning to decode the complexities of financial language. Similarly, the application of graph convolutional networks for financial statement fraud detection by Jiang et al. [2019] demonstrates

the versatility of NLP techniques in ensuring data integrity within the financial sector. The pioneering approach to cross-modal sentiment analysis by Wang et al. [2024] highlights the innovative applications and potential cross-disciplinary methodologies that could further enrich financial analysis. Additionally, the exploration of transformer models for financial market predictions by Zerveas et al. [2021] opens new avenues for applying advanced machine learning techniques to financial time series analysis. Despite these strides, the exploration of coreference resolution's role in enhancing sentiment analysis accuracy within financial texts signifies a fertile area for future research, aiming to fine-tune the precision of NLP applications in finance.

Propensity Score Matching (PSM) is a statistical technique used to estimate the effect of a treatment by accounting for the covariates that predict receiving the treatment [Rosenbaum and Rubin, 1983]. In the context of financial sentiment analysis, PSM is employed to match companies with similar characteristics, such as size, industry, and financial performance, but with different exposure such as sentiment.

Our method extends the conventional application of PSM and DID to financial sentiment analysis by integrating them with advanced NLP techniques, such as coreference resolution and domain-adapted language models. By leveraging PSM, we ensure that the companies being compared have similar characteristics, reducing the bias in the estimation of sentiment impact. Furthermore, the application of DID allows us to control for any time-invariant unobserved factors that may influence company performance, strengthening the causal inference of the impact of financial sentiments.

The combination of PSM and DID with coreference resolution-enhanced financial sentiment analysis represents a novel approach to understanding the complex relationship between sentiments and company performance. This innovative methodology enables us to isolate the causal effect of sentiments more effectively, providing a more nuanced and accurate understanding of how sentiments drive financial outcomes.

In summary, while notable advancements have been achieved in the domains of coreference resolution and sentiment analysis within the broader

ambit of NLP research, their application to financial text analysis—especially when combined with empirical financial methodologies like PSM and DID presents a promising domain for future scholarly exploration. The literature review underscores the necessity for more robust engagement with the specific challenges and opportunities presented by financial texts, as well as a call for studies that critically assess the limitations of existing research and propose comprehensive frameworks for integrating advanced NLP technologies into financial analytics.

4.3 Methodology

Our methodology advances the analytical rigor in financial sentiment analysis by integrating and enhancing coreference resolution techniques with machine learning algorithms. This integrated approach aims to refine the accuracy of entity recognition and sentiment measurement, thereby addressing the intricacies of financial sentiment analysis. We adopt a multi-step process that combines proprietary algorithms with public domain models to achieve this objective. The methodological framework is structured in the following sections.

4.3.1 Coreference Resolution Pre-Processing

We begin by assembling an extensive dataset of financial news articles and company filings, obtained from the Factiva database, which encompasses a wide range of companies included in the S&P 500 index. Each article is linked to one or more companies either through direct mentions within the text or via metadata information provided by Factiva. The collected data then undergoes a series of preprocessing steps, such as tokenization, stop word removal, and lemmatization, to clean and prepare the text for subsequent analysis.

Inspired by the groundbreaking research of Bohnet et al. [2023], we adopt their state-of-the-art coreference resolution approach which utilizes a text-to-text (seq2seq) transition-based system. This system, built upon

powerful pre-trained encoder-decoder architectures like T5, processes documents incrementally. For our experiments, we utilized the publicly available mT5 model fine-tuned by Bohnet et al. (2023) on the CoNLL-2012 English dataset, leveraging its proven performance for coreference resolution on our financial corpus.

Central to this coreference resolution pipeline is the transition-based system that processes a document one sentence at a time. The system maintains a state representing the current coreference clusters identified so far. For each sentence, it takes the text as input along with the current state (identified mentions and clusters), and the seq2seq model predicts a sequence of actions to identify mentions within the sentence and link them to existing clusters or form new ones. The action space conceptually involves operations like identifying a mention span, linking or appending a new mention to an existing cluster, and signaling the end of processing for the current sentence (SHIFT action).

To see how this process helps attribute sentiment correctly at the entity level, let's walk through the following illustration:

- **Previous Context/State:** After processing sentence 1 ("Alpha Corp. reported its quarterly earnings."), the system has identified clusters: C_1 = {"Alpha Corp."} and C_2 = {"its quarterly earnings"}.
- **Input Sentence 2:** "The company exceeded expectations, and it forecasts continued growth."
- **Predicted Action Sequence (Conceptual):**
  1. Identify mention: "The company".
  2. Predict Action: `Append(mention=‘‘The company’’, cluster=`C_1`)`. Links "The company" to the "Alpha Corp." cluster C_1. State updated.
  3. Identify mention: "it".
  4. Predict Action: `Append(mention=‘‘it’’, cluster=`C_1`)`. Links "it" also to the "Alpha Corp." cluster C_1. State updated.

5. Identify mention: “continued growth”.
6. Predict Action: `Append(mention=‘‘continued growth’’, cluster = NewCluster)`. Creates C_3. State updated.
7. Predict Action: `SHIFT`.

- **Updated State:** The key cluster for the company is now C_1 = {“Alpha Corp.”, “The company”, “it”}. Other clusters are C_2 = {“its quarterly earnings”} and C_3 = {“continued growth”}.

Here, the positive sentiment associated with “exceeded expectations” (linked implicitly to “The company”) and “forecasts continued growth” (linked explicitly to “it”) can now be correctly attributed to “Alpha Corp.” through the resolved coreference chain C_1. This demonstrates the importance of CR for accurate entity-level sentiment attribution.

This approach offers several advantages for our work, including the elimination of a separate mention detection step, the ability to directly leverage the power of advanced language models like mT5, and efficient handling of long documents through incremental processing. This pre-processing step is crucial for accurately attributing sentiment to the correct entities in the subsequent analysis phase.

4.3.2 Entity-Specific Sentiment Analysis

Leveraging the enhanced coreference resolution output, we perform sentiment analysis using FinBERT, a financial sector-specific adaptation of the BERT model. This step involves calculating sentiment scores for each entity within a given timeframe, based on the contextual analysis of sentences where the company is mentioned directly or identified through coreference resolution. The sentiment score for an entity e in a document d can be computed as:

$$S(e,d) = \frac{1}{N}\sum_{i=1}^{N} s(x_i) \tag{4.1}$$

where $S(e, d)$ represents the overall sentiment score for entity e in document d, N is the number of mentions of entity e in document d, and $s(x_i)$ is the sentiment score of the i-th mention of entity e, obtained from FinBERT.

4.3.3 Empirical Analysis Framework

The study employs a robust empirical framework to account for potential biases and confounders inherent in observational data. We combine Propensity Score Matching (PSM) with a Difference-in-Differences (DID) approach. The PSM methodology, first introduced by Rosenbaum and Rubin [1983] , is used to construct a valid control group by matching treated firms (i.e., those experiencing a sentiment shock) with non-treated firms that have similar observable characteristics. The propensity score, defined as the conditional probability of receiving the treatment given a vector of pre-treatment covariates, is estimated using a logistic regression model:

$$e(X) = Pr(D = 1|X) = \frac{1}{1 + e^{-\beta X}} \tag{4.2}$$

where D is the binary treatment indicator, X is the vector of covariates, and β is the vector of coefficients. This matching process ensures that the treated and control groups have balanced covariates, allowing for a more accurate estimation of the impact of financial sentiments on company performance.

On the other hand, the Difference-in-Differences (DID) approach is a quasi-experimental technique used to estimate the effect of a specific intervention or treatment by comparing the changes in outcomes over time between a treated group and a control group [Card and Krueger, 1994]. The basic formula for DID estimation is:

$$\delta = (\bar{Y}^t_{post} - \bar{Y}^t_{pre}) - (\bar{Y}^c_{post} - \bar{Y}^c_{pre}) \tag{4.3}$$

where δ is the DID estimator, $\bar{Y}^t_{post}$ and $\bar{Y}^t_{pre}$ are the average outcomes for the treated group in the post-treatment and pre-treatment periods, respectively, and $\bar{Y}^c_{post}$ and $\bar{Y}^c_{pre}$ are the average outcomes for the control group in

the post-treatment and pre-treatment periods, respectively. DID controls for time-invariant unobserved heterogeneity, which is crucial in financial sentiment analysis, as it helps to isolate the causal impact of sentiments on company performance by accounting for any pre-existing differences between the treated and control groups.

The novelty of our approach lies not in the development of these econometric methods themselves, but in their specific application to this problem domain. We are, to our knowledge, the first to use a PSM-DID framework to rigorously estimate the causal impact of an entity-specific financial sentiment signal that has been purified using state-of-the-art coreference resolution techniques. This combination allows for a more robust causal inference than traditional correlation or regression analyses of textual sentiment.

4.3.4 Robust Analysis: Mediation, DID, and Multivariate Regression Techniques

In addition to the primary PSM-DID analysis, our study employs mediation effect regression to investigate the potential mechanisms through which sentiment influences financial outcomes. Mediation analysis, as formalized by Baron and Kenny [1986] , aims to identify and explain the process underlying an observed relationship between an independent variable and a dependent variable via the inclusion of a third, mediator variable. The mediation effect is estimated using a series of regression equations:

$$Y = \alpha_1 + \beta_1 X + \epsilon_1 \tag{4.4}$$

$$M = \alpha_2 + \beta_2 X + \epsilon_2 \tag{4.5}$$

$$Y = \alpha_3 + \beta_3 X + \gamma M + \epsilon_3 \tag{4.6}$$

where Y is the outcome variable (financial performance), X is the independent variable (sentiment), M is the mediator variable, α_i are the intercepts, β_i and γ are the coefficients, and ϵ_i are the error terms.

Equation (4.4) represents the total effect of sentiment (X) on financial outcomes (Y). Equation (4.5) represents the effect of sentiment (X) on the

mediator variable (M). Equation (4.6) represents the effect of sentiment (X) and the mediator (M) on financial outcomes (Y), while controlling for the direct effect of sentiment.

The mediation effect, also known as the indirect effect, is quantified by the product of coefficients β_2 and γ. It measures the extent to which the mediator variable (M) accounts for the relationship between sentiment (X) and financial outcomes (Y). A significant mediation effect suggests that sentiment influences financial outcomes through the mediator variable.

To assess the robustness of our findings, we conduct a series of additional analyses. These include:

Post-financial sentiment analysis DID regression: We apply the Difference-in-Differences (DID) approach to the sentiment scores obtained from our analysis. This helps to control for potential confounding factors and strengthens the causal inference of the impact of sentiment on financial outcomes. Multivariate regression analyses: We employ multivariate regression models to account for multiple independent variables simultaneously. This allows us to control for potential confounding factors and assess the relative importance of sentiment in predicting financial outcomes. By conducting these robustness checks, we aim to validate the stability and reliability of our findings, ensuring that the observed effects of sentiment on financial outcomes are not driven by spurious relationships or unaccounted-for factors.

The novel contribution in this section is the application of this mediation framework to dissect the causal chain initiated by a CR-enhanced sentiment signal. By testing potential mediators, we can move beyond simply asking if sentiment has an effect, to asking how it exerts its influence on the market. The subsequent robustness checks, including post-financial sentiment analysis DID regression and standard multivariate regression analyses, employ conventional econometric techniques to validate the stability and reliability of our primary findings.

4.3.5 Exploring Dynamic Sentiment Transitions and Spatial Correlations

The research extends to examining the dynamics of sentiment over time and space. To model sentiment transitions, we employ a Markov Transition Probability Matrix, a standard tool in stochastic modeling for describing the probabilities of moving from one state to another in a sequence [Norris, 1998]. The matrix P is defined as:

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix} \tag{4.7}$$

where p_{ij} represents the probability of transitioning from state i to state j, and $\sum_{j=1}^{n} p_{ij} = 1$ for all i.

To analyze spatial correlations in sentiment (e.g., how sentiment about one company might affect sentiment about its geographically or economically linked peers), we use a Spatial Durbin Model (SDM) [Mur and Angulo, 2006]. The SDM is a spatial econometric model that incorporates both a spatially lagged dependent variable (ρWy) and spatially lagged independent variables ($WX\theta$), allowing it to capture both direct and spillover effects:

$$y = \rho Wy + X\beta + WX\theta + \epsilon \tag{4.8}$$

where y is the dependent variable, X is the matrix of independent variables, W is the spatial weight matrix, ρ is the spatial autoregressive coefficient, β and θ are the coefficient vectors, and ϵ is the error term.

This comprehensive methodological approach, underpinned by sophisticated NLP and machine learning techniques, sets a new benchmark in financial sentiment analysis. It not only enhances the precision of sentiment measurement but also provides a granular understanding of the interplay between financial news sentiment and market dynamics. Future research directions include exploring the potential of emerging coreference resolution techniques and their applications in financial analyses, further enriching the analytical landscape of financial sentiment analysis.

4.4 Financial Datasets

4.4.1 Factiva Database for Financial News Analysis

The Factiva database, produced by Dow Jones, is an essential resource for financial analysis. With access to over 30,000 sources worldwide, including news articles, company filings, financial data, and industry reports from more than 200 countries in 30+ languages, Factiva provides a comprehensive view of global financial markets. Its powerful search engine and advanced filtering options streamline information retrieval, while its data export capabilities in CSV, Excel, and XML formats make it invaluable for researchers and analysts incorporating financial news data into their work. Factiva's extensive coverage, sophisticated search features, and flexible data export options make it a crucial tool for anyone requiring up-to-date financial information and insights.

4.4.2 Text Data Processing

To construct a comprehensive corpus of finance-related articles, we gathered raw newspaper articles from the Factiva database, spanning the period from 2020 to 2022. This data collection effort focused on articles directly linked to companies listed within the S&P 500 index, ensuring the financial relevance of the corpus. To further refine the data sources, we restricted our selection to articles published by prominent financial news providers, such as the Wall Street Journal, the New York Times, USA Today, and the Washington Post. Each article was meticulously linked to at least one corresponding company, and some articles were associated with multiple S&P 500 companies due to their broader coverage of industry-wide developments. This initial data collection process yielded a substantial dataset comprising 200,642 article-company pairs.

To enhance the quality and relevance of the dataset for subsequent financial analysis, we implemented a rigorous data cleaning process. We conducted a manual review of the most frequently occurring headlines to

identify and remove articles that deviated from the desired focus on finance-related information. This eliminated opinion pieces, product reviews, and stock market summaries, ensuring the dataset's alignment with the objectives of our study. Furthermore, we employed the Harvard General Inquirer2 word lists "Legal" and "@Econ" to systematically count the frequency of legal and finance-related terms within each article. Articles with less than 5% of their words classified as legal or finance-related were excluded, ensuring that the corpus remained centered on information pertinent to investors. Additionally, adhering to the recommendations of Tetlock et al. [2008] and Loughran and McDonald [2016], we removed articles containing fewer than 50 words, as textual analysis of extremely short texts often yields unreliable results. These meticulous filtering steps resulted in a refined dataset comprising 150,374 article-company pairs, providing a robust foundation for our subsequent analyses.

Finally, we applied standard text-cleaning methods commonly utilized in natural language processing and finance to prepare the article bodies for further analysis. These methods involved the systematic removal of URLs, email addresses, numbers, punctuation, special symbols, and common English stop words. Additionally, we standardized all words to lowercase, replaced multiple white spaces with a single space, expanded contractions, and removed possessives ("'s") and hyphens. These preprocessing steps ensured that the text data was consistent, structured, and free from extraneous elements, thereby enhancing the effectiveness of our subsequent computational analyses.

4.4.3 Data summary

To compile the necessary data for our analysis, we obtained S&P 500 constituents, stock returns, and trading volumes from the CRSP (Center for Research in Security Prices) database, which provides comprehensive historical stock market data. Fundamental data and accounting information were retrieved from Compustat, a database of financial, statistical, and market information on active and inactive global companies. Fama & French

factors [Fama and French, 1993] and risk-free rates were sourced from Kenneth French's Data Library, a standard resource in empirical finance for asset pricing model factors. Additionally, we utilized Harvard General Inquirer's "Econ@" and "Legal" lists to identify economics, business, and legal terms, respectively. For sentiment analysis, we employed Loughran & McDonald's negative and positive word lists [Loughran and McDonald, 2011]. To count company mentions and their locations within articles, we used a search vector composed of company names and their variations, as identified from CRSP common names and our newspaper article database.

Assessing the impact of our coreference resolution method—specifically, its efficacy in identifying instances where a text segment pertains to a particular company—on sentiment metrics necessitates the availability of company-level sentiment estimates within the same article. As discussed previously, a single article may encompass discussions on multiple companies with varying tones. Consequently, obtaining an unbiased estimation of company-level tone becomes imperative to gauge the extent to which coreference resolution enhances the alternative approach of uniformly applying the same tone to all companies featured in a given article.

Table 4.1: Descriptive Statistics for the Analyzed Variables.

Variable Name	Unit	Sample Size	Mean	Std. Dev.	Min	Max	Skewness	Kurtosis
AQTB	%	33768	0.06	1.35	-9.85	9.92	0.15	5.15
SE	%	33768	104.85	11.50	70.20	140.50	0.51	3.10
RTY	%	33768	41.25	15.80	2.15	92.50	0.35	2.55
WERT	%	33768	39.50	16.10	1.85	95.80	0.41	2.80
POIUQ	%	33768	35.10	25.40	0.25	195.50	1.85	6.90
UUYTR	%	33768	15.40	9.10	1.10	70.10	1.15	5.10
YYUTR	%	33768	6.80	6.20	0.50	55.20	1.05	4.50
Age	Year	33768	7.80	7.34	0.000	21.280	0.457	1.934

The descriptive statistics in Table 4.1 reveal the characteristics of the variables used in the analysis. The variable 'AQTB' (Abnormal Return) is nearly symmetric and centered close to zero, but exhibits a distribution with heavier tails than a normal distribution, which is a common characteristic of financial returns. The market and sentiment indicators, 'SE' (Degree of Bull Market), 'RTY' (Financial Positive Index), and 'WERT' (Fiscal Negative Index), show distributions that approach normality. Their low skewness

and kurtosis suggest a relatively balanced spread of values with a limited presence of extreme outliers. In contrast, the remaining variables all exhibit positive skewness and are leptokurtic, indicating distributions with heavier tails. The mediator variable 'POIUQ' (Economic Benefits) is the most skewed and displays the highest kurtosis, which signifies that while its values are often low, there is a notable tail of high-impact events. Similarly, 'UUYTR' (Numerical Density) and 'YYUTR' (Coreference Complexity Factor) are positively skewed, suggesting their distributions are concentrated at lower values with a tail extending towards higher values. The 'Age' variable also shows a positive skew, reflecting the presence of many long-established firms in the S&P 500. This analysis highlights the varied statistical properties of the data, which is crucial for the subsequent regression modeling.

4.5 Results

Our analysis commenced with the collection of an extensive dataset of news headlines directly related to individual companies within the S&P 500, covering the period from January 1, 2020, to December 31, 2022. The initial dataset comprised 418,703 headlines, which, upon undergoing our refined entity recognition process, expanded to include 576,191 instances of enriched news texts. These texts were meticulously analyzed for sentiment extraction, employing a FinBERT model specifically tuned for the financial domain. This model, selected for its superior accuracy in financial sentiment analysis, facilitated a comprehensive evaluation of sentiment across the dataset, allowing us to accurately gauge the prevailing market sentiments associated with these entities over the specified period.

The correlation test results in Table 4.2 illustrate the complex interrelationships between various variables, highlighting both direct and inverse correlations. Notably, WERT shows a strong positive correlation with AQTB (0.331), suggesting that as one variable increases, so does the other, whereas it exhibits a significant negative correlation with UUYTR (-0.385), indicating an inverse relationship. The high positive correlation between UUYTR

Table 4.2: Correlation Test Results.

	AQTB	SE	RTY	WERT	POIUQ	UUYTR	YYUTR	Age
AQTB	1							
SE	-0.041	1						
RTY	-0.050	-0.013	1					
WERT	0.331	-0.074	-0.098	1				
POIUQ	-0.182	0.131	0.005	-0.260	1			
UUYTR	-0.146	0.105	-0.039	-0.385	0.850	1		
YYUTR	0.126	-0.040	0.119	0.293	-0.318	-0.351	1	
Age	-0.153	0.002	-0.020	-0.127	0.318	0.208	-0.186	1

and POIUQ (0.850) stands out, suggesting a very strong direct relationship between these variables. Conversely, POIUQ and WERT have a notable negative correlation (-0.260), indicating that as one increases, the other tends to decrease. The correlations between other variables, such as RTY with AQTB and SE, are relatively weak, suggesting less direct influence on each other. Age shows a moderate positive correlation with POIUQ (0.318) and a mild inverse correlation with YYUTR (-0.186), highlighting demographic influences on these variables. These correlations reveal the intricate dynamics at play within the dataset, providing insights into how changes in one variable might affect another, which is crucial for understanding the underlying patterns and potential causal relationships.

Table 4.3: PSM-DID Pre-matching Balance Test.

Variable Name	Group	Mean	Mean Difference	T-value	P-value
POIUQ	Experimental Control	14.591 13.882	-4.704	0.537	0.676
WERT	Experimental Control	1.291 1.153	0.612	0.039	1.234

Table 4.3 presents a PSM-DID pre-matching balance test comparing experimental and control groups across two variables, POIUQ and WERT. For POIUQ, despite the experimental group having a higher mean (14.591) compared to the control group (13.882), resulting in a mean difference of -4.704, the T-value (0.537) and P-value (0.676) indicate that this difference is not statistically significant. In the case of WERT, the experimental group's mean is slightly higher (1.291) than that of the control group (1.153), with a mean

difference of 0.612. However, the extremely high P-value (1.234) alongside a negligible T-value (0.039) suggests that this observed difference lacks statistical significance, questioning the impact of the intervention. This analysis demonstrates that, at least on the surface, the interventions may not have had a significant effect on the measured outcomes, as indicated by the high P-values, suggesting the need for further investigation or reconsideration of the intervention's efficacy.

Table 4.4: PSM-DID Post-matching Balance Test.

Variable Name	Group	Mean	Mean Difference	T-value	P-value
POIUQ	Experimental Control	12.583 11.052	-5.320	0.462	0.670
WERT	Experimental Control	1.876 1.243	0.474	0.038	0.684

The results from a Propensity Score Matching-Difference in Differences (PSM-DID) post-matching balance test are summarized in Table 4.4, aimed at evaluating the comparability of experimental and control groups across variables POIUQ and WERT. This test is crucial for causal inference studies, ensuring that observed differences post-treatment are attributable to the intervention. For POIUQ, the experimental group exhibits a mean of 12.583 compared to the control group's 11.052, with a mean difference of -5.320, a T-value of 0.462, and a P-value of 0.670. These statistics indicate no significant difference between groups, highlighting an effective matching process. Similarly, WERT shows a mean of 1.876 for the experimental group and 1.243 for the control, with a mean difference of 0.474, a T-value of 0.038, and a P-value of 0.684, reinforcing the matching process's success. The non-significant P-values for both variables (> 0.05) confirm the efficacy of the PSM-DID method in creating statistically comparable groups, laying a solid foundation for attributing observed outcome differences to the treatment effect with minimal confounding from pre-treatment disparities.

Table 4.5 presents the results of a Propensity Score Matching-Difference in Differences (PSM-DID) baseline regression analysis, aimed at evaluating the impact of an intervention across different time frames. The coefficients

Table 4.5: PSM-DID Baseline Regression.

	(1) [-36,36]	(2) [-24,24]	(3) [-18,18]
SE*time	-0.154*** (-3.425)	-0.082** (-2.021)	-0.070** (-1.819)
SE	0.226*** (2.040)	0.337 (2.081)	0.424** (1.852)
Time	-0.031 (-1.200)	-0.025** (-0.912)	-0.026*** (-1.149)
Year	Y	Y	Y
Industry	Y	Y	Y
Cons	0.560*** (6.300)	0.571*** (9.198)	0.788*** (7.634)
N	20440	20828	19264
R^2	0.110	0.130	0.168

for SE*time, significant across all models, suggest a varying effect of the intervention over time. The significance levels, denoted by asterisks, underscore the robustness of these findings. The SE coefficient shows variability in the intervention's impact, indicating its effects become more pronounced and consistent over narrower time frames. The Time variable shows a negative but only sometimes significant relationship, hinting at a potential overall downward trend. The inclusion of Year and Industry as control variables ensures that the analysis accounts for fixed effects. The constant term is highly significant across all models, indicating a substantial effect not explained by the included variables. The sample size and R^2 values provide insights into the robustness and explanatory power of the regression.

In the presented analysis of Table 4.6, employing a Propensity Score Matching-Difference in Differences (PSM-DID) fixed effects model, seven distinct specifications were systematically explored to assess the influence of various predictors on the dependent variable. Each model specification consistently incorporated the AQTB variable, alongside the interaction term SE×time, revealing a statistically significant negative impact across all configurations. This finding signifies a consistent diminishment in the outcome variable attributable to the interaction between time and standard error, with significance levels rigorously denoted as $p < 0.10$, $p < 0.05$, and $p < 0.01$ for *, **, and ***, respectively. Furthermore, the SE variable

Table 4.6: PSM-DID Fixed Effects Model.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	AQTB	AQTB	AQTB	AQTB	AQTB	AQTB	AQTB
SE*time	−0.235**	−0.215*	−0.232**	−0.246**	−0.234**	−0.232**	−0.210**
	(−2.72)	(−2.16)	(−2.91)	(−2.34)	(−2.73)	(−2.54)	(−2.62)
SE	0.173**	0.146*	0.176**	0.126*	0.135**	0.146**	0.164**
	(2.00)	(1.91)	(2.01)	(1.96)	(2.05)	(2.01)	(1.96)
time	−0.0870***	−0.0574***	−0.0920***	−0.0550***	−0.0622***	−0.0922***	−0.0784***
	(−3.42)	(−3.34)	(−3.40)	(−3.34)	(−3.91)	(−3.40)	(−3.63)
YYUTR	0.00143						0.00252
	(0.87)						(1.23)
RTY		−0.00714**					−0.00564**
		(−2.35)					(−2.46)
Age			−0.0421				−0.0151
			(−0.21)				(−0.21)
WERT				0.136***			0.142***
				(7.51)			(8.00)
POIUQ					−0.164***		−0.143***
					(−4.27)		(−3.30)
UUYTR						−0.000520	0.269**
						(−0.00)	(2.25)
Year	Y	Y	Y	Y	Y	Y	Y
Industry	Y	Y	Y	Y	Y	Y	Y
Cons	0.879***	1.228***	0.820**	0.654***	2.864***	0.901***	2.460***
	(4.23)	(5.28)	(2.35)	(3.49)	(6.23)	(5.41)	(3.82)
N	17253	17253	17253	17253	17253	17253	17253
R^2	0.053	0.053	0.053	0.053	0.053	0.053	0.053

itself exhibited a positive and significant influence across all models, indicating that an increase in the standard error correlates with an elevation in the dependent variable. The inclusion of additional variables such as YYUTR, RTY, Age, WERT, POIUQ, and UUYTR, each introduced in individual model specifications, exerted varying degrees of influence on the dependent variable, thereby underscoring their predictive power. Notably, the significant positive effect of WERT and the negative impact of POIUQ in specific models highlight the nuanced interplay between these variables and the outcome of interest. The incorporation of Year and Industry as fixed effects served to control for unobservable heterogeneity, thereby enhancing the reliability of the estimations. Despite the consistent sample size (N = 17253) and an R^2 value of 0.053 across all models indicating a degree of explained variance, a considerable portion of the outcome vari-

ability remains unexplained, pointing towards the complex and multifaceted nature of the dependent variable's determinants. This analysis underscores the critical importance of considering both the temporal dynamics and the intrinsic variability within the data to elucidate the underlying mechanisms influencing the outcome variable, thereby contributing to a more nuanced understanding of the subject matter within the scientific discourse.

Table 4.7: Mediation Effect Regression Model.

	(1) AQTB	(2) AMI	(3) AQTB	(4) Indhold	(5) AQTB
On	-0.0306*** (-3.04)	-0.0124*** (-2.93)	-0.0478*** (-3.41)	1.4743*** (-7.09)	-0.0516*** (-3.71)
YYUTR RTY	0.00532*** (-5.12) 0.00159*** (-1.81)	-0.00472*** (-13.28) 0.00614*** (-18.97)	0.00218*** (-2.41) 0.0049*** (-5.05)	0.25092*** (-12.92) 0.15855*** (-13.53)	0.00438*** (-3.75) 0.00188*** (-1.98)
Age WERT	0.0207*** (-10.76) 0.134***	-0.0457*** (-34.91) -0.169***	0.0104*** (-3.54) 0.089***	3.2331*** (-49.26) 7.452***	0.0232*** (-8.28) 0.127***
POIUQ UUYTR	-0.0734*** (-7.43) 0.499***	-0.4153*** (-83.05) 0.457***	-0.2386*** (-21.61) 0.682***	2.794*** (-11.39) 2.125***	-0.0928*** (-5.77) 0.349***
AMI	(-8.22)	(-18.65)		(-2.05 (-11.58)	(-7.01)
Indhold			-0.424 (-34.80)		-0.000273 (-0.94)
Year Industry	Y Y	Y Y	Y Y	Y Y	Y Y
Cons	1.686*** (-17.35)	-3.969*** (-65.64)	0.00401 (-0.03)	-55.69*** (-23.33)	1.670*** (-16.35)
N R^2	32017 0.242	32017 0.172	32017 0.462	32017 0.328	32017 0.196

Table 4.7 offers a comprehensive examination of the mediation effects within a regression model framework, across five distinct specifications. This analysis aims to uncover the nuanced roles played by variables such as AQTB, AMI, and Indhold in influencing the model's outcomes. Notably, the presence of significant predictors is confirmed through their coefficients and statistical significance levels, with asterisks ($p < 0.01$) marking a high degree of significance. Particularly in models (1), (3), and (5), AQTB emerges as a central variable, consistently demonstrating a negative and statistically significant impact on the dependent variable. This pattern highlights AQTB's

critical mediation role, with coefficients like -0.0306*** in model (1) and -0.0516*** in model (5) reinforcing the variable's substantial influence across various contexts.

Conversely, AMI, featured in models (2) and (5), displays a diverse impact. Model (2) reveals a significant negative effect (-0.0124***), suggesting AMI's capacity to significantly alter the mediation pathway. In stark contrast, Indhold, predominantly included in model (4), exhibits a strong positive effect (1.4743***), indicating a significant mediating role distinct from those observed with AQTB and AMI. The analysis further delves into the effects of other pivotal variables like YYUTR, RTY, Age, WERT, and POIUQ UUYTR, each uniquely contributing to the mediation effect across the models. For instance, the variables Age and WERT show significant effects in both directions, underscoring the complexity of their roles in the mediation process. The inclusion of Year and Industry as fixed effects across all models serves to mitigate potential confounding influences, thereby bolstering the reliability of the findings. The variation in constants across the models, from 1.686*** in model (1) to -55.69*** in model (4), reflects the distinct baseline levels inherent to each model's context. With a consistent sample size (N=32017) and R^2 values ranging from 0.172 in model (2) to 0.462 in model (3), the analysis highlights the models' differential explanatory capacities and the intricate relationships at play. These insights afford a deeper comprehension of the mediation effects exerted by various variables, enriching our understanding of the mechanisms underpinning the observed outcomes.

The robustness check outlined in Table 4.8 evaluates the stability of the effects across different sample sizes and time frames, specifically focusing on the full sample, a mid-range period ($[-36, 36]$), and a short-range period ($[-18, 18]$). The SE×time interaction demonstrates a statistically significant negative impact across all time frames, with the effect size increasing as the window narrows, indicating a stronger interaction effect in more immediate proximity to the event ($-0.274*$ in the $[-18, 18]$ period). The SE variable shows variability in its significance across different periods, highlighting its fluctuating influence.

Table 4.8: Robustness Check.

	-1 full sample	-2 [-36,36]	-3 [-18,18]
SE*time	-0.126*** (-2.55)	-0.235*** (-2.14)	-0.274* (-2.29)
SE time	0.066** (-2.59)	0.169 (-1.89)	0.141* (-2.01)
YYUTR	-0.0363** (-1.78)	-0.0316 (-1.79)	-0.1220*** (-3.47)
RTY Age	0.00294*** (-5.92)	0.00013 (-0.06)	0.00231 (-1.14)
	0.00190*** (-3.51)	0.00146 (-1.23)	-0.00606** (-1.71)
WERT	0.0228 (-1.08)	-0.0911 (-2.59)	-0.0110 (-0.22)
POIUQ	0.108*** (-23.40)	0.168*** (-15.32)	0.176*** (-6.10)
UUYTR	-0.00929 (-0.89)	-0.01150 (-0.59)	-0.15498** (-3.16)
	0.0801*** (-3.66)	-0.0762 (-0.81)	0.3146** (-2.78)
Year Industry	Y Y	Y Y	Y Y
Cons	1.862*** (-15.47)	1.357*** (-6.36)	3.168*** (-5.5)
N R^2	56523 0.1	27664 0.1	14333 0.1

Notably, the variable POIUQ consistently exhibits a significant positive effect across all models, underscoring its robustness ($0.108 * **$, $0.168 * **$, and $0.176 * **$), whereas other variables like YYUTR and RTY Age display mixed results in terms of significance, suggesting a more complex relationship with the dependent variable. The inclusion of Year and Industry as fixed effects across all specifications ensures control for potential confounding factors, enhancing the credibility of the findings. The constants across models ($1.862***$, $1.357***$, $3.168***$) further affirm the presence of a baseline effect that remains significant regardless of the time frame considered. The sample sizes and R^2 values (0.1 across all models) indicate a moderate explanatory power, with the models capturing a consistent portion of the variability in the dependent variable across varying contexts. This robustness check thus validates the stability and reliability of the observed effects, contributing to a nuanced understanding of the underlying dynamics at play.

Table 4.9: Post-Financial Sentiment Analysis DID Regression.

	(1) AQTB1	(2) AQTB1	(3) AQTB1	(4) AQTB1	(5) AQTB1	(6) AQTB1	(7) AQTB1
SE*time	-0.180**	-0.281**	-0.174**	-0.225**	-0.286**	-0.296**	-0.176**
	(-2.47)	(-2.30)	(-1.78)	(-2.49)	(-1.65)	(-1.84)	(-2.29)
SE	0.175**	0.185**	0.242**	0.167**	0.270**	0.187**	0.159**
	(1.862)	(1.751)	(2.288)	(2.688)	(2.576)	(2.441)	(1.883)
time	-0.038**	-0.036**	-0.042**	-0.048**	-0.062**	-0.037**	-0.054**
	(-2.66)	(-2.43)	(-1.67)	(-2.03)	(-2.30)	(-1.71)	(-2.40)
YYUTR	0.00142						0.00161
	(0.76)						(0.71)
RTY		-0.0057**					-0.00457*
		(-2.36)					(-1.72)
Age			-0.0251				-0.00347
			(-0.21)				(-0.07)
WERT				0.136***			0.162***
				(8.96)			(7.41)
POIUQ					-0.165***		-0.163***
					(-4.25)		(-3.16)
UUYTR						-0.173	0.176
						(-1.45)	(1.31)
Year	Y	Y	Y	Y	Y	Y	Y
Industry	Y	Y	Y	Y	Y	Y	Y
Cons	0.942***	1.143***	0.894**	0.725***	2.718***	0.269***	2.314***
	(5.95)	(4.75)	(4.50)	(5.80)	(4.60)	(5.80)	(5.75)
N	17413	17413	17413	17413	17413	17413	17413
R^2	0.042	0.042	0.042	0.042	0.042	0.042	0.042

The analysis encapsulated in Table 4.9 showcases a Difference in Differences (DID) regression focusing on the post-financial sentiment across seven distinct models, all under the umbrella of the AQTB1 variable. The SE×time interaction across these models consistently demonstrates a statistically significant negative impact, with coefficients ranging from -0.180** to -0.296**, highlighting the temporal and conditional variability in sentiment post-financial events. This effect is underscored by the consistent significance of the SE variable itself, indicating a robust relationship between standard errors and the outcome variable across all models.

Further analysis reveals nuanced relationships with other variables across the models. For instance, YYUTR and RTY exhibit conditional significance, indicating varying influences on the outcome based on the model. Specifically, RTY shows a negative effect in models (2) and (7), suggesting

sector-specific temporal dynamics. Additionally, WERT and POIUQ variables emerge with strong positive and negative impacts respectively, indicating their crucial roles in driving sentiment post-financial events. Notably, the constant terms across models signify the baseline sentiment level, with significant variability indicating the diverse contexts of the financial sentiment analysis. The uniformity in sample size (N=17413) and R^2 values (0.042) across all models suggests a consistent explanatory power, albeit modest, underscoring the complexity of financial sentiment dynamics as captured by the DID regression analysis.

Table 4.10: Multivariate Regression Results.

	RTY			Top5		
	Low -1	Mid -2	High -3	Low -4	Mid -5	High -6
On	0.0023 (-0.07)	-0.0157 (-0.89)	-0.0806*** (-3.45)	-0.0049 (-0.17)	-0.0581** (-1.43)	-0.0652*** (-3.48)
YYUTR	-0.0023 (-0.88)	0.0151*** (-5.87)	0.0112*** (-6.41)	-0.0003 (-0.19)	0.0057*** (-3.85)	0.0109*** (-5.48)
RTY	0.0001 (-0.02)	0.0009 (-0.37)	0.0001 (-0.06)	-0.0025 (-1.00)	-0.0033 (-1.22)	0.0015 (-1.08)
Age	0.0123*** (-2.63)	0.0234*** (-3.24)	0.0357*** (-6.28)	0.0173*** (-2.68)	0.0264*** (-6.23)	0.0249*** (-3.94)
WERT	0.2157*** (-19.07)	0.1361*** (-5.83)	0.1304*** (-9.37)	0.2832*** (-13.30)	0.1579*** (-7.23)	0.0928*** (-5.45)
POIUQ	-0.0612*** (-1.94)	-0.0432** (-1.73)	-0.0834*** (-4.05)	-0.0264 (-1.72)	-0.1224*** (-4.54)	-0.0669*** (-2.36)
UUYTR	0.2746*** (-4.63)	-0.0099 (-0.16)	0.6785*** (-8.13)	0.4079*** (-3.60)	0.3687*** (-3.52)	0.2294*** (-2.01)
Year Industry	Y Y	Y Y	Y Y	Y Y	Y Y	Y Y
Cons	1.642*** (-6.82)	1.257*** (-5.62)	1.426*** (-6.85)	1.735*** (-5.14)	2.045*** (-6.19)	1.676*** (-5.05)
N R2	10915 0.159	12027 0.141	10895 0.163	11455 0.208	11677 0.208	11853 0.198

The multivariate regression results, as summarized in Table 4.10, offer a nuanced exploration of the effects across different groups delineated by RTY and Top5 categories (Low, Mid, High). The analysis demonstrates significant variability in the impact of the examined variables. Specifically, the 'On' variable's effect transitions from non-significant in the Low category to highly significant and negative in the High category for both RTY

and Top5, signifying an escalating adverse impact with increasing category level. The YYUTR variable exhibits a positive and significant influence in the higher categories across both dimensions, suggesting an augmenting positive contribution. Conversely, Age and WERT consistently show a significant positive impact across all categories, indicating their overarching beneficial effects irrespective of group classification. The POIUQ variable reveals a mixed pattern, with its effect ranging from significantly negative to non-significant, highlighting the complexity of its influence. The UUYTR variable, notably, presents a stark contrast in its impact between the RTY and Top5 dimensions, with a significant positive effect in the higher categories, particularly notable in the RTY High category with a coefficient of 0.6785***. Controls for Year and Industry are uniformly applied, ensuring the robustness of the analysis against external confounding factors. The models' constants indicate a significant baseline effect across all categories, with variability reflective of the differing impacts per category. The analysis is further supported by the sample sizes and R2 values, which suggest a reasonable model fit and explanatory power across the board. This detailed examination underscores the intricacies and differential impacts of various predictors within the multivariate regression framework, providing valuable insights into the underlying dynamics at play.

The regression analysis captured in Table 4.11 delineates the nuanced impact of various predictors across different segments of the Z-index and S-index. The 'On' variable's negative impact is notably significant in the low segments of both indices, with stronger significance in the S-index (-0.0679***), indicating a pronounced sensitivity of financial sentiment in environments characterized by lower index values. This effect attenuates and loses statistical significance as we move to the mid and high segments, suggesting a diminishing influence of immediate factors with increasing index values.

Variables such as YYUTR and Age exhibit a consistently positive impact across most segments, emphasizing their stabilizing influence on financial sentiment. Specifically, YYUTR shows a significant positive effect in the

Table 4.11: Post-Financial Sentiment Analysis Multivariate Regression Results.

	Z-index			S-index		
	Low (1)	Mid (2)	High (3)	Low (4)	Mid (5)	High (6)
On	-0.0580**	-0.0119	-0.0333	-0.0679***	-0.0068	-0.0237
	(-2.39)	(-0.5)	(-1.42)	(-2.19)	(-0.25)	(-0.68)
YYUTR	0.0113***	0.0089***	-0.0020	0.0107***	0.0135***	-0.0023
	(5.47)	(2.86)	(-1.6)	(7.54)	(4.88)	(-1)
RTY	0.0040***	0.0005	0.0013	0.0019	-0.0006	0.0003
	(-2.88)	(0.41)	(0.81)	(2.69)	(-0.42)	(0.12)
Age	0.0264***	0.0367***	0.0184	0.0386***	0.0184***	0.0191***
	(6.64)	(6.52)	(-3.43)	(8.33)	(4.26)	(2.54)
BM	0.1059***	0.1872	0.2400	0.1485	0.1569***	0.2182***
	(8.69)	(13.8)	(8.84)	(7.42)	(8.39)	(9.98)
POIUQ	-0.0473***	-0.1129***	-0.0791	-0.0698***	-0.0376**	-0.1085***
	(-3.15)	(-3.51)	(-3.91)	(-3.92)	(-2.69)	(-2.99)
UUYTR	0.5321***	0.4621***	0.0628	0.7788***	0.2330	0.0739
	(9.39)	(2.81)	(0.66)	(6.63)	(2.67)	(0.72)
Year	Y	Y	Y	Y	Y	Y
Industry	Y	Y	Y	Y	Y	Y
Cons	0.883***	1.791***	2.154***	1.075***	1.379***	2.225***
	(3.64)	(6.83)	(6.31)	(4.23)	(5.49)	(6.61)
N	12026	12073	12033	12051	12002	12079
R^2	0.276	0.276	0.276	0.276	0.276	0.276

low and mid segments of both indices, with its influence slightly waning in the high segments. Age, on the other hand, maintains its significance across all segments, underscoring the critical role of temporal factors in shaping financial sentiment. The BM variable stands out with its significant positive effect across all segments, particularly highlighting its increasing importance in higher index values. In contrast, the POIUQ variable consistently exerts a negative influence across the board, with UUYTR displaying a significant positive impact mainly in the lower segments of both indices. This comprehensive analysis, supported by controls for Year and Industry, provides a deep dive into the dynamics of financial sentiment across different market conditions, as evidenced by the uniform R^2 values (0.276) across all models, indicating a consistent explanatory power of the regression framework.

The Markov Transition Probability Matrix, showcased in Table 4.12, elucidates the intricacies of spatial entity transitions across different temporal

Table 4.12: Markov Transition Probability Matrix After Coreference Resolution.

Spacial Lags	t \t+1	n	1	2	3	4
1	1	678	0.6254	0.3054	0.0062	0.1524
	2	421	0.2015	0.5214	0.0048	0.3625
	3	201	0.0023	0.3145	0.1425	0.4955
	4	65	0.01425	0.0241	0.1552	0.5162
2	1	152	0.5218	0.6254	0.2565	0.6029
	2	431	0.0062	0.4429	0.1144	0.7824
	3	405	0.0047	0.5932	0.9524	0.8145
	4	152	0.0001	0.6041	0.1152	0.9933
3	1	15	0.5026	0.2014	0.2004	0.4582
	2	99	0.0954	0.3026	0.0036	0.9246
	3	272	0.0262	0.0426	0.0042	0.8556
	4	130	0.0002	0.4814	0.0142	0.2651
4	1	6	0.2059	0.5241	0.6254	0.4157
	2	42	0.0841	0.8012	0.7152	0.3311
	3	182	0.0847	0.6231	0.4821	0.2514
	4	569	0.0042	0.5471	0.2695	0.3369

phases, structured around four spatial lags. This analytical framework is instrumental in quantifying the probabilities of transitioning from one state at time t to another at time $t + 1$. It is pivotal for deciphering the underlying mechanisms of spatial dynamics, highlighting the tendencies of states to either maintain their current status or evolve into new states, with these probabilities significantly influenced by spatial proximity.

The matrix reveals that certain transitions have notably high probabilities, indicating a pronounced tendency for entities to either remain in their current state or move to a particular future state. For example, a significant transition probability is observed at spatial lag 2, where the transition probability from state 1 to state 4 is notably high (0.6029), suggesting a considerable likelihood of advancement or change to a distinct state over time. Conversely, lower transition probabilities across various segments imply a degree of unpredictability in state evolution, underscoring the complex interplay between spatial dynamics and external influences. The disparities in sample sizes across spatial lags and states introduce an additional layer of complexity, where smaller sample sizes may result in less reliable estimates of transition probabilities. Thus, this matrix serves as an essential analytical tool for comprehending the temporal shifts of spatial entities, offering

invaluable insights for predictive modeling and strategic decision-making in spatial research.

Table 4.13: Empirical Analysis of Factors Influencing Sentiment Identifiability Post-Coreference Resolution

	W (Wide-coverage)		P (Primary-focus)		S (Short & Specific)	
Variable	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
EE_{t-1}	0.427^{***}	0.425^{***}	0.265^{***}	0.426^{***}	0.403^{***}	0.825^{***}
	(8.62)	(11.27)	(8.66)	(9.82)	(8.92)	(9.98)
FD	-0.265^{***}	-0.235^{***}	-0.159^{***}	-0.226^{***}	-0.127^{***}	-0.154^{***}
	(−6.03)	(−5.20)	(−4.62)	(−5.61)	(−5.53)	(−6.23)
$\ln(ER_1)$	-0.135^{**}	-0.152^{**}	-0.091^{*}	-0.110^{**}	-0.103^{**}	-0.125^{**}
	(−2.69)	(−2.85)	(−1.92)	(−2.15)	(−2.06)	(−2.30)
$\ln(ER_1)^2$		0.025^{*}		0.015		0.020^{*}
		(1.80)		(1.15)		(1.95)
$\ln(ER_2)$	-0.088^{*}	-0.095^{*}	−0.050	−0.065	−0.070	-0.080^{*}
	(−1.75)	(−1.85)	(−1.10)	(−1.30)	(−1.50)	(−1.70)
$\ln(ER_2)^2$		0.010		0.008		0.012
		(0.90)		(0.70)		(1.10)
IS	-0.362^{***}	-0.350^{***}	-0.325^{***}	-0.316^{***}	-0.227^{***}	-0.215^{***}
	(−3.69)	(−3.60)	(−3.86)	(−3.70)	(−3.62)	(−3.50)
FDI	0.694^{***}	0.703^{***}	0.262^{***}	0.285^{***}	0.541^{***}	0.525^{***}
	(3.99)	(4.10)	(3.92)	(4.05)	(4.01)	(3.90)
TRADE	-0.159^{**}	-0.145^{*}	−0.065	−0.082	-0.043^{*}	-0.056^{*}
	(−2.15)	(−1.90)	(−0.92)	(−1.10)	(−1.70)	(−1.85)
UL	-0.726^{***}	-0.833^{***}	-0.522^{***}	-0.670^{***}	-0.452^{***}	-0.598^{***}
	(−4.28)	(−4.50)	(−3.51)	(−4.03)	(−3.80)	(−4.15)
$(UL)^2$		0.140^{***}		0.105^{**}		0.090^{**}
		(3.06)		(2.50)		(2.30)
ρ	0.525^{***}	0.530^{***}	0.401^{***}	0.415^{***}	0.358^{***}	0.365^{***}
	(3.85)	(3.90)	(3.10)	(3.20)	(2.90)	(3.00)
θ	-0.066^{**}	-0.072^{**}	-0.041^{*}	-0.045^{*}	−0.031	−0.033
	(−2.15)	(−2.20)	(−1.85)	(−1.90)	(−1.10)	(−1.15)
Log-L	−370.26	−365.15	−280.50	−275.80	−310.90	−305.25
Obs	470	470	470	470	470	470

Notes: Significance levels: *** p<0.01, ** p<0.05, * p<0.1. Values in parentheses are t-statistics. This table presents factors influencing the identifiability of sentence-level financial sentiment (derived post-CR and FinBERT) under different news contexts: W (Wide-coverage news), P (Primary-focus news), and S (Short & Specific news). Model 2 adds quadratic terms to Model 1's linear specification for variables ER1, ER2, and UL.

The empirical analysis detailed in Table 4.13 provides a nuanced understanding of the factors impacting the identifiability of financial sentiment encoding fields, subsequent to the crucial step of coreference resolution. This table examines these relationships under different news contexts—namely,

sentences from Wide-coverage news (W), Primary-focus news (P), and Short and Specific news (S). For each context, Model 1 assesses linear effects of the explanatory variables, while Model 2 extends this by incorporating quadratic terms for Exchange Rates (ER1, ER2) and Uncertainty Level (UL) to explore potential non-linear relationships. The dependent variable, for the models presented in this table, is operationalized as the absolute t-statistic of the CR-enhanced sentence-level sentiment score when it is used to explain short-term stock abnormal returns in an event-study style regression. A higher absolute t-statistic signifies a stronger and more statistically reliable link between our derived sentiment and the market's reaction, indicating greater effectiveness of the sentiment signal.

Several key insights emerge from the analysis. A consistent and notably strong positive influence on sentiment signal effectiveness (the dependent variable Y) is observed from lagged emotional expression (EE_{t-1}). Across all models and news contexts, its coefficients (e.g., ranging from 0.265^{***} to 0.825^{***}) are highly significant. This suggests that a clear history of emotional articulation regarding an entity robustly enhances the statistical power with which its current sentiment, once accurately attributed by CR, explains market reactions. This underscores the persistence of sentiment signals and the importance of historical emotional context.

Furthermore, higher Foreign Direct Investment (FDI) is consistently associated with more effective sentiment signals (e.g., coefficients 0.262^{***} to 0.703^{***}), possibly reflecting that increased transparency or more standardized reporting associated with international investment leads to sentiment that is more readily and reliably priced by the market.

Conversely, factors indicative of instability or opacity tend to diminish the statistical link between our CR-enhanced sentiment and market reactions. For instance, Financial Distress (FD), interpreted here as a measure of firm distress, exhibits consistently negative and significant coefficients (e.g., from -0.265^{***} to -0.127^{***}). This suggests that in situations of higher financial distress, the specific sentiment signals from news, even if accurately attributed by CR, may be overshadowed or their relationship with returns

becomes less clear-cut. Likewise, a higher Uncertainty Level (UL) consistently correlates with lower sentiment signal effectiveness, as evidenced by its strong negative and significant linear coefficients across all models (e.g., from -0.452^{***} to -0.833^{***}). The introduction of quadratic terms for UL in Model 2 often yields a significant positive coefficient for $(UL)^2$ (e.g., 0.140^{***} in W, Model 2), suggesting a potential U-shaped or convex relationship where the negative impact of uncertainty on sentiment effectiveness might diminish or change its nature at very high levels of uncertainty.

Variables such as Industrial Sector (IS) and Trade Openness (TRADE) also show significant effects, often indicating that certain sectoral characteristics or the complexities introduced by high trade volumes can modulate how effectively sentiment signals translate into statistically robust market reactions. Exchange Rates ($ln(ER_1), ln(ER_2)$) present more complex relationships, with their linear and quadratic terms in Model 2 sometimes reaching significance, pointing towards non-linear influences.

In summary, these findings underscore that once coreference resolution provides an essential foundation by correctly attributing textual statements to their respective entities and FinBERT provides sentence-level sentiment scores, the ultimate market effectiveness or 'identifiability' of these sentiment signals is further shaped by a multifaceted interplay of historical sentiment expression, the firm's financial health and international exposure, prevailing uncertainty levels, and broader sectoral or economic conditions. CR is thus pivotal not merely for basic textual accuracy, but for enabling a more granular and nuanced subsequent analysis of these determinants of sentiment clarity.

To directly evaluate the potential practical value of the Coreference Resolution (CR) enhanced approach for downstream predictive tasks, we compared the prediction accuracy of several machine learning models using two different types of sentiment feature inputs. The 'Orig.' columns in Table 4.14 refer to the performance of the predictive models (e.g., Random Forest) when trained on sentiment features derived directly from the original news texts, without CR processing. The 'CR' columns refer to the performance

Table 4.14: Overall Accuracy Comparison: Original vs. CR-Enhanced Sentiment Inputs (%)

Model	Forecast Window Length	2020		2021		2022	
		Orig.	CR	Orig.	CR	Orig.	CR
Decision Tree	1	**39.4954**	39.2130	44.3554	**44.6210**	**55.7030**	55.5100
	2	35.2904	**35.5200**	40.5934	**40.7500**	**62.8749**	62.6300
	3	**43.9781**	43.6500	**37.9803**	37.8000	35.4467	**35.5100**
Random Forest	1	49.3658	**50.8200**	48.4512	**49.9500**	**36.6807**	36.4100
	2	61.2547	**63.0500**	65.1161	**67.2300**	55.5706	**57.1400**
	3	**45.3285**	45.1500	67.3392	**68.8500**	33.2652	**34.9600**
Gradient Boosting	1	**48.8226**	48.5300	52.8277	**53.1500**	**53.8428**	53.7100
	2	38.3814	**38.8100**	**66.0697**	65.9200	63.1274	**63.2500**
	3	**44.5560**	44.1200	39.5780	**39.8300**	**41.6134**	41.4000
Deep Forest	1	57.9512	**60.5800**	69.0657	**71.3200**	38.8935	**41.2100**
	2	**59.8429**	59.5500	43.5652	**45.1800**	45.3127	**47.6600**
	3	48.8290	**51.1300**	50.7141	**52.4500**	**36.4096**	36.3000

of the same models when trained on our CR-enhanced sentiment features. The 'Forecast Window Length' refers to the number of days after the news publication for which abnormal returns are being predicted. This allows us to assess the short-term predictive power of the sentiment signals.

The findings suggest that the benefit of CR-enhanced sentiment features tends to vary depending on the predictive model employed. For relatively simpler models, such as Decision Tree and Gradient Boosting, the advantage offered by the CR-enhanced approach appears limited and inconsistent. In numerous settings, the accuracy differences are marginal, and in several instances, features derived from the original text yield comparable or slightly better performance. This might indicate that these models do not fully leverage the nuanced information potentially unlocked by CR.

The picture becomes somewhat clearer when examining more complex ensemble models. For the Random Forest model, utilizing CR-enhanced sentiment inputs resulted in higher accuracy in a majority, though not all, of the tested settings, with performance gains being generally modest. Similarly, the Deep Forest model, while often benefiting from the CR-enhanced features (e.g., showing noticeable improvements in Window 1 across years), also exhibits instances where the original text features performed comparably or slightly better.

While these results show variability, reflecting the inherent noise and complexity in financial forecasting, the overall tendency suggests a potential advantage for employing CR-enhanced sentiment features, particularly when coupled with more sophisticated machine learning models like Random Forest and Deep Forest that might be better equipped to exploit the richer information. The observed improvements, though often modest and context-dependent, point towards the value of advanced NLP preprocessing in enhancing the inputs for downstream financial applications. This suggests that enhancements in front-end information quality through techniques like Coreference Resolution can contribute to improving the capabilities of these predictive systems, warranting further investigation.

The analysis presented in Table 4.15 is designed to evaluate the impact of Coreference Resolution (CR) on understanding spatial dynamics in financial sentiment. It meticulously compares the direct and indirect effects estimated using a Spatial Durbin Model (SDM) framework applied to both original texts and algorithmic texts (which have undergone CR processing). By examining results across a series of models (incorporating baseline regression and lagged explanatory variables) and under distinct spatial weight matrices – representing relationships based on geographic distance (W_d^H), economic linkages (W_e), geo-economic factors (W_{ge}^H), and other specified criteria (W_w^H) – we can assess how CR influences the measured interplay of factors. These factors include indicators such as industrial structure (IS), capital-labor ratio (KL), economic openness (OS), environmental/social metrics (ES), urbanization/unemployment (UR, $UR2$), technological innovation (TI), and crucially, the text-derived emotional expression (EE).

The direct effects quantify the local impact of these factors. Comparing these effects between the original and algorithmic text analyses reveals how accurately attributing sentiment via CR can refine our understanding of local influences. Similarly, the indirect effects unveil the spatial spillover – the extended influence certain factors wield on the broader network. Critically, comparing these indirect effects across the two text types allows us to gauge how CR impacts the measurement of these spatial dependencies.

Table 4.15: Analysis of Direct and Indirect Effects in Original and Algorithmic Texts in Spatial Durbin models

	Original Text				Algorithmic Text			
	Model1	Model2	Model3	Model4	Model5	Model6	Model7	Model8
Spatial Weight Matrix	W_d^H	W_e	W_{ge}^H	W_w^H	W_d^H	W_e	W_{ge}^H	W_w^H
Direct Effect								
IS	-0.03585^c	-0.02552^a	-0.0165	0.0275^c	-0.06585^a	-0.3625^a	-0.35525^a	-0.155^b
	(0.04254)	(0.0063)	(0.0023)	(0.04558)	(0.2248)	(0.7865)	(0.275)	(0.5588)
KL	-0.9855^a	-0.9525^a	-0.8585^a	-0.7695^a	-0.2489^a	-0.2255^a	-0.6925^a	-0.7725^b
	(0.1436)	(0.02546)	(0.17552)	(0.2494)	(0.2245)	(0.12235)	(0.4362)	(0.062953)
OS	-0.04865	-0.04955	-0.02658	-0.04582	-0.0782^a	-0.04955^a	-0.75485^a	0.4275^a
	(0.14455)	(0.0468)	(0.03652)	(0.01448)	(0.0365)	(0.3058)	(0.4698)	(0.4953)
ES	-0.3475^a	-0.2256^a	-0.2985^a	-0.3659^a	0.55	-0.5956^b	-0.4955^a	-0.5625^a
	(0.06956)	(0.07955)	(0.0425)	(0.004485)	(0.01425)	(0.04955)	(0.01253)	(0.2765)
UR	-0.4365^a	-0.9585^a	-0.6252^a	-0.6255^a	-0.225^a	0.0445^a	-0.9925^a	-0.9525^a
	(0.12455)	(0.2365)	(0.1959)	(0.2904)	(0.2095)	(0.07855)	(0.2459)	(2.63)
UR2	-0.4892^a	0.05425^h	-0.1285^a	0.0825^a	0.0125^a	-0.0788^a	-0.4485^a	-0.8655^a
	(0.1654)	(0.01958)	(0.06258)	(0.04555)	(0.0148)	(0.0047)	(0.6358)	(0.3624)
TI	0.03552^a	0.04265^a	-0.05265^a	0.02251^a	-0.03355^b	-0.7695^a	-0.495^a	-0.4925^b
	(0.00755)	(0.00695)	(0.00795)	(0.00455)	(0.0075)	(0.3085)	(0.0075)	(0.543)
EE	0.3654^a	0.2686^a	0.8551^a	0.36925^a	0.256^a	-1.075^a	-1.4625^a	0.2125^a
	(0.0246)	(0.0175)	(0.0699)	(0.0122)	(0.00485)	(0.7453)	(0.2258)	(0.885)
Indirect Effect								
IS	-0.495^a	-0.2756^a	-0.0854^a	-0.0258^b	-0.5255^a	0.2548^a	-0.3625^a	-0.7525^b
	(0.06485)	(0.0295)	(0.01245)	(0.0454)	(0.0955)	(0.382)	(0.2355)	(0.2525)
KL	0.5395	1.23655^a	0.6495^a	-0.9954^a	-0.2755^a	-0.8051^a	-0.9525^a	-0.7965^a
	(0.7598)	(0.0255)	(0.2356)	(0.1972)	(0.6895)	(0.01556)	(0.2033)	(0.365)
OS	0.3495	-0.0355	-0.049	-0.05429^a	-0.06825^a	-0.1525^a	-0.955^a	0.955^a
	(0.1366)	(0.06584)	(0.0448)	(0.2955)	(0.0492)	(0.4365)	(0.5553)	(0.6683)
ES	-2.6259^a	-0.2655^a	-0.2275^a	-0.4455^a	-0.7525^b	-0.6925^a	0.4475^a	-0.9525^a
	(0.5365)	(0.04855)	(0.0494)	(0.0655)	(0.2213)	(0.1664)	(0.2265)	(0.4542)
UR	2.1452^b	1.985^a	0.652^a	-0.9522^a	-0.70585^a	-0.7572^a	-0.8525^a	-0.9348^a
	(1.633)	(0.27585)	(0.3452)	(0.4955)	(0.6958)	(0.7955)	(0.269)	(0.7265)
UR2	-0.4595^b	-0.135^a	-0.2415^a	-0.2769^a	-0.3485^a	0.4785^a	-0.4635^a	-0.2245^c
	(1.3665)	(0.04485)	(0.2123)	(0.0065)	(0.3658)	(0.1556)	(0.9373)	(0.958)
TI	0.3524^a	0.026485	0.0748	0.0325^b	0.1585^a	0.6955^a	-0.31255^a	-0.9345^a
	(0.007582)	(0.0025)	(0.0044)	(0.0405)	(0.05525)	(0.4942)	(0.2755)	(0.278)
EE	0.4895^b	-0.34588	-0.3775	0.00399	0.49555^b	-0.6825^a	-0.2455^a	-1.2425^a
	(0.19355)	(0.004252)	(0.2473)	(0.0606)	(0.1785)	(0.7553)	(0.225)	(0.293)

For instance, comparing the indirect effect of Emotional Expression (EE) revealed varied impacts of CR depending on the spatial context: under the W_{H_d} matrix (Model 1 vs 5), the positive spillover effect remained statistically significant and slightly increased, while under the W_{H_w} matrix (Model 4 vs 8), a previously insignificant effect became a strong, statistically significant negative spillover after CR processing. These differing results underscore how CR, by improving data accuracy, allows for a more nuanced and context-dependent analysis of sentiment propagation.

The robustness checks using various models and spatial weight matrices

further validate these comparisons. Through this dual lens of direct and indirect effects, applied specifically to contrast results before and after CR processing, the study navigates the intricate landscape of textual analysis. This approach offers insightful perspectives on the multifaceted impacts of various factors, highlighting the added value of CR in achieving a more precise understanding of both local effects and complex spatial spillovers within financial narratives.

Table 4.16: 2022 Random Forest Financial Sentiment Prediction Monthly Accuracy Rates

Month	Decision Tree	Random Forest	Gradient Boosting	Deep Forest
1	0.523765	0.489412	0.424471	0.604235
2	0.342759	0.338307	0.457743	0.577461
3	0.521510	0.403135	0.374210	0.438578
4	0.540769	0.575384	0.428718	0.458693
5	0.481688	0.503785	0.519130	0.450895
6	0.464310	0.489007	0.526102	0.345763
7	0.359579	0.571369	0.601011	0.387369
8	0.368216	0.623084	0.406784	0.358568
9	0.404515	0.371655	0.375508	0.510969
10	0.433238	0.379943	0.570258	0.379943
11	0.477293	0.466554	0.440581	0.375302
12	0.453767	0.359024	0.316803	0.406346

Table 4.17: Quarterly Accuracy of Financial Sentiment Prediction Using Deep Forest, 2020-2022

Quarter	Decision Tree	Random Forest	Gradient Boosting	Deep Forest
2020 Q1	0.603871	0.632724	0.452903	0.368817
2020 Q2	0.431831	0.526760	0.448591	0.608381
2020 Q3	0.578934	0.510188	0.353291	0.444357
2020 Q4	0.648224	0.474277	0.569901	0.458553
2021 Q1	0.407742	0.491355	0.633635	0.534452
2021 Q2	0.507399	0.635454	0.506047	0.419556
2021 Q3	0.400581	0.398643	0.403856	0.655039
2021 Q4	0.578320	0.495189	0.470798	0.426302
2022 Q1	0.378313	0.447990	0.341240	0.519236
2022 Q2	0.368281	0.538015	0.633899	0.573390
2022 Q3	0.579007	0.423115	0.586681	0.381625
2022 Q4	0.563893	0.599917	0.376804	0.678470

Tables 4.16, 4.17, and 4.18 demonstrate varied model performances over monthly, quarterly, and lunar quarterly periods from 2020 to 2022. A notable observation is the fluctuation in accuracy rates across these models, with the Deep Forest model often showing robust performance, particularly in scenarios demanding complex pattern recognition over longer temporal spans.

Table 4.18: Quarterly Accuracy of Financial Sentiment Prediction Using Deep Forest, 2020-2022 (Lunar Calendar)

Lunar Quarter	Decision Tree	Random Forest	Gradient Boosting	Deep Forest
2020 Q1	0.632342	0.455911	0.572788	0.647398
2020 Q2	0.368763	0.621806	0.438261	0.532977
2020 Q3	0.366022	0.529140	0.448101	0.531326
2020 Q4	0.550000	0.703248	0.559140	0.601911
2021 Q1	0.396667	0.413694	0.607297	0.436959
2021 Q2	0.373541	0.486178	0.425304	0.460967
2021 Q3	0.555022	0.338021	0.407478	0.428522
2021 Q4	0.434396	0.406748	0.588629	0.509407
2022 Q1	0.382000	0.573000	0.440316	0.393105
2022 Q2	0.677727	0.600069	0.524919	0.398249
2022 Q3	0.587005	0.603554	0.579061	0.641726
2022 Q4	0.614984	0.647670	0.695637	0.407500

The Deep Forest model's peak accuracy in the lunar quarterly analysis of 2022's fourth quarter, as indicated in Table 4.18, highlights its potential in capturing the nuanced dynamics of financial markets influenced by a myriad of factors beyond traditional western calendars.

Further examination reveals the Random Forest model exhibiting a strong showing in early 2020, as seen in Table 4.17, but its performance appears to dip in certain months of 2022 according to Table 4.16. This variability underscores the challenges inherent in financial sentiment prediction, where market sentiments can drastically shift due to unforeseen global events or economic indicators. Additionally, the consistent performance of Gradient Boosting across different periods, with a notable increase in accuracy in the later lunar quarters of 2022, suggests its effectiveness in adapting to the cyclic nature of financial sentiments as influenced by the lunar calendar. These observations collectively stress the importance of model selection tailored to specific temporal and cultural contexts within the domain of financial sentiment analysis, potentially guiding more accurate and context-aware predictions in future applications.

The data presented in Table 4.19 provide a comprehensive overview of the effectiveness of Random Forest in predicting financial sentiment across various types of information. It is evident that composite data prediction outperforms both text and keyword predictions in all categories, with industry reports yielding the highest accuracy at 7.85. This suggests a significant

Table 4.19: Stepwise Analysis Table of Predictive Factors for Financial Sentiment using Random Forest

Type	Text Prediction	Composite Data Prediction	Keyword Prediction
News Articles	4.64	7.21	2.91
Company Documents	4.57	6.51	3.55
Financial Data	4.05	6.04	3.09
Industry Reports	5.69	7.85	4.15

advantage in utilizing a broad array of data points, combining qualitative and quantitative insights, for sentiment analysis. The relatively lower accuracy of keyword prediction across all types underscores the complexity of financial sentiment, which cannot be captured through simple keyword spotting alone. Interestingly, news articles and company documents, while offering rich textual content, still benefit substantially from the composite approach, indicating the nuanced nature of financial narratives. This analysis underscores the value of leveraging diverse data sources and analytical techniques to enhance the predictive accuracy of financial sentiment models, suggesting that a multi-dimensional approach is crucial for capturing the intricate dynamics of market sentiment.

4.5.1 Methodological Walkthrough: An Illustrative Case Study

This subsection provides a concrete, step-by-step illustration of the entire research methodology, from raw text to the final vector of variables used in the empirical analysis. The purpose is to demystify the data processing pipeline and demonstrate its application in a real-world scenario, thereby validating the framework's practical utility and rigor.

4.5.1.1 Case Selection and Justification

To ground the walkthrough in a tangible example, a specific company and news event have been selected based on criteria that align with the study's scope.

- **Focal Company:** Ford Motor Company (Ticker: F). Ford is a traditional, non-technology, S&P 500 constituent with extensive and continuous news coverage. It is frequently discussed alongside direct competitors, making it an ideal candidate to demonstrate the necessity of *entity-specific* sentiment analysis.

- **Comparative Company:** General Motors Company (Ticker: GM). As Ford's primary domestic rival, GM is often mentioned in the same news articles, allowing for a demonstration of how the methodology separates distinct corporate narratives within a single text.

- **Event and News Article:** The event is the public reaction to the release of full-year 2021 earnings reports. The chosen text is a news article published on **February 7, 2022**, which discusses the earnings of both Ford and GM.

4.5.1.2 Step 1: Textual Data Processing and Coreference Resolution (CR)

The process begins with the raw text from the selected news article. A key excerpt for this demonstration is:

> "Ford reported $17.9 billion in net income for the previous year, its highest since 2011... GM broke its previous record set in 2015, reporting $10 billion in net income... Despite the banner profits, however, Wall Street reacted coolly, pushing down Ford's share price over 9 percent on Friday, on news that the company fell well short of analysts' earnings-per-share forecasts."

The seq2seq transition-based model processes this text to link pronouns and mentions into coreference chains. The results are shown in Table 4.20.

4.5.1.3 Step 2: Entity-Specific Sentiment and Textual Feature Calculation

From the resolved text, the sentiment indices and textual control variables are calculated.

Table 4.20: Walkthrough - Coreference Resolution and Sentiment Attribution Example

Original Sentence	**Identified Mentions**	**Resolved Coreference Chain (Entity)**	**FinBERT Sentiment Score (Positive/Negative/Neutral)**
"Ford reported $17.9 billion in net income for the previous year, its highest since 2011..."	'Ford', 'its'	**Ford Motor Company**	0.85 / 0.05 / 0.10
"...GM broke its previous record set in 2015, reporting $10 billion in net income..."	'GM', 'its'	**General Motors**	0.82 / 0.06 / 0.12
"...pushing down Ford's share price over 9 percent on Friday, on news that the company fell well short of analysts' earnings-per-share forecasts."	'Ford's', 'the company'	**Ford Motor Company**	0.03 / 0.91 / 0.06

Sentiment Indices ('RTY' and 'WERT') The aggregated sentiment probabilities for Ford are calculated by averaging the scores from the two relevant sentences:

- Positive Score Probability: $(0.85 + 0.03)/2 = 0.44$
- Negative Score Probability: $(0.05 + 0.91)/2 = 0.48$

To align with the scaling used in the main empirical analysis, these probabilities are multiplied by 100. Thus, for this article, **'RTY' = 44.0%** and **'WERT' = 48.0%**.

Textual Control Variables ('UUYTR' and 'YYUTR') These variables are calculated from the raw text of the excerpt (65 tokens):

- **'UUYTR' (Numerical Density):** The excerpt contains 8 numerical tokens ("17.9", "billion", "2011", "10", "billion", "2015", "9", "percent"). The density is $8/65 \approx 0.123$, or **12.3%**.

- **'YYUTR' (Coreference Complexity):** The excerpt contains 2 pronouns ("its", "its") that require resolution. The complexity factor is $2/65 \approx 0.031$, or **3.1%**.

4.5.1.4 Step 3: Financial Data Extraction and Variable Construction

Next, financial data for February 7, 2022, is gathered to construct the remaining variables.

Dependent Variable ('AQTB') The abnormal return is calculated using the market model, $AR_{it} = R_{it} - (\hat{\alpha}_i + \hat{\beta}_i R_{mt})$, with parameters estimated over a 250-day window. For Ford, this yielded $\hat{\alpha}_F = 0.0005$ and $\hat{\beta}_F = 1.25$.

Table 4.21: Walkthrough - Financial Data Calculation for Ford (F) on Feb 7, 2022

Variable	Symbol	Value	Source / Calculation
Ford's Daily Return	R_{it}	2.05%	(Close: 17.93 - Prev. Close: 17.57) / 17.57
S&P 500 Daily Return	R_{mt}	-0.37%	(Close: 4483.87 - Prev. Close: 4500.53) / 4500.53
Expected Return	$E(R_{it})$	-0.41%	$0.0005 + 1.25 \times (-0.0037)$
Abnormal Return	**'AQTB'**	**2.46%**	$0.0205 - (-0.0041)$

Financial Control and Mediator Variables The remaining variables are constructed as follows:

- **'SE' (Degree of Bull Market):** Defined as the ratio of the S&P 500's closing price to its 252-day moving average, expressed as a percentage. On Feb 7, 2022, the S&P 500 closed at 4483.87, and its 252-day moving average was approx. 4320. Thus, 'SE' = (4483.87 / 4320) * 100 = **103.8%**.

- **Size (Firm Size):** The natural logarithm of Ford's market capitalization at year-end 2021 ($83.00 billion). Thus, Size = $\ln(83,000,000,000) = 25.14$.

- **B/M (Book-to-Market Ratio):** The reciprocal of the Price-to-Book ratio. For fiscal year 2021, Ford's P/B was 2.3, so B/M = 1/2.3 = 0.43.
- **'POIUQ' (Mediator):** Calculated as 'RTY' multiplied by abnormal trading volume. Assuming Ford's abnormal volume on this day was 1.2 (i.e., 20% above average), then 'POIUQ' = $44.0 \times 1.2 =$ **52.8%**.

4.5.1.5 Step 4: Data Integration and Forecast Windows

The final step is to assemble the complete data vector for the observation and define the forecast windows for the dependent variable. Table 4.22 shows the integrated data vector for Ford on the event date.

Table 4.22: Walkthrough - Final Assembled Data Vector for Ford (F)

Firm ID	Date	'RTY'	'WERT'	'SE'	Size	B/M	...	'AQTB' (t)
F	2022-02-07	0.44	0.48	1	25.14	0.43	...	2.46%

The analysis in this thesis examines the predictive power of sentiment over different time horizons. This is accomplished by shifting the dependent variable to future dates.

- **Forecast Window 1 (1-day):** The dependent variable is the abnormal return on the next trading day, $AQTB_{t+1}$. On Feb 8, 2022, Ford's abnormal return was -0.38%.
- **Forecast Window 2 (2-day):** The dependent variable is the Cumulative Abnormal Return (CAR) over the next two days, $CAR(t+1, t+2) = AQTB_{t+1} + AQTB_{t+2}$. This was calculated as $-0.38\% + (-2.86\%) = -3.24\%$.
- **Forecast Window 3 (3-day):** The dependent variable is the CAR over the next three days, $CAR(t+1, t+3) = AQTB_{t+1} + AQTB_{t+2} + AQTB_{t+3}$. This was calculated as $-0.38\% + (-2.86\%) + 0.59\% = -2.65\%$.

This walkthrough example demonstrates the rigorous, multi-stage process of transforming unstructured news text into a structured dataset suitable for sophisticated econometric analysis, bridging the gap between advanced NLP and empirical finance.

4.6 Conclusions

Our extensive empirical analysis, incorporating a diverse set of models and methodologies, has culminated in several pivotal findings within the domain of financial sentiment analysis. Through the meticulous examination of financial news across various platforms, including news articles, company documents, and industry reports, we have delineated the intricate relationship between textual sentiment and financial market movements. The utilization of advanced machine learning algorithms, notably Random Forest and Deep Forest models, has been instrumental in deciphering the nuanced interplay between sentiment and market performance, offering novel insights into predictive accuracy dynamics.

Our findings reveal a notable variance in model performance across different periods, with the Deep Forest model demonstrating exceptional proficiency, particularly in 2021, where it achieved a peak accuracy of 0.690657. This highlights the model's superior capability in capturing complex patterns within financial sentiment data, suggesting its potential for offering nuanced insights into market dynamics.

Moreover, our analysis has unveiled the strategic significance of employing a composite data approach in sentiment prediction. The stepwise analysis revealed that composite data prediction consistently outperforms both text and keyword predictions across all information types, with industry reports achieving the highest accuracy. This emphasizes the superior efficacy of integrating multifaceted data sources, blending qualitative and quantitative insights, to navigate the complexities inherent in financial sentiment.

Additionally, the variability observed in the Random Forest model's performance across different periods, particularly in 2022, elucidates the challenges posed by market volatility and unforeseen global events. This variability accentuates the necessity for adaptive and sophisticated predictive models capable of adjusting to rapid market changes and external economic indicators.

In summary, our comprehensive study significantly advances the understanding of financial sentiment analysis, highlighting the paramount importance of model selection tailored to specific temporal and cultural contexts. The insights gleaned from our research not only validate the effectiveness of advanced machine learning techniques in financial sentiment prediction but also pave the way for future explorations aimed at refining these predictive models further. As we continue to delve into the complexities of financial markets, the integration of nuanced analytical approaches will undoubtedly play a crucial role in enhancing the accuracy and applicability of financial sentiment analysis, offering a robust framework for informed decision-making in the ever-evolving landscape of financial analytics.

Chapter 5

Conclusions and Future Directions

5.1 Conclusions

This doctoral thesis has made significant contributions to the field of Coreference Resolution and its applications in the financial domain. The comprehensive survey presented in Chapter 2 provided a valuable overview of the development trends in Coreference Resolution over the past decade, categorizing models into feature-based, neural network-based, knowledge-based, and transformer-based approaches. This survey serves as a valuable resource for researchers and practitioners, offering insights into the evolution of Coreference Resolution techniques and the current state-of-the-art methods.

In Chapter 3, we introduced a novel Multi-task Learning model for Gold-two-mention Coreference Resolution, which outperformed existing methods on multiple datasets by jointly learning mention identification and linking tasks. The implementation of a dynamic weight balancing mechanism in the co-reference resolver allowed for adaptive balancing between the two tasks during training, optimizing the model's performance and generalizability across different datasets and domains. The superiority of the Multi-task Learning approach over traditional single-task learning methods was demonstrated, highlighting the benefits of leveraging the interdependencies between mention identification and linking tasks to improve overall performance.

Chapter 4 presented a comprehensive investigation into the application of Coreference Resolution techniques within the financial domain, focusing on their integration with financial sentiment analysis. Using advanced language models, such as FinBERT, and machine learning techniques, we improved the accuracy and granularity of sentiment attribution in financial texts. The introduction of novel analytical techniques, such as mediation effect regression, uncovered the underlying mechanisms through which sentiment influenced financial performance. Rigorous robustness checks, including DID regression from post-financial sentiment analysis and multivariate regression, validated the stability and reliability of our findings. The quantitative analysis of financial texts post-coreference resolution, utilizing advanced models to investigate the interaction of textual factors and their impact on sentiment identifiability, further demonstrated the critical role of Coreference Resolution in improving the accuracy and depth of financial sentiment analysis. The evaluation of financial sentiment prediction models, particularly the Random Forest model, across original and algorithmically processed texts highlighted the enhanced predictive capabilities of our proposed methodological framework.

The contributions of this thesis advance the field of Coreference Resolution and demonstrate its practical applications in the financial domain. The proposed Multi-task Learning model and the dynamic weight balancing mechanism offer novel approaches to improving the performance of Coreference Resolution systems. The in-depth exploration of Coreference Resolution's application in financial sentiment analysis contributes to a deeper understanding of the role of natural language processing in financial analysis and offers new perspectives on leveraging linguistic techniques for more accurate and comprehensive market assessments. These findings pave the way for future research and development in Coreference Resolution and its applications in various domains, including finance.

5.2 Prospective Future Directions

Coreference Resolution (CR), while having undergone significant advancements in recent years, continues to present a complex landscape of challenges

and potential advancements in both the academic realm and practical applications.

5.2.1 Enhancing CR Models with Linguistic and Cognitive Insights

Incorporating linguistic theories and cognitive insights into deep learning-based CR models could lead to more linguistically informed and cognitively plausible systems. Future research could explore the integration of knowledge about discourse salience[Miltsakaki, 2007], and other language-specific phenomena into the multi-task learning framework proposed in this thesis. Additionally, investigating the role of pragmatics and world knowledge in resolving complex coreference chains could further enhance the performance of CR models, particularly in scenarios that require common sense reasoning and context-aware interpretation.

5.2.2 Adapting CR Models to Low-Resource Languages and Domains

Developing CR models for low-resource languages and domain-specific applications remains a challenge. Future research could investigate techniques such as cross-lingual transfer learning, few-shot learning, and unsupervised or semi-supervised approaches to adapt CR models to these settings. Moreover, exploring the potential of multilingual pre-trained language models and their adaptation to low-resource languages could open up new possibilities for developing robust and efficient CR systems in resource-constrained environments. Additionally, investigating the effectiveness of domain adaptation techniques, such as adversarial learning and domain-specific fine-tuning, could help bridge the gap between general-purpose CR models and their application in specialized domains.

5.2.3 Advancing the Integration of CR with Other NLP Tasks

Recent research has demonstrated the benefits of integrating CR with various NLP tasks, such as named entity recognition[Martins et al., 2019], dia-

logue systems[Xu and Choi, 2022], machine translation[Yehudai et al., 2023] and text summarization[Liu et al., 2021]. Future work could focus on developing innovative architectures, training strategies, and evaluation metrics to effectively integrate CR with a wide range of NLP tasks. Additionally, exploring the potential of joint learning frameworks that simultaneously optimize CR and downstream tasks could lead to more efficient and effective models, reducing the need for cascading errors and enabling seamless integration of CR in real-world applications. Furthermore, investigating the impact of CR on the interpretability and explainability of downstream NLP models could provide valuable insights into the role of coreference information in various language understanding tasks.

5.2.4 Improving the Interpretability and Explainability of CR Models

As CR models become increasingly complex, there is a growing need for interpretable and explainable models that can provide insights into their decision-making process. Future research could focus on developing techniques to visualize and interpret the attention mechanisms and feature representations learned by CR models, enhancing their trustworthiness and facilitating their deployment in real-world applications. This could involve the development of novel visualization tools, the application of model-agnostic interpretation methods, and the integration of human-in-the-loop approaches to refine and validate the interpretations provided by CR models. Furthermore, investigating the relationship between model interpretability and performance could help strike a balance between transparency and accuracy in CR systems.

5.2.5 Scaling CR Models for Large-Scale Applications

Investigating the scalability of CR models for handling massive amounts of textual data is essential, given the advent of large-scale pre-trained language models. Future work could explore techniques for efficient training and inference of CR models on large-scale datasets, such as knowledge distillation,

model compression, and distributed computing. Additionally, research into memory-efficient architectures, such as sparse attention mechanisms and hierarchical processing, could help alleviate the computational burden associated with processing long documents and large corpora. Exploring the potential of edge computing and federated learning could also enable the deployment of CR models in resource-constrained environments and facilitate their application in privacy-sensitive scenarios.

5.2.6 The Role of Coreference Resolution in the LLM Era

The recent and rapid proliferation of Large Language Models (LLMs) such as GPT, LLaMA, and PaLM has fundamentally reshaped the landscape of Natural Language Processing. These models exhibit remarkable zero- and few-shot learning capabilities across a wide array of tasks, including a strong conceptual understanding of coreference. However, the assertion that LLMs render specialized, explicit Coreference Resolution (CR) systems obsolete is premature[Shore et al., 2025]. Instead, the future of CR lies in the synergistic integration of traditional, structured approaches with the broad world knowledge and generative power of LLMs. Several key challenges inherent to current LLMs underscore the continued relevance of dedicated CR mechanisms and point toward promising avenues for future research.

A primary challenge is the issue of reliability and hallucination. LLMs are known to generate plausible but factually incorrect or inconsistent outputs, a significant problem when applying them to structured prediction tasks like CR [Bang et al., 2025]. For high-stakes domains such as the financial analysis explored in this thesis, where precision and factual grounding are paramount, the risk of an LLM inventing a coreferential link or misattributing a statement is unacceptably high. Dedicated CR systems, especially state-of-the-art seq2seq models, are trained specifically on the structured task of forming valid coreference chains and are less prone to such ungrounded generation.

Furthermore, LLMs face practical limitations regarding context length and consistency. While context windows are expanding, processing extremely long documents—such as annual financial reports, legal contracts,

or literary texts—remains a challenge. Specialized CR systems are often designed with mechanisms to handle long-range dependencies that can span thousands of tokens, a domain where LLMs may still lose track of referential chains[Liu et al., 2025]. Another significant issue is prompt sensitivity. The performance of an LLM on a CR task can vary dramatically based on the specific phrasing of the input prompt, leading to brittle and inconsistent results [Gan et al., 2025]. This contrasts with fine-tuned, task-specific models that offer more robust and predictable performance. Finally, the generative nature of LLM outputs complicates evaluation using traditional, link-based CR metrics like the CoNLL F1 score, necessitating the development of new evaluation paradigms to accurately assess their capabilities.

These challenges illuminate several promising future research directions that build upon the contributions of this thesis:

- **Hybrid Architectures:** A compelling avenue is the development of hybrid systems that combine the strengths of both approaches. A structured, reliable CR model could first generate a high-precision scaffold of entity chains within a document. An LLM could then be used to reason over this structured output, enriching it with its vast world knowledge to resolve highly ambiguous cases or perform downstream tasks like relation extraction with greater accuracy. This leverages the reliability of the specialized model and the reasoning capacity of the LLM [Chun et al., 2025].

- **LLMs for Data Augmentation and Resource Creation:** The generative power of LLMs can be harnessed to address the long-standing data bottleneck in CR. LLMs can be prompted to generate vast quantities of high-quality, diverse synthetic training data for coreference, a particularly valuable strategy for low-resource languages or specialized domains where annotated data is scarce [Ding et al., 2024].

- **Advanced Prompting and Fine-Tuning Strategies:** Research is needed to move beyond simple question-answering or infilling prompts for CR. Future work could explore structured prompting techniques

or fine-tuning methods that explicitly teach LLMs to generate outputs in a format that is directly compatible with standard CR evaluation, thereby bridging the gap between generative capabilities and structured prediction [Arslan et al., 2025].

In conclusion, the LLM era does not signify the end of coreference resolution as a distinct research field. Rather, it reframes its role. The principles of rigorous evaluation, task-specific architectural design, and the importance of structured linguistic representation, which are central to this thesis, remain critically relevant. The future of CR will likely involve leveraging LLMs as powerful components within more complex, reliable, and interpretable systems, ensuring that the precision required for real-world applications is not sacrificed for generative flexibility.

Appendix A

Definition and Classification of Empirical Variables

This appendix provides precise operational definitions, data sources, and model roles for all variables used in the empirical analysis in Chapter 4, supplementing the information in Table 4.1. This serves to clarify the construction of the dataset and the underlying causal framework of the study.

Table A.1: Glossary and Classification of Empirical Variables

Variable	Description	Source(s)	Role in Model
'AQTB'	**Abnormal Return.** The component of a stock's daily return not explained by market-wide movements. It is the standard measure for isolating firm-specific price changes in response to new information (events).	CRSP (for daily stock and market returns).	**Dependent (Target) Variable**
'RTY'	**Financial Positive Index.** The aggregated entity-specific positive sentiment score derived from news text, representing the strength of positive news content, expressed as a percentage (0-100).	Factiva (text), FinBERT (model).	**Independent Variable**
'WERT'	**Fiscal Negative Index.** The aggregated entity-specific negative sentiment score derived from news text, representing the strength of negative news content, expressed as a percentage (0-100).	Factiva (text), FinBERT (model).	**Independent Variable**
'SE'	**Degree of Bull Market.** A continuous variable capturing the prevailing market trend, calculated as the ratio of the S&P 500's closing price to its 252-day moving average, expressed as a percentage.	CRSP.	**Control Variable**
Size	**Firm Size.** The natural logarithm of the firm's market capitalization. Controls for the well-known size effect in asset pricing.	CRSP/Compustat.	**Control Variable**

Continued on next page

Table A.1: Glossary and Classification of Empirical Variables (Continued)

Variable	Description	Source(s)	Role in Model
B/M	**Book-to-Market Ratio.** A valuation ratio comparing a company's book value to its market value. Controls for the value effect in asset pricing.	Compustat (Book Value), CRSP (Market Cap).	**Control Variable**
'POIUQ'	**Economic Benefits from Positive Index.** A measure of market reaction or attention, calculated as the product of 'RTY' and the abnormal trading volume ratio for the event day, expressed as a percentage.	Factiva, FinBERT, CRSP.	**Mediator Variable**
'UUYTR'	**Numerical Density of Corpus.** The proportion of tokens in a news article that are numerical, expressed as a percentage. This controls for the style of reporting (e.g., fact-based vs. narrative-driven).	Factiva.	**Control Variable**
'YYUTR'	**Coreference Complexity Factor.** A measure of textual complexity related to referential ambiguity, calculated as the density of pronouns within the text, expressed as a percentage. Higher values indicate a greater need for CR.	Factiva.	**Control Variable**
'Age'	**Firm Age.** The number of years the firm has been listed in the CRSP database. Controls for life-cycle effects where younger firms may be more volatile or receive different types of media coverage.	CRSP.	**Control Variable**

Discussion of Variable Roles and the Causal Framework

The classification of variables in Table A.1 codifies the empirical strategy of this thesis. The framework is designed to test a primary causal hypothesis while accounting for established confounding factors and exploring potential mechanisms.

- **Primary Causal Hypothesis:** The study's central hypothesis posits a causal relationship where the entity-specific sentiment expressed in news (the **independent variables**, 'RTY' and 'WERT') influences the firm's subsequent abnormal stock performance (the **dependent variable**, 'AQTB'). A positive coefficient is expected for 'RTY' and a negative coefficient for 'WERT', indicating that positive news leads to positive abnormal returns and vice versa.

- **Justification for Control Variables:** The financial econometrics literature has identified several firm characteristics that systematically explain stock returns. To isolate the unique impact of the sentiment signal, it is essential to include these factors as **control variables**. The inclusion of 'Size' and 'B/M' accounts for the well-documented size and value effects in asset pricing [Fama and French, 1993]. The 'SE' variable controls for the broader market environment, while 'UUYTR', 'YYUTR', and 'Age' control for characteristics of the text and the firm that could otherwise confound the relationship between sentiment and returns.

- **Role of Mediator Variables:** The mediation effect regression models presented in Section 4.5 (Table 4.7) explore the pathways through which sentiment may operate. In these models, variables such as 'POIUQ' are treated as potential **mediator variables**. This framework tests whether the effect of sentiment (X) on abnormal returns (Y) is channeled through these intermediate factors (M). For instance, a significant mediation effect for 'POIUQ' would suggest that positive news does not impact returns directly, but rather by first generating unusually high trading volume, which then leads to price changes. This analysis moves beyond asking *if* sentiment has an effect, to asking *how* it exerts its influence on the market.

Bibliography

Artem Abzaliev. On GAP coreference resolution shared task: Insights from the 3rd place solution. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 107–112, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3816. URL https://www.aclweb.org/anthology/W19-3816.

Oshin Agarwal, Sanjay Subramanian, Ani Nenkova, and Dan Roth. Evaluation of named entity coreference. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 1–7, Minneapolis, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2801. URL https://aclanthology.org/W19-2801.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1034. URL https://www.aclweb.org/anthology/P15-1034.

Rahul Aralikatte, Heather Lent, Ana Valeria Gonzalez, Daniel Herschcovich, Chen Qiu, Anders Sandholm, Michael Ringaard, and Anders Søgaard. Rewarding coreference resolvers for being consistent with world knowledge. 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 1229–1235, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1118. URL https://aclanthology.org/D19-1118.

Tuğba Pamay Arslan, Emircan Erol, and Gülşen Eryiğit. Corefinst: Leveraging llms for multilingual coreference resolution, 2025. URL https://arxiv.org/abs/2509.17505.

John Atkinson, Gonzalo Salas, and Alejandro Figueroa. Improving opinion retrieval in social media by combining features-based coreferencing and memory-based learning. *Information Sciences*, 299: 20 – 31, 2015a. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2014.12.021. URL http://www.sciencedirect.com/science/article/pii/S0020025514011608.

John Atkinson, Gonzalo Salas, and Alejandro Figueroa. Improving opinion retrieval in social media by combining features-based coreferencing and memory-based learning. *Information Sciences*, 299:20–31, 2015b.

Sandeep Attree. Gendered ambiguous pronouns shared task: Boosting model confidence by evidence pooling. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 134–146, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3820. URL https://aclanthology.org/W19-3820.

Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer, 1998.

David Bamman, Olivia Lewke, and Anya Mansoor. An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.6.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. HalluLens: LLM hallucination benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1176. URL https://aclanthology.org/2025.acl-long.1176/.

Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020. URL https://arxiv.org/abs/2004.05150.

Santanu Bhattacharjee, Rejwanul Haque, Gideon Maillette de Buy Wenniger, and Andy Way. Investigating query expansion and coreference resolution in question answering on bert. In *International conference on applications of natural language to information systems*, pages 47–59. Springer, 2020.

Bernd Bohnet, Chris Alberti, and Michael Collins. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226, 2023. doi: 10.1162/tacl_a_00543. URL https://aclanthology.org/2023.tacl-1.13.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic, November 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.225. URL https://aclanthology.org/2021.findings-emnlp.225.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E. Peters, Arie Cattan, and Ido Dagan. Cross-document language modeling. *CoRR*, abs/2101.00406, 2021b. URL https://arxiv.org/abs/2101.00406.

David Card and Alan B Krueger. Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *The American Economic Review*, 84(4):772–793, 1994.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. Streamlining cross-document coreference resolution: Evaluation and modeling. *CoRR*, abs/2009.11032, 2020. URL https://arxiv.org/abs/2009.11032.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. Cross-document coreference resolution over predicted mentions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5100–5107, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.453. URL https://aclanthology.org/2021.findings-acl.453.

Craig G Chambers and Ron Smyth. Structural parallelism and discourse coherence: A test of centering theory. *Journal of Memory and Language*, 39(4):593–608, 1998.

Guanyi Chen, Kees Van Deemter, and Chenghua Lin. Modelling pro-drop with the rational speech acts model. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 57–66. Association for Computational Linguistics (ACL), 2018.

Changwoo Chun, Daniel Rim, and Juhee Park. LLM ContextBridge: A hybrid approach for intent and dialogue understanding in IVSR. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert, Kareem Darwish, and Apoorv Agarwal, editors, *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 794–806, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-industry.66/.

Kevin Clark and Christopher D. Manning. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1136. URL https://aclanthology.org/P15-1136.

Kevin Clark and Christopher D. Manning. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1245. URL https://aclanthology.org/D16-1245.

Agata Cybulska and Piek Vossen. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/840_Paper.pdf.

Zeyu Dai, Hongliang Fei, and Ping Li. Coreference aware representation learning for neural named entity recognition. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4946–4953. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/687. URL https://doi.org/10.24963/ijcai.2019/687.

Parag Pravin Dakle, Takshak Desai, and Dan Moldovan. A study on entity resolution for email conversations. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 65–73, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.8.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. 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 5925–5932, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1606. URL https://aclanthology.org/D19-1606.

Hal Daumé III and Daniel Marcu. A large-scale exploration of effective global features for a joint entity detection and tracking model. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 97–104, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL https://aclanthology.org/H05-1013.

Ernest Davis, Leora Morgenstern, and Charles L. Ortiz. The first winograd schema challenge at ijcai-16. *AI Magazine*, 38(3):97–98, September 2017. ISSN 0738-4602. doi: 10.1609/aimag.v38i4.2734. Publisher Copyright: © 2017, Association for the Advancement of Artificial Intelligence.

Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pages 1039–1047, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL https://aclanthology.org/P08-1118.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. Data augmentation using LLMs: Data perspectives, learning paradigms and challenges. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1679–1705, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.97. URL https://aclanthology.org/2024.findings-acl.97/.

Xiaowen Ding and Bing Liu. Resolving object and attribute coreference in opinion mining. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 268–276, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL https://www.aclweb.org/anthology/C10-1031.

Vladimir Dobrovolskii. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.605. URL https://aclanthology.org/2021.emnlp-main.605.

Greg Durrett and Dan Klein. A Joint Model for Entity Analysis: Coreference, Typing, and Linking. *Transactions of the Association for Computational Linguistics*, 2:477–490, 11 2014a. ISSN 2307-387X. doi: 10.1162/tacl_a_00197. URL https://doi.org/10.1162/tacl_a_00197.

Greg Durrett and Dan Klein. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490, 2014b. doi: 10.1162/tacl_a_00197. URL https://aclanthology.org/Q14-1037.

Alon Eirew, Arie Cattan, and Ido Dagan. WEC: Deriving a large-scale cross-document event coreference dataset from Wikipedia. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2498–2510, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.198. URL https://www.aclweb.org/anthology/2021.naacl-main.198.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M. Strassel. Overview of linguistic resources for the TAC KBP 2015 evaluations: Methodologies and results. In *Proceedings of the 2015 Text Analysis Conference, TAC 2015, Gaithersburg, Maryland, USA, November 16-17, 2015, 2015*. NIST, 2015. URL https://tac.nist.gov/publications/2015/additional.papers/TAC2015.KBP_resources_overview.proceedings.pdf.

Joe Ellis, Jeremy Getman, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M. Strassel. Overview of linguistic resources for the TAC KBP 2016 evaluations: Methodologies and results. In *Proceedings of the 2016 Text Analysis Conference, TAC 2016, Gaithersburg, Maryland, USA, November 14-15, 2016*. NIST, 2016. URL https://tac.nist.gov/publications/2016/additional.papers/TAC2016.KBP_resources_overview.proceedings.pdf.

Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. A generalized knowledge hunting framework for the Winograd schema challenge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 25–31, New Orleans, Louisiana, USA, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-4004. URL https://www.aclweb.org/anthology/N18-4004.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1386. URL https://www.aclweb.org/anthology/P19-1386.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1102. URL https://aclanthology.org/P19-1102.

Eugene F Fama and Kenneth R French. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56, 1993.

Elisa Ferracane, Iain Marshall, Byron C. Wallace, and Katrin Erk. Leveraging coreference to identify arms in medical abstracts: An experimental study. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis*, pages 86–95, Auxtin, TX, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-6112. URL https://aclanthology.org/W16-6112.

Yujian Gan, Massimo Poesio, and Juntao Yu. Assessing the capabilities of large language models in coreference: An evaluation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1645–1665, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.145/.

Yujian Gan, Yuan Liang, Yanni Lin, Juntao Yu, and Massimo Poesio. Improving LLMs' learning of coreference resolution. In Frédéric Béchet, Fabrice Lefèvre, Nicholas Asher, Seokhwan Kim, and Teva Merlin, editors, *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 311–321, Avignon, France, August 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.sigdial-1.25/.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 2017.

Abbas Ghaddar and Phillippe Langlais. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1021.

Aaryan Gupta, Vinya Dengre, Hamza Abubakar Kheruwala, and Manan Shah. Comprehensive review of text-mining applications in finance. *Financial Innovation*, 6(1):1–25, 2020.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA, June 2006. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N06-2015.

Yin Jou Huang, Jing Lu, Sadao Kurohashi, and Vincent Ng. Improving event coreference resolution by learning argument compatibility from unlabeled data. 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)*, pages 785–795, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1085. URL https://aclanthology.org/N19-1085.

Jianguo Jiang, Jiuming Chen, Tianbo Gu, Kim-Kwang Raymond Choo, Chao Liu, Min Yu, Weiqing Huang, and Prasant Mohapatra. Anomaly detection with graph convolutional networks for insider threat and fraud detection. In *MILCOM 2019-2019 IEEE Military Communications Conference (MILCOM)*, pages 109–114. IEEE, 2019.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. BERT for coreference resolution: Baselines and analysis. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 5803–5808, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1588. URL https://aclanthology.org/D19-1588.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020. URL https://aclanthology.org/2020.tacl-1.5.

Colm Kearney and Sha Liu. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185, 2014.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1023. URL https://aclanthology.org/N18-1023.

Sopan Khosla and Carolyn Rose. Using type information to improve entity coreference resolution. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 20–31, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.codi-1.3. URL https://aclanthology.org/2020.codi-1.3.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Yuval Kirstain, Ori Ram, and Omer Levy. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 14–19, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.3. URL https://aclanthology.org/2021.acl-short.3.

Vid Kocijan, Oana-Maria Camburu, Ana-Maria Cretu, Yordan Yordanov, Phil Blunsom, and Thomas Lukasiewicz. WikiCREM: A large unsupervised corpus for coreference resolution. 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 4303–4312, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1439. URL https://www.aclweb.org/anthology/D19-1439.

Kelvin Leonardi Kohsasih, B. Herawan Hayadi, Robet, Carles Juliandy, Octara Pribadi, and Andi. Sentiment analysis for financial news using rnn-lstm network. In *2022 4th International Conference on Cybernetics and Intelligent System (ICORIS)*, pages 1–6, 2022. doi: 10.1109/ICORIS56080.2022.10031595.

M. Hari Krishna, K. Rahamathulla, and Ali Akbar. A feature based approach for sentiment analysis using svm and coreference resolution. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 397–399, 2017. doi: 10.1109/ICICCT.2017.7975227.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Gourab Kundu, Avi Sil, Radu Florian, and Wael Hamza. Neural cross-lingual coreference resolution and its application to entity linking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 395–400, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2063. URL https://aclanthology.org/P18-2063.

Tuan Lai, Heng Ji, Trung Bui, Quan Hung Tran, Franck Dernoncourt, and Walter Chang. A context-dependent gated module for incorporating symbolic semantics into event coreference resolution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3491–3499, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.274. URL https://aclanthology.org/2021.naacl-main.274.

Tuan Manh Lai, Trung Bui, and Doo Soon Kim. End-to-end neural coreference resolution revisited: A simple yet effective baseline. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8147–8151, 2022. doi: 10.1109/ICASSP43922.2022.9746254.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2019.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1018. URL https://www.aclweb.org/anthology/D17-1018.

Kenton Lee, Luheng He, and Luke Zettlemoyer. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2108. URL https://aclanthology.org/N18-2108.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

Hector J. Levesque. The winograd schema challenge. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI, 2011. URL http://www.aaai.org/ocs/index.php/SSS/SSS11/paper/view/2502.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.211. URL https://aclanthology.org/2021.findings-emnlp.211.

Xiao Li, Kees Van Deemter, and Chenghua Lin. Statistical NLG for generating the content and form of referring expressions. In *Proceedings of the 11th International Conference on Natural Language Generation*. Association for Computational Linguistics (ACL), 2018.

Qika Lin, Rui Mao, Jun Liu, Fangzhi Xu, and Erik Cambria. Fusing topology contexts and logical rules in language models for knowledge graph completion. *Information Fusion*, 90:253–264, 2023. ISSN 1566-2535.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.713. URL https://aclanthology.org/2020.acl-main.713.

Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, 56(12):14439–14481, 2023.

Yanming Liu, Xinyue Peng, Jiannan Cao, Yanxin Shen, Tianyu Du, Sheng Cheng, Xun Wang, Jianwei Yin, and Xuhong Zhang. Bridging context gaps: Leveraging coreference resolution for long contextual understanding, 2025. URL https://arxiv.org/abs/2410.01671.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019a. URL http://arxiv.org/abs/1907.11692.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b.

Zhengyuan Liu, Ke Shi, and Nancy Chen. Coreference-aware dialogue summarization. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–519, Singapore and Online, July 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.sigdial-1.53.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. Finbert: A pre-trained financial language representation model for financial text mining. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4513–4519. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/622. URL https://doi.org/10.24963/ijcai.2020/622. Special Track on AI in FinTech.

Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65, 2011.

Tim Loughran and Bill McDonald. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230, 2016.

Jing Lu and Vincent Ng. Event coreference resolution: A survey of two decades of research. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5479–5486. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/773. URL https://doi.org/10.24963/ijcai.2018/773.

Jing Lu and Vincent Ng. Conundrums in entity coreference resolution: Making sense of the state of the art. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6620–6631, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.536. URL https://www.aclweb.org/anthology/2020.emnlp-main.536.

Jing Lu and Vincent Ng. Constrained multi-task learning for event coreference resolution. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4504–4514, Online, June 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.356. URL https://www.aclweb.org/anthology/2021.naacl-main.356.

Jing Lu and Vincent Ng. Conundrums in event coreference resolution: Making sense of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1368–1380, Online and Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.103. URL https://aclanthology.org/2021.emnlp-main.103.

Yaojie Lu, Hongyu Lin, Jialong Tang, Xianpei Han, and Le Sun. End-to-end neural event coreference resolution. *Artificial Intelligence*, 303:103632, 2022. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103632. URL https://www.sciencedirect.com/science/article/pii/S0004370221001831.

Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/H05-1004.

Xiaoqiang Luo and Sameer Pradhan. Evaluation metrics. In *Anaphora Resolution*, pages 141–163. Springer Berlin Heidelberg, 2016. doi: 10.1007/978-3-662-47909-4_5. URL https://doi.org/10.1007%2F978-3-662-47909-4_5.

Rui Mao and Xiao Li. Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13534–13542, 2021.

Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. Joint learning of named entity recognition and entity linking. In Fernando Alva-Manchego, Eunsol Choi, and Daniel Khashabi, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 190–196, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-2026. URL https://aclanthology.org/P19-2026.

Eleni Miltsakaki. A rethink of the relationship between salience and anaphora resolution. In *Proceedings of the 6th discourse anaphora and anaphor resolution colloquium*, pages 91–96, 2007.

Teruko Mitamura, Zhengzhong Liu, and Eduard H. Hovy. Overview of TAC-KBP 2016 event nugget track. In *Proceedings of the 2016 Text Analysis Conference, TAC 2016, Gaithersburg, Maryland, USA, November 14-15, 2016*. NIST, 2016. URL https://tac.nist.gov/publications/2016/additional.papers/TAC2016.KBP_Event_Nugget_overview.proceedings.pdf.

Teruko Mitamura, Zhengzhong Liu, and Eduard H Hovy. Events detection, coreference and sequencing: What's next? overview of the tac kbp 2017 event track. In *TAC*, 2017.

Ruslan Mitkov. *Anaphora Resolution: The State of The Art*. Citeseer, 1999.

Nafise Sadat Moosavi and Michael Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1060. URL https://www.aclweb.org/anthology/P16-1060.

Jesús Mur and Ana Angulo. The spatial durbin model and the common factor tests. *Spatial Economic Analysis*, 1(2):207–226, 2006.

Vincent Ng. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-1142.

James R Norris. *Markov chains*. Number 2. Cambridge university press, 1998.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5706. URL https://www.aclweb.org/anthology/W16-5706.

Haoruo Peng, Kai-Wei Chang, and Dan Roth. A joint framework for coreference resolution and mention head detection. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 12–21, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/K15-1002. URL https://aclanthology.org/K15-1002.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

Massimo Poesio, Roland Stuckardt, and Yannick Versley. *Anaphora Resolution - Algorithms, Resources, and Applications*. Theory and Applications of Natural Language Processing. Springer, 2016. ISBN 978-3-662-47908-7. doi: 10.1007/978-3-662-47909-4. URL https://doi.org/10.1007/978-3-662-47909-4.

Massimo Poesio, Juntao Yu, Silviu Paun, Abdulrahman Aloraini, Pengcheng Lu, Janosch Haber, and Derya Cokal. Computational models of anaphora. *Annual Review of Linguistics*, 9:561–587, 2023.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W12-4501.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2025.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, Cambridge, MA, October 2010a. Association for Computational Linguistics. URL https://aclanthology.org/D10-1048.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A multi-pass sieve for coreference resolution. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2010b.

Altaf Rahman and Vincent Ng. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1071.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Marta Recasens and Eduard Hovy. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510, 2011.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1008.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2002. URL https://www.aclweb.org/anthology/N18-2002.

Santosh Kumar Sahu, Anil Mokhade, and Neeraj Dhanraj Bokde. An overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance: Recent progress and challenges. *Applied Sciences*, 13(3):1956, 2023.

Micah Shlain, Hillel Taub-Tabib, Shoval Sadde, and Yoav Goldberg. Syntactic search by example. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 17–23, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.3. URL https://aclanthology.org/2020.acl-demos.3.

Amber Shore, Russell Scheinberg, Ameeta Agrawal, and So Young Lee. Correct-detect: Balancing performance and ambiguity through the lens of coreference resolution in llms, 2025. URL https://arxiv.org/abs/2509.14456.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15 (56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Rosemary J. Stevenson, Rosalind A. Crawley, and David Kleinman. Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, 9(4):519–548, 1994.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 656–664, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/P09-1074.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162, 2020. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2020.01.010. URL https://www.sciencedirect.com/science/article/pii/S1566253519303677.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *CoRR*, abs/2107.02137, 2021. URL https://arxiv.org/abs/2107.02137.

Yee Teh, Victor Bapst, Wojciech M Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. Distral: Robust multitask reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4496–4506, 2017.

Paul C Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The journal of finance*, 63(3):1437–1467, 2008.

Raghuveer Thirukovalluru, Nicholas Monath, Kumar Shridhar, Manzil Zaheer, Mrinmaya Sachan, and Andrew McCallum. Scaling within document coreference to long texts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3921–3931, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.343. URL https://aclanthology.org/2021.findings-acl.343.

Yuval Varkel and Amir Globerson. Pre-training mention representations in coreference models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8534–8540, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.687. URL https://www.aclweb.org/anthology/2020.emnlp-main.687.

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

Patrick Verga and Andrew McCallum. Row-less universal schema. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, pages 63–68, San Diego, CA, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1312. URL https://www.aclweb.org/anthology/W16-1312.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, 1995. URL https://www.aclweb.org/anthology/M95-1005.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

Lan Wang, Junjie Peng, Cangzhi Zheng, Tong Zhao, et al. A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning. *Information Processing & Management*, 61 (3):103675, 2024.

Yu Wang, Yilin Shen, and Hongxia Jin. An end-to-end actor-critic-based neural coreference resolution system. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7848–7852, 2021. doi: 10.1109/ICASSP39728.2021.9413579.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6: 605–617, 2018. doi: 10.1162/tacl_a_00240. URL https://www.aclweb.org/anthology/Q18-1042.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6: 287–302, 2018.

Terry Winograd. Understanding natural language. *Cognitive Psychology*, 3(1):1–191, 1972a. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(72)90002-3. URL https://www.sciencedirect.com/science/article/pii/0010028572900023.

Terry Winograd. Understanding natural language. *Cognitive Psychology*, 3(1):1–191, 1972b.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. Learning global features for coreference resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1114. URL https://aclanthology.org/N16-1114.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.622. URL https://www.aclweb.org/anthology/2020.acl-main.622.

Patrick Xia, João Sedoc, and Benjamin Van Durme. Incremental neural coreference resolution in constant memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8617–8624, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.695. URL https://aclanthology.org/2020.emnlp-main.695.

Frank Z Xing, Erik Cambria, and Roy E Welsch. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73, 2018.

Liyan Xu and Jinho D. Choi. Revealing the myth of higher-order inference in coreference resolution. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8527–8533, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.686. URL https://aclanthology.org/2020.emnlp-main.686.

Liyan Xu and Jinho D. Choi. Online coreference resolution for dialogue processing: Improving mention-linking on real-time conversations. In Vivi Nastase, Ellie Pavlick, Mohammad Taher Pilehvar, Jose Camacho-Collados, and Alessandro Raganato, editors, *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 341–347, Seattle, Washington, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.starsem-1.30. URL https://aclanthology.org/2022.starsem-1.30.

Nishant Yadav, Nicholas Monath, Rico Angell, and Andrew McCallum. Event and entity coreference using trees to encode uncertainty in joint decisions. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 100–110, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.crac-1.11. URL https://aclanthology.org/2021.crac-1.11.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.582. URL https://www.aclweb.org/anthology/2020.emnlp-main.582.

Asaf Yehudai, Arie Cattan, Omri Abend, and Gabriel Stanovsky. Evaluating and improving the coreference capabilities of machine translation models. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 980–992, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.69. URL https://aclanthology.org/2023.eacl-main.69.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. Neural mention detection. In *LREC*, 2020a.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. Pairwise representation learning for event coreference, 2020b. URL https://arxiv.org/abs/2010.12808.

Amir Zeldes. The gum corpus: Creating multilayer resources in the classroom. *Lang. Resour. Eval.*, 51(3):581–612, September 2017. ISSN 1574-020X. doi: 10.1007/s10579-016-9343-x. URL https://doi.org/10.1007/s10579-016-9343-x.

Yutao Zeng, Xiaolong Jin, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. Event coreference resolution with their paraphrases and argument-aware embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3084–3094, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.275. URL https://aclanthology.org/2020.coling-main.275.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2114–2124, 2021.

Hongming Zhang, Yan Song, Yangqiu Song, and Dong Yu. Knowledge-aware pronoun coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 867–876, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1083. URL https://www.aclweb.org/anthology/P19-1083.

Rui Zhang, Cícero Nogueira dos Santos, Michihiro Yasunaga, Bing Xiang, and Dragomir Radev. Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 102–107, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2017. URL https://aclanthology.org/P18-2017.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://www.aclweb.org/anthology/N18-2003.

Pengfei Zhu, Zhuosheng Zhang, Jiangtong Li, Yafang Huang, and Hai Zhao. Lingke: a fine-grained multi-turn chatbot for customer service. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 108–112, Santa Fe, New Mexico, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-2024.