UNIVERSITY OF CALIFORNIA SAN DIEGO

Improving Explainability in Personalized Systems by Extraction and Understanding

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science

by

Jiacheng Li

Committee in charge:

Professor Julian McAuley, Chair
Professor Jingbo Shang, Co-Chair
Professor Taylor Berg-Kirkpatrick
Professor Zhiting Hu

2023

The Dissertation of Jiacheng Li is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To the boundless potential harbored within the realm of artificial intelligence.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am immensely grateful for the privilege I had to follow my academic ambitions. This endeavor would have remained a far-fetched dream without the persistent backing and help from numerous individuals.

I would like to extend my profound gratitude to Prof. Julian McAuley and Prof. Jingbo Shang, my advisors, for their ceaseless support, guidance, and countless insightful discussions throughout my doctoral voyage. Their invaluable counsel has not only honed my research acumen but also bolstered my confidence in navigating other professional avenues beyond academic research.

My heartfelt appreciation goes out to all the members of my dissertation committee Prof. Taylor Berg-Kirkpatrick and Prof. Zhiting Hu for offering me indispensable advice and encouragement which significantly shaped my research trajectory.

Subsequently, my research would not have taken shape without the tremendous contribution from the many mentors I had the privilege to work with during my works: Yifan Gao (Amazon), Shuyang Li (Meta), Xian Li and Chenwei Zhang (Amazon), Jin Li and Ming Wang (Amazon), Chun-nan Hsu (UCSD), Tong Zhao (Amazon), and Haibo Ding (Bosch). Their support was instrumental in the development of my research acumen and professional growth.

My Ph.D. journey was enriched by my incredible collaborators: William Hogan, Ming Wang, Zhankui He, An Yan, Ranak Roy Chowdhury, Yannis Katsis, Tyler Baldwin, and all other co-authors and labmates. Their assistance and insights have been exceedingly invaluable all through.

Lastly, the sacrifices, guidance, and blessings of my parents and wife have been my cornerstone. Without them, achieving my dreams would have remained elusive. Their unwavering belief in me has been the driving force behind my perseverance, and for that, I am eternally grateful. This acknowledgment is a small token of my gratitude towards all who have contributed towards making this significant milestone achievable.

I am also grateful to my co-authors who kindly approved the following publications and

material to be included in my dissertation:

Chapter 2 incorporates material from the publication "Weakly Supervised Named Entity Tagging with Learnable Logical Rules" by Jiacheng Li, Haibo Ding, Jingbo Shang, Julian McAuley, Zhe Feng, which was published in *Annual Meeting of the Association for Computational Linguistics*, 2021. The author of this dissertation was the principal investigator and the lead author of this paper.

Chapter 3 incorporates material from the publication "UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining" by Jiacheng Li, Jingbo Shang, Julian McAuley, which was published in *Annual Meeting of the Association for Computational Linguistics*, 2022. The author of this dissertation was the principal investigator and the lead author of this paper.

Chapter 4 incorporates material from the publication "Justifying Recommendations Using Distantly-labeled Reviews and Fine-grained Aspects" by Jianmo Ni, Jiacheng Li, Julian McAuley, which was published in *Conference on Empirical Methods in Natural Language Processing*, 2019. The author of this dissertation was one of the primary authors of this paper.

Chapter 5 incorporates material from the publication "UCEpic: Unifying Aspect Planning and Lexical Constraints for Generating Explanations in Recommendation" by Jiacheng Li, Zhankui He, Jingbo Shang, Julian McAuley, which was published in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2023. The author of this dissertation was the principal investigator and the lead author of this paper.

Chapter 6 incorporates material from the publication "Personalized Showcases: Generating Multi-Modal Explanations for Recommendations" by An Yan*, Zhankui He*, Jiacheng Li*, Tianyang Zhang, Julian McAuley, which was published in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023. The author of this dissertation was one of the primary authors of this paper.

VITA

2018	B.Eng in Information Security, Nanjing University of Posts & Telecommunications
2020	M.S. in Computer Science, University of California San Diego
2023	Ph.D. in Computer Science, University of California San Diego

PUBLICATIONS

Jianmo Ni, **Jiacheng Li**, Julian McAuley. Justifying Recommendations using Distantly-Labeled Reviews and Fine-grained Aspect. *Empirical Methods in Natural Language Processing* (EMNLP), 2019.

Jiacheng Li, Yujie Wang, Julian McAuley. Time Interval Aware Self-Attention for Sequential Recommendation. *International Conference on Web Search and Data Mining* (WSDM), 2020.

Yang Jiao*, **Jiacheng Li***, Jiaman Wu, Dezhi Hong, Rajesh Gupta, Jingbo Shang. SeNsER: Learning Cross-Building Sensor Metadata Tagger. *Empirical Methods in Natural Language Processing* (EMNLP), 2020.

Jiacheng Li, Haibo Ding, Jingbo Shang, Julian McAuley, Zhe Feng. Weakly Supervised Named Entity Tagging with Learnable Logical Rules. *Annual Meeting of the Association for Computational Linguistics* (ACL), 2021.

Jiacheng Li, Jingbo Shang, Julian McAuley. UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining. *Annual Meeting of the Association for Computational Linguistics* (ACL), 2022.

Jiacheng Li, Tong Zhao, Jin Li, Jim Chan, Christos Faloutsos, George Karypis, Soo-Min Pantel, Julian McAuley. Coarse-to-Fine Sparse Sequential Recommendation. *International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR), 2022.

Jiacheng Li, Yannis Katsis, Tyler Baldwin, Ho-Cheol Kim, Andrew Bartko, Julian McAuley, Chun-Nan Hsu. SPOT: Knowledge-Enhanced Language Representations for Information Extraction. *31st ACM International Conference on Information and Knowledge Management* (CIKM), 2022.

William Hogan, **Jiacheng Li**, Jingbo Shang. Fine-grained Contrastive Learning for Relation Extraction. *Empirical Methods in Natural Language Processing* (EMNLP), 2022.

Ranak Roy Chowdhury, **Jiacheng Li**, Xiyuan Zhang, Dezhi Hong, Rajesh Gupta, Jingbo Shang. PrimeNet: Pre-training for Irregular Multivariate Time Series. *AAAI Conference on Artificial*

Intelligence (AAAI), 2023.

An Yan*, Zhankui He*, **Jiacheng Li***, Tianyang Zhang, Julian McAuley. Personalized Showcases: Generating Multi-Modal Explanations for Recommendations. *International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR), 2023.

Jiacheng Li, Zhankui He, Jingbo Shang, Julian McAuley. UCEpic: Unifying Aspect Planning and Lexical Constraints for Generating Explanations in Recommendation. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD), 2023.

Jiacheng Li, Ming Wang, Jin Li, Jinmiao Fu, Xin Shen, Jingbo Shang, Julian McAuley. Text Is All You Need: Learning Language Representations for Sequential Recommendation. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD), 2023.

William Hogan, **Jiacheng Li**, Jingbo Shang. Open-world Semi-supervised Generalized Relation Discovery Aligned in a Real-world Setting. *Empirical Methods in Natural Language Processing* (EMNLP), 2023.

ABSTRACT OF THE DISSERTATION

Improving Explainability in Personalized Systems by Extraction and Understanding

by

Jiacheng Li

Doctor of Philosophy in Computer Science

University of California San Diego, 2023

Professor Julian McAuley, Chair
Professor Jingbo Shang, Co-Chair

The development of personalized systems, driven by sophisticated machine learning models, has notably enriched user experiences across various digital interfaces. However, these systems often obscure the rationale behind personalized recommendations, creating a pressing need for enhanced explainability. We present a comprehensive framework aimed at bridging this explainability gap by systematically extracting, understanding, and demonstrating key information to users.

First, the framework starts with the extraction of information using named entity tagging. This step facilitates the identification and extraction of significant entities and terms from vast

datasets. The precision in extraction is crucial as it directly impacts the quality of understanding and explanation in subsequent phases. Upon successful extraction, the framework transitions to the understanding phase, where the unsupervised contrastive learning model, UCTopic, is employed. This model analyzes the extracted phrases, diving deep into their semantic and thematic contexts. It generates context-aware phrase representations and mines topics, thereby elucidating the thematic essence and semantic correlations encapsulated within the data. Finally, we leverage the strength of various models to generate coherent and intuitive explanations. These generative models can be categorized based on topic, keyphrase, or multi-modality. The generated explanations provide a clear rationale behind the recommendations, making them easily interpretable and relatable to the users.

In summation, our research improves the level of transparency and interpretability inherent in personalized systems. The empirical assessments show the effectiveness of our research in supporting explainability, thereby having a more transparent and user-aligned experience. Through this endeavor, our research substantially improves explainability in personalized systems forward, resulting in a more intuitive and user-friendly interaction paradigm.

Chapter 1

Introduction

The advancement of personalized systems, propelled by the nuanced algorithms of machine learning models, has substantially enriched user experiences across different scenarios. These systems, commonly referred to as recommender systems, harness vast amounts of data to understand user preferences and behaviors, providing suggestions that enhance engagement and satisfaction. They play a vital role in having a more intuitive and personalized interaction between users and items.

However, a notable downside to these sophisticated systems is the black-box rationale behind the personalized recommendations they provide. The complex nature of the algorithms employed can leave users confused about why certain recommendations are made, which could potentially hinder trust and acceptance. This opacity has brought a pressing need for enhanced explainability in recommender systems. Explainability in this context refers to the ability of a system to provide clear, understandable insights into recommended items, allowing users to understand the recommendations made. We attempt to improve the explainability of recommender systems by extracting key information, comprehending content, and generating explanations based on natural languages.

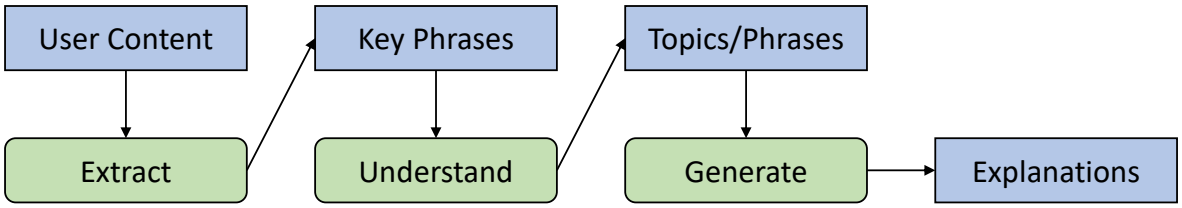


Figure 1.1. The overall framework of improving explanation generation by extraction and understanding.

1.1 Overall Framework

In our research endeavor, we aim to augment the explainability of recommender systems by generating textual explanations. Upon recommending an item to a user, this system is designed to explain the rationale behind the recommendation, leveraging both user content and item features. The user content is usually derived from historical user-generated text or interactions, such as user reviews, while item features might be sourced from item metadata or reviews associated with the item.

Nonetheless, a significant challenge arises from the nature of user-generated text, like user reviews, which tends to be noisy and less informative. The textual data originating from users often contains informal language, irrelevant information, or may lack a clear expression of user preferences, making it a less reliable source for understanding user preferences and item features. Previous methodologies exhibit a limited capability in incorporating user preference and item features from these texts, primarily due to the sparsity of useful information within them.

A promising avenue for enhancement lies in distilling knowledge from user-generated text by extracting key pieces of information. For instance, despite the noisy nature of user-generated text, there might be recurring phrases or keywords that signal user preferences or highlight specific features of items. Extracting such recurring or emphasized information could provide a clear understanding of both the user preference and the item attributes. This distilled information can serve as a robust foundation for generating explanations that are more aligned

with the user preferences and the item features.

Besides, the concept of controllable generation emerges as an important aspect of explanation generation. Controllable generation entails the ability to steer the generation process in a way that the resultant explanations are tailored to certain criteria or adhere to specified themes. This is particularly crucial in a personalized system where the objective is to align explanations with individual user preferences and specific item attributes. By implementing controllable generation mechanisms, the system can ensure that the explanations are not only accurate but also personalized, making the recommendations more transparent and relatable to the users.

To enhance the explainability of the personalized system, we explore augmenting the effectiveness of information comprehension and the controllability of generation through a three-pronged approach: extraction, understanding, and controllable generation. As depicted in Figure 1.1, our initial step involves extracting key phrases from user content, thereby addressing the challenge of information sparsity. Following this extraction, we engage a robust model to understand the semantics encapsulated within these extracted phrases, aiming to derive topics or high-quality representations of phrases. These topics and representations furnish high-quality aspects that serve to guide the generation process, ensuring that the explanations generated are insightful, relevant, and finely tailored to cater to individual user preferences and item attributes.

1.2 Information Extraction

User preferences and item attributes are important in personalized systems. To furnish convincing explanations within such systems, AI systems typically need to first thoroughly grasp the intersections between user personas and item attributes, as reflected in existing user content (e.g., reviews) and item metadata. However, contemporary AI systems often struggle to incorporate user content in the aforementioned scenarios due to its substantial volume and lower information density. Consequently, it is unfeasible to incorporate the entirety of this content within AI systems.

Existing techniques for explanation generation in systems typically address the aforementioned issue through the following three methods:

1. Concatenation and Truncation: This approach straightforwardly concatenates all available content as input until the model input limitation is reached. Subsequent content is typically disregarded, resulting in a loss of potentially valuable information.

2. Selection: By selectively harnessing pertinent information present in the data (e.g., prioritizing the most recent reviews to represent current user preferences), this method retains only a fraction of the content, yet still provides useful information for personalization. However, its efficacy heavily relies on the particular selection strategy employed, which is usually heuristic in nature.

3. Encoding: Rather than directly assimilating text content, certain methods first encode the text to obtain low-dimensional representations. These representations are then utilized to comprehend user preferences and item attributes. This hierarchical approach, however, falls short in capturing fine-grained information essential for generating explanations.

In our research, we discovered that user preferences and item attributes predominantly manifest in keywords rather than extensive sentences. Therefore, employing information extraction techniques (e.g., named entity tagging) can significantly aid in comprehending information from massive content without notable loss of information.

However, existing methods for named entity tagging have the following limitations in personalized systems:

1. They often require a large amount of manual, domain-specific labels, which is challenging in emerging domains due to the required manual effort and deep understanding of the target data.

2. Some rely on manual lexicons or heuristic rules provided by domain experts as weak

supervision, which can be limiting and labor-intensive. Especially for personalized systems, the target entities can be customized.

3. Existing neural network-based models typically lack explainability, which can hinder interaction with humans when selecting customized entities for personalized systems.

Ideally, the information extraction for personalized systems should be automated, scalable, and require minimal manual intervention, while maintaining high accuracy and relevance. Furthermore, it should allow users to effortlessly customize target entities in alignment with the requirements of personalized systems. My research succinctly tackles the above-mentioned challenges through the implementation of learnable logical rules. In this framework, users need only provide a handful of seed rules for named entity tagging and can achieve flexible target entity recognition by adjusting the learned rules.

1.3 Content Understanding

The objective of content understanding, as described in our research framework, is to address the critical task of understanding the semantics embedded within extracted keywords. As our exploration of personalized systems, the key to meaningful interactions between users and items often hinges on a well-rounded comprehension of these semantics. High-quality representations of keywords not only act as a key part of enhanced explanation generation but also serve as a bridge to a more intuitive and enriched user-item engagement within personalized systems. The two primary facets of this endeavor are detailed below:

Semantics Comprehension with Keyword Sparsity:

- The nature of user-generated content in personalized systems, such as reviews and feedback, often results in a sparse set of extracted keywords. This sparsity presents a challenge as each keyword carries a weight of meaning and intent reflective of user preferences and item attributes.

- Understanding the semantics of different keywords is thus important, as it reflects the underlying interactions between users and items. This comprehension helps us have a more nuanced analysis and enables the system to make more informed and relevant explanations for recommendations.

- The semantic understanding also helps bridge the gap between the black-box nature of AI systems and the tangible insights required for enhancing user trust and acceptance.

Normalized Keywords for Controllable Generation:

- Normalized keywords, categorized into coherent topics, serve as structured guidance for controllable explanation generation. This control is important in practical scenarios where the relevance and accuracy of explanations directly impact user satisfaction and system efficacy.

- By employing normalized keywords, the system can generate explanations that are both insightful and tailored to the specific context of user-item interactions. This, in turn, augments the personalization aspect, making the recommendations more relevant to individual user preferences.

- Furthermore, the structured nature of normalized keywords facilitates a systematic approach to explanation generation, making the process more transparent, controllable, and ultimately, more aligned with the goal of enhanced explainability in personalized systems.

Unfortunately, existing methods for keyword or keyphrase understanding often either combine unigram representations in a context-agnostic manner or need supervision from task-specific datasets or distant annotations with knowledge bases. Manual or distant supervision limits the ability to represent out-of-vocabulary phrases, especially for domain-specific datasets (e.g., user-generated content).

In our research, we aim to propose a method capable of learning keyword or keyphrase representations autonomously, without any supervision from human annotations or distant

labels. Given the ability of contrastive learning in representation learning, we explore its applicability to phrase understanding. Our method is proposed for rapid domain adaptation, making it a robust candidate for enhancing personalized systems. This adaptability allows for an intimate understanding of user and item attributes across varied domains, which is important for understanding personalized user-item interactions and generating insightful explanations.

1.4 Explanation Generation

Based on the extracted information and understanding of the content, the objective is to generate explanations for personalized systems using natural languages. Previous methods primarily focus on incorporating personalized information and generating explanations in natural languages. However, for personalized systems, it is crucial to control the generation to cater to customized needs. In our research, the endeavor is to enhance explanation generation from three distinct dimensions:

1. **Topic-based controllable generation**. Tailoring explanations based on specific topics allows for a more directed and relevant generation of explanations. It ensures that the generated content aligns with the particular interests or inquiries of the users, thus improving the user experience and satisfaction.

2. **Keyphrase-based controllable generation**. The significance of keyphrases in explanation generation lies in their ability to encapsulate core information in detail. Previous explanation generation methods suffer from generalized languages for different items. By focusing on keyphrases, the system can generate explanations that are both concise and informative, providing a quick insight into the rationale behind the recommendations.

3. **Image-based controllable generation**. We introduce a novel task named personalized showcases, aiming to provide both textual and visual information to explain recommendations. Our approach first selects a personalized image set most relevant to a user's

interest in a recommended item and then generates natural language explanations accordingly based on the selected images. This method proposes a personalized multi-modal framework capable of generating diverse and visually-aligned explanations through contrastive learning, which has shown to yield more expressive, diverse, and visually-aligned explanations compared to previous methods.

These three dimensions aim to ensure that the explanations generated are not only insightful but also aligned with the personalized needs and preferences of the users, thereby enhancing the overall user experience and trust in the recommender systems. Through a combination of topic-based, keyphrase-based, and image-based controllable generation, the research seeks to bridge the gap between generic explanation models and the personalized explanations required for more intuitive and satisfying user interaction within personalized systems.

1.5 Summary

This thesis is organized into six chapters, each introducing a critical facet of our proposed framework encompassing the processes of extraction, understanding, and generation. A significant portion of this thesis is dedicated to the exploration of explanation generation, with Chapters 4, 5, and 6 exploring the controllable generation predicated on topics, keyphrases, and images respectively.

Chapter 2 involves user-generated content extraction. It unveils a novel methodology, named TALLOR, that combines the bootstrap of high-quality logical rules to tutor a neural tagger autonomously. We further discuss compound rules, a combination of simple rules, aiming to augment the precision of boundary detection whilst fostering a diverse spectrum of pseudo labels. To ensure the superior quality of pseudo labels and thwart the overfitting malaise of the neural tagger, we architect a dynamic label selection strategy.

Chapter 3 delves into the comprehension of the extracted kernels of information. It contains the introduction of UCTOPIC, an innovative unsupervised contrastive learning framework

designed for context-aware phrase representations and topics. UCTopic is rigorously pretrained on a grand scale to understand whether the contexts of two phrase mentions are semantically congruent. The cornerstone of this pretraining is the construction of positive pairs, birthed from our phrase-centric hypotheses. Nonetheless, we found a performance decay when finetuning on datasets with small topic numbers, instigated by traditional in-batch negatives. This propels us to propose a Cluster-Assisted Contrastive Learning (CCL) paradigm, which significantly attenuates the noise in negatives by cherry-picking them from clusters, thereby enriching the phrase representations for topics.

Chapter 4 contains topic-based controllable explanation generation. Within this chapter, we unveil two personalized generation models nurtured with this data: (1) a reference-based Seq2Seq model infused with aspect-planning, capable of crafting justifications encompassing various aspects, and (2) an aspect-conditional masked language model, good at generating diverse justifications based on templates extracted from justification chronicles.

Chapter 5 contains hardly constrained controllable explanation generation. This chapter introduces UCEpic, a sophisticated model that produces high-quality personalized explanations for recommendation outcomes by unifying aspect planning and lexical constraints within an insertion-based generation framework. Compared to recommendation explanation generators solely steered by aspects, UCEpic embraces specific information distilled from keyphrases, thereby significantly enhancing the diversity and informativeness of the explanations generated.

Chapter 6 introduces a method of multi-modal explanation generation. This chapter contains a framework that combines both textual and visual information to explain our recommendations. The text generation is controlled by the accompanying images, allowing the proposed framework to spawn a wide array of diverse and visually-aligned explanations through the lens of contrastive learning.

Chapter 2

Weakly Supervised Named Entity Tagging with Learnable Logical Rules

In this chapter, we extensively explore the domain of information extraction for personalized systems, focusing primarily on weakly supervised named entity tagging. Previous methodologies predominantly concentrate on disambiguating entity types based on contextual indicators and expert-defined rules, while operating under the assumption that entity spans are pre-determined. In contrast, we propose a novel method, denoted as TALLOR, that autonomously bootstraps high-quality logical rules to train a neural tagger. Specifically, we introduce compound rules, which are formulated by combining simple rules, to enhance the precision of boundary detection and to generate a more diverse set of pseudo labels. Additionally, we architect a dynamic label selection strategy to maintain the quality of pseudo labels, thereby mitigating the risk of overfitting the neural tagger.

Empirical evaluations conducted on three distinct datasets illustrate that our method surpasses other weakly supervised approaches and even competes with a state-of-the-art distantly supervised tagger equipped with a lexicon of over 2,000 terms, all while initiating from a mere 20 simple rules. This demonstrates the efficacy of TALLOR as a tool for swiftly developing taggers in emerging domains and tasks.

Furthermore, our method provides the flexibility for customized named entity extraction by allowing the input of tailored seed rules. The rules derived from this method have the potential

to elucidate the predicted entities, which is a crucial feature for personalized systems. This chapter, therefore, not only introduces an innovative methodology for information extraction but also lays the foundation for deeper comprehension and customization in named entity tagging, contributing significantly to the advancement of personalized systems.

2.1 Introduction

In this chapter, we explore the domain of information extraction for personalized systems, with a particular focus on weakly supervised named entity tagging. Although supervised training methodologies for entity tagging systems often yield accurate results, they need a significant volume of manual, domain-specific labels. This requirement poses a challenge when applying such methodologies to emerging domains and tasks. To mitigate the manual effort, previous works have leveraged manual lexicons [Shang et al., 2018b, Peng et al., 2019] or heuristic rules provided by domain experts [Fries et al., 2017, Safranchik et al., 2020] as a form of weak supervision. For instance, LinkedHMM [Safranchik et al., 2020] has demonstrated performance comparable to supervised models using 186 heuristic rules, supplemented by a lexicon encompassing over two million terms. However, crafting complete and accurate rules or lexicons in emerging domains is a demanding task that requires a substantial manual effort and a profound understanding of the target data, leaving the question of building accurate entity tagging systems with reduced manual effort as an open problem.

We explore methodologies capable of autonomously deriving new rules from unlabeled data, utilizing a minimal set of seed rules (e.g., 20 rules). Such methodologies are highly desirable in real-world applications due to their rapid deployment ability to new domains or customized entity types, coupled with the effectiveness, interpretability, and simplicity of the learned rules for non-experts to rectify incorrect predictions. As depicted in Figure 2.1, new rules can be derived from seed rules. Specifically, we propose a novel iterative learning method, TALLOR, which is designed to accurately derive rules for training a neural tagger in an automated manner.

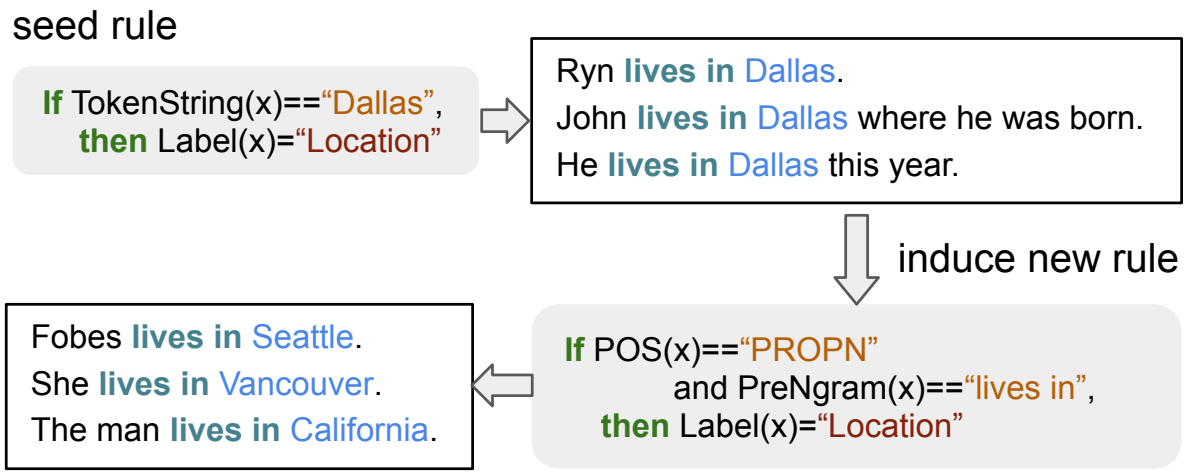


Figure 2.1. Examples of a seed logical rule and a newly induced rule from labeled data for recognizing locations. 'x' denotes a token span from a given sentence.

This method addresses two critical issues during the learning process: (1) the simultaneous detection of entity boundaries and prediction of their types using rules, and (2) the generation of accurate and diverse pseudo labels from rules.

While previous works [Niu et al., 2003, Huang and Riloff, 2010, Gupta and Manning, 2014] with a small set of seed rules as supervision predominantly focus on disambiguating entity types, assuming entity spans are pre-specified or merely represent syntactic chunks (e.g., noun phrases), we observe that syntactic chunks often misalign with target entity spans. For instance, in a given sentence from CoNLL2003: "Germany's representative to the European Union's veterinary committee...", the noun phrases [1] are "Germany's representative" and "the European Union's veterinary committee", whereas the gold entities are "Germany" and "European Union". This misalignment is evidenced further when comparing noun phrases extracted from spaCy with ground truth entity boundaries, as shown in Table 2.1, illustrating that a majority of target entities are missed when utilizing noun phrases as entity candidates.

To concurrently address entity boundary detection and type classification, we initially define five types of simple logical rules considering the lexical, local context, and syntax

[1] Noun phrases are extracted using spaCy noun chunks.

Table 2.1. Boundary detection performance from our method and parsing based noun phrases.

	Noun phrase			TALLOR		
	P	R	F_1	P	R	F_1
BC5CDR	17.1	50.1	25.5	69.8	67.8	68.7
CHEM	3.2	35.6	5.8	63.0	60.2	61.6
CoNLL	4.1	47.3	7.5	86.9	86.7	86.8

information of entities. Recognizing that simple logical rules often fall short in accurately detecting entity boundaries, we propose the derivation of compound logical rules, which are formulated from multiple simple rules and logical connectives (e.g., "and"). For instance, in the sentence "John lives in Dallas where he was born", the simple rule "lives in _", which is a preceding context clue, will match multiple token spans such as "Dallas", "Dallas where", "Dallas where he", etc. In contrast, compound logical rules can both accurately detect entity boundaries and classify their types. For example, employing both the preceding context and the part-of-speech (POS) tag rule (e.g., "lives in _" and POS is a proper noun) can correctly identify the Location entity "Dallas".

Although the objective is to derive accurate rules, automatically acquired rules can be inherently noisy. To ensure the quality of generated pseudo labels, we design a dynamic label selection strategy to select highly accurate labels, enabling the neural tagger to learn new entities rather than overfitting to the seed rules. Specifically, we maintain a high-precision label set during the learning process. For each learning iteration, we first automatically estimate a filtering threshold based on the high-precision set, subsequently filtering out low-confidence pseudo-labels by considering both their maximum and average distances to the high-precision set. Highly confident labels are incorporated into the high-precision set for the subsequent learning iteration. This dynamic selection strategy facilitates our framework in maintaining the precision of recognized entities while augmenting recall during the learning process, as demonstrated in our experiments.

We conduct evaluations of our method on three datasets, with experimental results

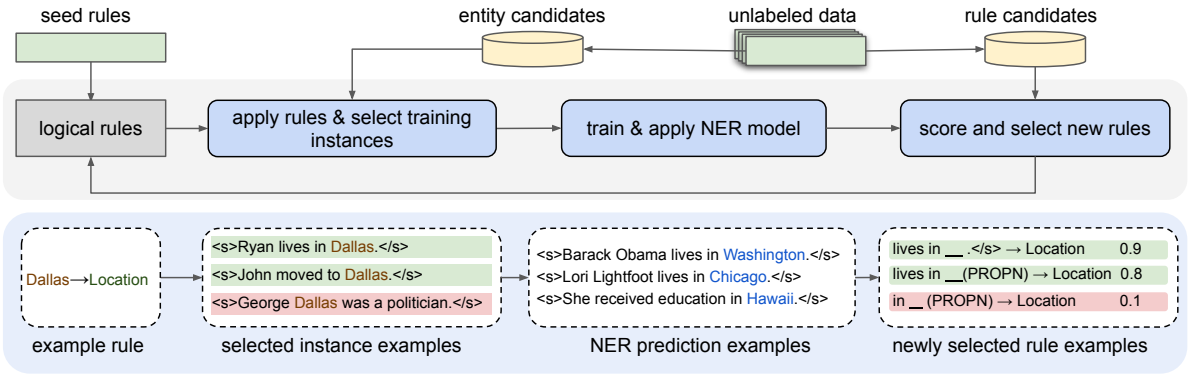


Figure 2.2. Overview of our proposed weakly-supervised NER framework by logical rules.

indicating that TALLOR outperforms existing weakly supervised methods, enhancing the average F_1 score by 60% across three datasets over methods using seed rules. Further analysis reveals that TALLOR can achieve performance akin to a state-of-the-art distantly supervised method with only 1% of the human effort [2]. Additionally, we conduct a user study concerning the explainability of learned logical rules, finding that annotators concur that an average of 79% (across three annotators) of the matched logical rules can be utilized to explain why a span is predicted as a target entity.

2.2 Method Overview

Figure 2.2 illustrates the flow within our iterative learning framework, which encompasses the subsequent stages. Initially, we derive all potential entity candidates and rule candidates from the unlabeled dataset. In each iterative cycle, logical rules are applied to the unlabeled data, subsequently refining a subset of high-quality, weak training instances. We then employ these instances to train a neural tagging model, which subsequently predicts labels for the unlabeled dataset. Subsequent to this, from the candidate rules, we identify and incorporate new precise logical rules based on these predictions. These newly integrated rules are then utilized to determine weak training labels for the succeeding iteration.

[2] In experiments, our method utilized 20 rules, while the other system employed a manually constructed lexicon of over 2000 terms.

2.3 Neural Tagging Model

In this section, we introduce the neural named entity recognition model (NER model) shown in Figure 2.2. Following the approach of Jiang et al. [2020], we approach tagging through span labeling. The principal methodology involves encoding each span within a consistent-length embedding and subsequently making predictions using this embedding. Specifically, for a given span and its related sentence, we commence by initializing all tokens within the sentence utilizing a pre-trained language model. Subsequently, a Bi-LSTM and Self-Attention layer are employed to derive the contextual embedding of the sentence. The span's embedding is then formulated by merging two components: a *content representation*, determined as the weighted mean of the span's token embeddings, and a *boundary representation*, which amalgamates the embeddings at the span's commencement and termination. The span's label prediction is then carried out using a multilayer perceptron (MLP).

2.4 Logical Rule Extraction

In our study, a logical rule is formulated as "if p then q" (represented as "$p \rightarrow q$"). [3] For entity tagging, q denotes one of the desired entity categories, while p encompasses any applicable matching criterion. For instance, the rule may be articulated as: "If a span is preceded by the tokens '*lives in*', then it is categorized as a Location." We have formulated five distinct types of elementary logical rules that accommodate the lexical, contextual, and syntactic attributes of a potential entity.

Simple Logical Rules A simple logical rule is characterized as a rule encompassing a single conditional predicate. We have formulated five distinct predicates to encapsulate prevalent logical conditions. Given a candidate entity,

1. `TokenString` matches its lexical string;

[3] The terms "heuristic rules" and "labeling rules" can be equated to logical rules, permitting their interchangeable use.

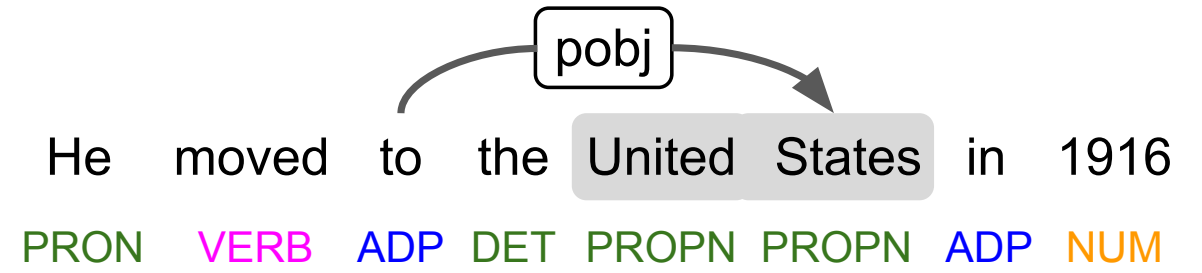


Figure 2.3. An example illustrated for rules.

2. `PreNgram` matches its preceding context tokens;

3. `PostNgram` matches its succeeding context tokens;

4. `POSTag` matches its part-of-speech tags;

5. `DependencyRel` matches the dependency relations of its headword.

Given a candidate entity "*United States*" in Figure 2.3, we can extract the following example logical rules for recognizing Locations:

`TokenString`=="*united state*"	→ Location,
`PreNgram`=="*move to the*"	→ Location,
`PostNgram`=="*in 1916*"	→ Location,
`POSTag`=="PROPN PROPN"	→ Location,
`DependencyRel`=="*to*" (via pobj)	→ Location.

Compound Logical Rules A compound logical rule integrates multiple conditional predicates and is connected by logical connectors such as `and` ($\wedge$), `or` ($\vee$), and `negation` ($\neg$). In our work, we concentrate on deriving compound logical rules interlinked with conjunctions ($\wedge$) to recognize entities with precision, given that single logical rules frequently fall short in detecting entity boundaries. For instance, the rule `PreNgram`=="*move to the*" could match multiple candidates like "*United*", "*United States*", and "*United States in*", among which many are imprecise. However, utilizing a compound rule, such as `PreNgram`=="*move to the*" $\wedge$ `POSTag`=="PROPN PROPN", enables the accurate identification of "*United States*" as a Location.

Prior to the training phase, we systematically enumerate and extract all feasible logical rules from unlabeled data, based on our pre-defined rule categories.

2.5 Applying Logical Rules

During each iteration, both initial (seed) and subsequently learned logical rules are applied to unlabeled entity candidates to derive a collection of weakly labeled instances. When an entity candidate aligns with multiple rules, which may be contradictory, the predominant rule determined by a majority vote is chosen as the definitive weak label.

Entity Candidate Identification In this research, tagging is approached as a span labeling task, as previously introduced. Prior to the learning phase, every token span up to a predetermined maximum length is enumerated from the unlabeled dataset to establish entity candidates.

Furthermore, it is observed that prevalent phrases, such as "*United States*", are infrequently divided into distinct entities, for instance, "*United*" and "*States*". Consequently, using the unsupervised AutoPhrase methodology Shang et al. [2018a], a compilation of prevalent phrases is generated. Subsequent to this, two consecutive spans that can constitute a prevalent phrase are combined to be recognized as a singular entity candidate.

2.6 Dynamic Training Label Selection

Once the learned rules are applied to the unlabeled data, there is a possibility of generating weak labels with inaccuracies. Such discrepancies can compromise the subsequent performance of our neural tagger. To address this, we propose the establishment of a high-precision entity set, which serves to retain accurately labeled training examples from each iteration. Drawing inspiration from the work of Zhang et al. [2020d], we devise a methodology to curate high-quality labels from the weakly generated ones through the integration of seed logical rules into the high-precision set. More precisely, for a given entity category i, its associated high-precision set

H_i, and a weakly labeled instance e_q, we initiate by calculating a confidence score for e_q being classified under category i. This is achieved by weighing both its maximum pair similarity to the high-precision set H_i (termed as **local score**) and its mean similarity to H_i (termed as **global score**). Subsequently, the weakly labeled instance e_q is incorporated into the high-precision set if its confidence score surpasses a predetermined threshold, which is also derived based on the high-precision set.

Instance Embedding The embedding for an entity instance is determined by averaging the embeddings of its constituent tokens. The embedding for each token is derived by taking the mean output from the initial three layers of a pre-established language model.

Local Score For a weakly labeled instance e_q and a reference instance e_i from the high-precision set, we initially calculate their similarity using the cosine similarity of their respective embeddings. Subsequently, the local confidence score of e_q being associated with category i is determined as the maximum similarity value when compared with all instances within the high-precision set.

Global Score The local score, derived from a singular instance within the high-precision set, offers the advantage of uncovering new entities. However, its reliability can occasionally be compromised. To address this, we introduce a more robust metric termed the global score to evaluate the likelihood of an instance e_q being associated with category i. To compute this, we initially select a subset E_s from the high-precision set H_i. The representative embedding, $\mathbf{x}_{E_s}$, of E_s is then determined by averaging the embeddings of all instances within E_s. This procedure is repeated N times to finalize the global score:

$$\text{score}_i^{glb} = \frac{1}{N} \sum_{1 \le j \le N} \cos(\mathbf{x}_{E_s}^j, \mathbf{x}_{e_q}) \tag{2.1}$$

To have a balance between exploratory capability and reliability, the ultimate confidence score for a weakly labeled instance's association with a category is determined by the geometric mean of its local and global scores.

Dynamic Threshold Estimation We assume that the thresholds for identifying high-quality weak labels may vary across distinct entity categories. Additionally, varying thresholds might be necessary across different iterations to effectively modulate between exploration and dependability. For instance, in the initial phases, the learning process might emphasize dependability, while later stages could prioritize exploration. Driven by this assumption, we advocate for a dynamic threshold in selecting high-quality weak labels. In this approach, an entity instance from the high-precision set is reserved, and its confidence score is evaluated against the remaining examples within that set. This process is iteratively executed T times, setting the threshold at the lowest observed value. For category i, it is calculated as:

$$\text{threshold} = \tau \cdot \min_{k \leq T, e_k \in H_i} \text{score}_i(e_k) \tag{2.2}$$

where e_k is the held-out entity instance and $\tau \in [0,1]$ is a temperature to control the final threshold.

2.7 Logical Rule Scoring and Selection

In each iteration, our neural tagging model is initially employed to predict labels for all textual spans. Subsequently, based on their prediction probabilities, we rank these spans and select the top 70%—noting that varying categories and datasets might necessitate distinct thresholds for label selection. Using a percentage-based approach allows for dynamic threshold adjustments across different categories, enhancing model resilience across diverse domains and categories. Following this, rule candidates are evaluated, and new rules are chosen based on their respective confidence scores. We utilize the *RlogF* methodology, as delineated by Thelen and Riloff [2002], to compute the confidence score for a rule, r:

$$F(r) = \frac{F_i}{N_i} \log_2(F_i) \tag{2.3}$$

where F_i denotes the count of spans that are both predicted with the category label i and matched by the rule r, while N_i represents the total count of spans that the rule r matches. This method holistically evaluates both the precision and reach of rules, where $\frac{F_i}{N_i}$ indicates the rule's precision, and $\log_2(F_i)$ represents its coverage capability.

2.8 Experiments

We conduct our experiments based on the following desiderata: (1) How does our proposed method perform compared to other weakly supervised methods? (2) What is the relation between our method's performance and different experimental settings? (3) How many distant supervisions can achieve similar performance as our seed rules?

2.8.1 Datasets

We conduct assessments of our approach using three distinct datasets. Note that the training set from each dataset is utilized as unlabeled data for our evaluations.

1. **BC5CDR** Li et al. [2016a]: Originating from the BioCreative V CDR task corpus, this dataset comprises 500 training, 500 development, and 500 testing articles from PubMed. It encompasses 15,953 chemical entities and 13,318 disease entities.

2. **CHEMDNER** Krallinger et al. [2015]: This dataset includes 10,000 PubMed abstracts, identifying a total of 84,355 chemical entities. The dataset is divided into training, development, and testing sets with 14,522, 14,572, and 12,434 sentences, respectively.

3. **CoNLL2003** Sang and Meulder [2003]: Derived from Reuters news articles, this dataset contains 14,041 training, 3,250 development, and 3,453 testing sentences. In our evaluation, we focused on the Person, Location, and Organization entities. We excluded the "Misc" category, as it doesn't correspond to a unified semantic category and thus can't be effectively represented by a limited set of seed rules.

2.8.2 Baselines

We compare our method to the following weakly supervised named entity recognition methods to show our effectiveness.

1. **Seed Rules**. We directly employed seed rules to the test set and evaluate their efficacy.

2. **CGExpan** [Zhang et al., 2020d] is a state-of-the-art lexicon expansion method by probing a language model. Given that `TokenString` seed rules can act as a foundational lexicon, we augmented its volume to 1,000 using this method, employing them as `TokenString` rules. We applied the top sets of 200, 500, 800, and 1,000 rules to test sets and documented optimal results.

3. **AutoNER** [Shang et al., 2018b] takes lexicons of typed terms and untyped mined phrases as input. The most effective expanded lexicon from CGExpan serves as typed terms, while the combination of this expanded lexicon and phrases sourced from AutoPhrase [Shang et al., 2018a] are used as untyped mined phrases.

4. **LinkedHMM** [Safranchik et al., 2020]: This innovative generative model integrates noisy rules for supervision and predicts entities utilizing a neural NER model. For our tests, we employed the CGExpan-expanded lexicon as tagging rules and phrases mined by AutoPhrase as linking rules.

5. **HMM-agg.** [Lison et al., 2020]: This model introduces a hidden Markov mechanism that first produces weak labels from labeling functions, followed by training a sequence tagging model. We transformed the CGExpan-expanded lexicon into labeling functions and then documented the tagging model's outcomes.

6. **Seed Rule + Neural Tagger**. This approach encompasses our framework minus iterative learning. After deploying seed rules, we utilized the weakly generated labels to train our neural tagger and then documented the tagger's outcomes.

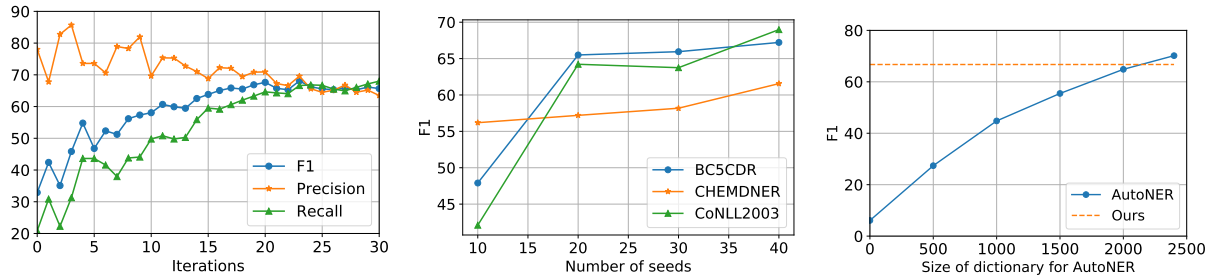
Table 2.2. Performance of baselines (in upper section), our method and its ablations (in lower section).

Methods	BC5CDR			CHEMDNER			CONLL2003		
	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
Seed Rules	**94.09**	3.81	7.33	**91.60**	13.19	23.07	**95.77**	2.76	5.36
LinkedHMM	10.18	15.60	12.32	23.99	10.77	14.86	19.78	31.51	24.30
HMM-agg.	43.70	21.60	29.00	49.60	18.40	26.80	52.00	8.50	14.60
CGExpan	40.96	24.75	30.86	45.70	25.58	32.80	55.97	28.7	37.95
AutoNER	42.22	30.66	35.52	66.83	27.59	39.05	32.07	5.98	10.08
Seed Rules + Neural Tagger	78.33	21.60	33.86	84.18	21.91	34.78	72.57	24.68	36.83
Self-training	73.69	29.55	42.19	85.06	20.03	32.42	72.80	24.83	37.03
Our Learned Rules	79.29	18.46	29.94	69.86	21.97	33.43	65.51	21.12	31.94
Ours w/o Autophrase	74.56	32.93	45.68	67.74	55.99	61.31	71.37	25.50	37.57
Ours w/o Instance Selection	58.70	63.37	60.95	42.64	48.25	45.27	58.51	58.8	58.65
TALLOR	66.53	**66.94**	**66.73**	63.01	**60.18**	**61.56**	64.29	**64.14**	**64.22**

7. **Self-training**. Initiation is with weak labels sourced from seed rule application. Subsequently, a self-training system is established using these weak labels for initial guidance, and our neural tagger as the foundational model.

2.8.3 Performance Comparison

We report the precision, recall, and micro-averaged F_1 scores across three datasets in Table 2.2. Our method demonstrates notable superiority over baseline techniques, registering an average F_1 improvement of 24 points across the three datasets in comparison to the most effective baseline. Our seed rules exhibit high precision, though the recall is limited. The lexicon expansion approach, CGExpan, identifies a broader range of entities. However, it compromises precision, enhancing recall at the expense of accuracy. Current weakly supervised techniques, namely AutoNER, LinkedHMM, and HMM-agg., struggle to effectively identify entities with either seed or CGExpan-expanded rules. These strategies predominantly rely on a high-precision lexicon. Yet, the precision of the lexicon expanded automatically doesn't satisfy this prerequisite. While seed rules are precise, they do not provide comprehensive coverage of diverse entities. Our non-iterative approach denoted as "Seed Rules + Neural Tagger", and the self-training strategy



(a) **(b)** **(c)**

Figure 2.4. (a) Iterations vs. performance of our method on BC5CDR. (b) Performance with different numbers of seed rules. (c) Performance of AutoNER with different sizes of manual lexicon and our method on BC5CDR.

both exhibit high precision. This is attributed to the credible pseudo-labels derived from seed rules. Intriguingly, the self-training model rooted in our neural tagger also manifests limited recall. We assume this is predominantly because the neural tagger overcompensates for the modest quantity of labels sourced from seed rules.

In an effort to understand the significance of various components within our framework, we conducted an ablation study. The outcomes are presented in Table 2.2, specifically in its lower section. The results highlight that while our learned rules are precise, they do not provide extensive coverage. When we exclude common phrases identified by Autophrase, denoted as "Ours w/o Autophrase", our method exhibits a marked decline in recall. This shows the important role common phrases play in enhancing coverage. Moreover, in the absence of the high-quality training instance selection step, referred to as "Ours w/o Instance Selection", there is a noticeable dip in precision. This emphasizes the criticality of the instance selection phase.

2.8.4 Performance vs. Different Settings

Performance vs. Iterations Figure 2.4a illustrates the progression of our method across various iterations. Notably, there is an enhancement in recall from 20% to beyond 60% throughout the learning trajectory, accompanied by a marginal reduction in precision. The optimal F_1 score is realized after 25 iterations.

Table 2.3. Number and ratio of different type rules.

Rule Type	BC5CDR	CHEMDNER	CoNLL
TokenStr	503 (41%)	1667 (44%)	779 (25%)
Pre $\wedge$ Post	203 (17%)	629 (17%)	956 (31%)
Pre $\wedge$ POS	288 (24%)	585 (16%)	455 (15%)
POS $\wedge$ Post	149 (12%)	418 (11%)	438 (14%)
Dep $\wedge$ POS	79 (6%)	432 (12%)	469 (15%)

Performance vs. Different Numbers of Seeds Figure 2.4b presents the outcomes of our approach when utilizing varying quantities of manually selected seed rules across three datasets. Our method exhibits consistent enhancement with an increasing number of seeds. With a mere 10 seeds, our approach achieves an F_1 score exceeding 55% on CHEMDNER, showing the framework's efficiency in a limited supervision context. On datasets like BC5CDR and CoNLL, leveraging 20 seeds yields superior results, around 65% F_1, suggesting that 20 seeds serve as an optimal baseline for developing a tagging system with limited manual intervention.

2.8.5 Comparison with Distant Supervision

AutoNER [Shang et al., 2018b] operates under distant supervision, utilizing a manually curated lexicon. We aimed to determine the number of terms required in AutoNER's lexicon for it to match the performance of our approach. Our experiments on BC5CDR employed just 20 seeds for our method. Conversely, for AutoNER, we incorporated an additional M terms from an expertly crafted lexicon [Shang et al., 2018b], sourced from the MeSH database and CTD Chemical and Disease vocabularies.

Figure 2.4c illustrates the performance metrics corresponding to varying M values. The findings reveal that AutoNER necessitates an addition of approximately 2000 terms to parallel our method's performance, achieving roughly 66% F_1 score. This shows the efficiency of our approach even in the absence of an expansive manually curated lexicon.

2.8.6 Error Analysis and Case Study

Error Analysis Table 2.3 presents the distribution of various rule types learned after completing all iterations, with abbreviations such as TokenStr, Pre, Post, POSTag, and Dep representing TokenString, PreNgram, PostNgram, POSTag, and DependencyRel, respectively. The analysis reveals that the TokenString rule dominates in domain-specific datasets like BC5CDR and CHEMDNER. However, in a more generalized domain, our model predominantly learns the PreNgram∧PostNgram rule.

In our subsequent error analysis of the BC5CDR dataset, we meticulously reviewed 100 entities incorrectly identified by our inferred rules, categorizing the nature of the discrepancies. Our findings indicate that a significant 56% of the errors arise due to challenges in differentiating between closely associated entity categories, such as chemicals versus medications. Additionally, 20% of the errors stemmed from improper identification of entity boundaries. A notable observation was the existence of spans like "HIT type II" and their corresponding sub-spans "HIT", both classified as disease entities. However, only the more extensive spans were annotated with gold labels. Such instances, where our rules predominantly recognize the sub-spans as diseases, account for 20% of the total errors. A detailed representation of each error type is provided in Table 2.4.

Case Study Given the intuitive nature of our logical rules as cues for entity recognition, we posited that these automatically derived rules might serve as comprehensible justifications for entity predictions. To evaluate this hypothesis, we undertook a user study centered on the explainability of these rules. In this study, we utilized the rules learned from the BC5CDR dataset, selecting 100 entities that were labeled by at least one logical rule, excluding TokenString rules, as their self-explanatory nature rendered them redundant for this evaluation. A snapshot of such examples can be found in Table 2.5.

For this evaluative process, we enlisted the expertise of two annotators lacking domain-specific knowledge and one specialist in biology. Their task was to determine if the auto-

generated logical rules were decipherable and could be employed to elucidate the rationale behind designating a text span as either a disease or a chemical. The findings from this manual assessment revealed a convergence in understanding: the two general annotators and the biology expert respectively discerned that 81%, 87%, and 70% of the entity predictions could be logically justified by the provided rules.

2.9 Conclusion

In this chapter, we investigate the construction of a tagger utilizing a limited set of foundational logical rules and unlabeled data. We introduce five categories of basic logical rules and formulate compound logical rules. These compound rules, derived from the basic ones, serve the dual purpose of identifying entity boundaries and concurrently classifying their respective types. Furthermore, we devised a dynamic label selection strategy to curate precise pseudo-labels, produced from the learned rules, to train an advanced tagging model. Empirical findings show the effectiveness of our approach, highlighting its superiority over previous weakly supervised methodologies.

Chapter 2, in part, is a reprint of the material as it appears in "Weakly Supervised Named Entity Tagging with Learnable Logical Rules." by Jiacheng Li, Haibo Ding, Jingbo Shang, Julian McAuley, Zhe Feng, which was published in *Annual Meeting of the Association for Computational Linguistics*, 2021. The dissertation author was the primary investigator and author of this paper.

Table 2.4. Gold entities are underlined, predicted entities are in red. **Error type "similar semantic concepts"** means that our rules cannot distinguish two closely related semantic concepts. **Error type "inaccurate boundary"** means our rules label incorrectly about the boundaries of entities. **Error type "nested entity"** means the error is due to multiple possible entities are nested. **NotEntity** means the predicted span is not an entity.

Examples	Predicted Labels	Gold Label
Error Type: Similar Semantic Concepts (56%)		
The aim of this work is to call attention to the risk of tacrolimus use in patients with SSc.	Disease	NotEntity
We recorded time to first dysrhythmia occurrence , respective times to 25 % and 50 % reduction of the heart rate (HR) and mean arterial pressure , and time to asystole and total amount of bupivacaine consumption.	Disease	NotEntity
The severity of pain due to etomidate injection , mean arterial pressure , heart rate , and adverse effects were also evaluated.	Disease	NotEntity
Error Type: Inaccurate Boundary (20%)		
Furthermore ameliorating effect of crocin on diazinon induced disturbed cholesterol homeostasis was studied.	Disease	Disease
Pretreatment with S. virgaurea extract for 5 weeks at a dose of 250 mg / kg followed by isoproterenol injection significantly prevented the observed alterations.	Chemical	Chemical
This depressive -like profile induced by METH was accompanied by a marked depletion of frontostriatal dopaminergic and serotonergic neurotransmission , indicated by a reduction in the levels of dopamine , DOPAC and HVA , tyrosine hydroxylase and serotonin , observed at both 3 and 49 days post - administration.	Chemical	Chemical
Error Type: Nested Entity (20%)		
Early postoperative delirium incidence risk factors were then assessed through three different multiple regression models.	Disease	Disease
The impact of immune - mediated heparin -induced thrombocytopenia type II (HIT type II) as a cause of thrombocytopenia.	Disease	Disease
Extensive literature search revealed multiple cases of coronary artery vasospasm secondary to zolmitriptan , but none of the cases were associated with TS.	Disease	Disease
Error Type: Others (4%)		
It is characterized by its intense urotoxic action , leading to hemorrhagic cystitis.	Disease	NotEntity
Famotidine is a histamine H2-receptor antagonist used in inpatient settings for prevention of stress ulcers and is showing increasing popularity because of its low cost .	Chemical	NotEntity
It is characterized by its intense urotoxic action , leading to hemorrhagic cystitis.	Disease	NotEntity

Table 2.5. Examples of learned rules and correctly labeled entities (in red) by the learned rules in BC5CDR dataset.

Labeled Entities and Sentences	Learned Logical Rules	Entity type
This occlusion occurred after EACA therapy in a patient with SAH and histopathological documentation of recurrent SAH.	`PreNgram`="a patient with" ∧ `PostNgram`="and"	Disease
We also analyzed published and unpublished follow-up data to determine the risk of PROPN ICH in antithrombotic users with MB.	`PreNgram`="the risk of" ∧ `POStag`=PROPN	Disease
3 weeks after initiation of amiodarone therapy for ADJ atrial NOUN fibrillation.	`PreNgram`="therapy for" ∧ `POStag`=ADJ NOUN	Disease
Although 25 mg of NOUN lamivudine was slightly less effective than 100mg (P=.011) and 300 mg (P=.005).	`PreNgram`="mg of" ∧ `POStag`=NOUN	Chemical
These results suggest that the renal protective effects of NOUN misoprostol is dose - dependent.	`PreNgram`="protective effect of" ∧ `POStag`=NOUN	Chemical

Chapter 3

UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining

In this chapter, we dive deep into phrase understanding. Here, we will focus on unsupervised phrase representation learning and topic mining. High-quality phrase representations play an important role in topic mining within documents. Existing methodologies for obtaining phrase representations either combine unigram representations without considering the surrounding context or need extensive annotations to incorporate such context-aware insights. In this chapter, we introduce a groundbreaking unsupervised contrastive learning framework, referred to as UCTopic, specifically for generating context-aware phrase representations and enhancing topic mining. The pretraining phase of UCTopic emphasizes understanding whether the contexts of two phrase mentions share semantic congruence. Central to this pretraining is the ingenious method of positive pair construction, which derives from our phrase-centric hypotheses. We also introduce the Cluster-Assisted Contrastive Learning (CCL) method. CCL alleviates the impact of noisy negatives by strategically choosing negatives from defined clusters, thereby refining the phrase representations relative to their topics. Empirical analyses underscore the efficacy of UCTopic. Notably, it surpasses its contemporary phrase representation counterparts, boasting a remarkable increase of 38.2% NMI on average across four distinct entity clustering assignments. Moreover, an extensive assessment of topic mining demonstrates that UCTopic

excels in isolating both coherent and multifaceted topical phrases.

3.1 Introduction

Topic modeling identifies abstract 'topics' within a collection of documents, each typically modeled as a distribution over terms. High-quality phrase representations are crucial for topic models to understand semantics and distinguish topics effectively. However, existing methods, such as those by Wang et al. [2021], Yu and Dredze [2015], Zhou et al. [2017], often fall short in integrating representations for phrases (multigram terms) and words (unigram terms) seamlessly. Some methodologies combine unigram embeddings to get context-free representations, leading to the extraction of semantically akin phrases, exemplified by "great food" and "good food". On the other hand, context-aware methods like DensePhrase [Lee et al., 2021] and LUKE [Yamada et al., 2020] need supervision from task-specific datasets or distant annotations, thereby limiting their efficacy in representing novel phrases, particularly in domain-specific contexts. The effectiveness of contrastive learning for unsupervised representation has recently been illuminated in both visual [Chen et al., 2020b] and textual [Gao et al., 2021a] domains.

We present UCTOPIC, an **U**nsupervised **C**ontrastive learning framework dedicated to phrase representations and **TOPIC** mining. Our target is to leverage contrastive learning to deepen the comprehension of phrase semantics within sentences. A key challenge was fabricating contrastive pairs apt for phrase representation learning. Traditional data augmentation techniques in NLP, such as back translation [Xie et al., 2020a], synonym replacement [Zhang et al., 2015], and text mix up [Zhang et al., 2018], are not tailored for this requirement. To address this, we propose two guiding assumptions about phrase semantics:

1. The phrase semantics are determined by their context.

2. Phrases that have the same mentions have the same semantics.

Refer to Figure 3.1 for an illustrative example. By adopting these assumptions, the masked sentences serve as positive pairs in our contrastive learning regime. The intuition

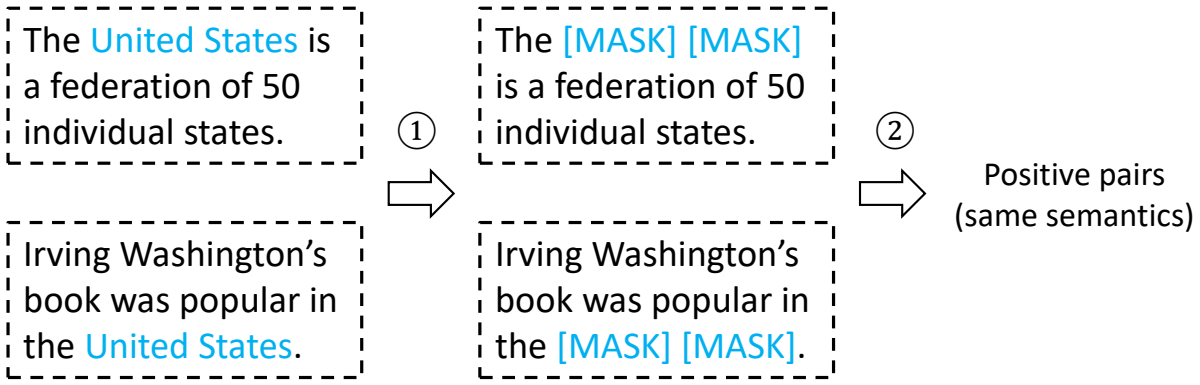


Figure 3.1. Two assumptions used in UCTOPIC to produce positive pairs for contrastive learning.

behind the two assumptions is that we expect the phrase representations from different sentences describing the same phrase should group together in the latent space. Masking the phrase mentions forces the model to learn representations from context which prevents overfitting and representation collapse [Gao et al., 2021a]. Based on the two assumptions, our context-aware phrase representations can be pre-trained on a large corpus via a contrastive objective without supervision.

During large-scale pre-training, we adhered to past works [Chen et al., 2017, Henderson et al., 2017, Gao et al., 2021a] and utilized in-batch negatives. However, the inadequacy of in-batch negatives became a problem during fine-tuning. To address this, we innovated the cluster-assisted contrastive learning (CCL) strategy, leveraging clustering results as pseudo-labels and drawing negatives from instances displaying high confidence within clusters. The cluster-assisted negative sampling offers two distinct benefits:

1. reducing potential positives from negative sampling compared to in-batch negatives;

2. the clusters are viewed as topics in documents, thus, cluster-assisted contrastive learning is a topic-specific finetuning process which pushes away instances from different topics in

the latent space.

Leveraging the dual foundational premises and the innovative cluster-assisted negative sampling described in this paper, we pre-train representations on an expansive dataset. This is followed by fine-tuning these representations on a domain-specific dataset for topic mining, using an unsupervised approach. To show the effectiveness of our phrase representations, we evaluate entity clustering across four datasets. The outcomes reveal that our pre-trained model, denoted as UCTOPIC, realizes a remarkable 53.1% (NMI) enhancement over LUKE. This performance delta further increases to an average of 73.2% (NMI) post the integration of data-specific features via CCL.

3.2 Contrastive Learning

Contrastive learning targets learning high-quality representations by pulling semantically similar instances towards each other and simultaneously repelling disparate ones within the embedded space [Hadsell et al., 2006]. Given a representative contrastive set $x, x^+, x_1^-, \ldots, x_{N-1}^-$ which contains one positive instance and $N-1$ counter instances, and the corresponding representations $\mathbf{h}, \mathbf{h}^+, \mathbf{h}1^-, \ldots, \mathbf{h}_{N-1}^-$ derived from the encoder, our approach is aligned with established contrastive learning frameworks [Sohn, 2016, Chen et al., 2020b, Gao et al., 2021a]. We leverage the cross-entropy measure as the pivotal objective function:

$$l = -\log \frac{e^{\text{sim}(\mathbf{h},\mathbf{h}^+)/\tau}}{e^{\text{sim}(\mathbf{h},\mathbf{h}^+)/\tau} + \sum_{i=1}^{N-1} e^{\text{sim}(\mathbf{h},\mathbf{h}_i^-)/\tau}} \tag{3.1}$$

where τ is a temperature hyperparameter and $\text{sim}(\mathbf{h}_1, \mathbf{h}_2)$ is the cosine similarity $\frac{\mathbf{h}_1^\top \mathbf{h}_2}{\|\mathbf{h}_1\| \cdot \|\mathbf{h}_2\|}$.

3.3 Phrase Encoder

We employ the transformer-based model, LUKE [Yamada et al., 2020], as our phrase encoder $\mathbf{E}$. LUKE is a state-of-the-art pre-trained language model capable of producing repre-

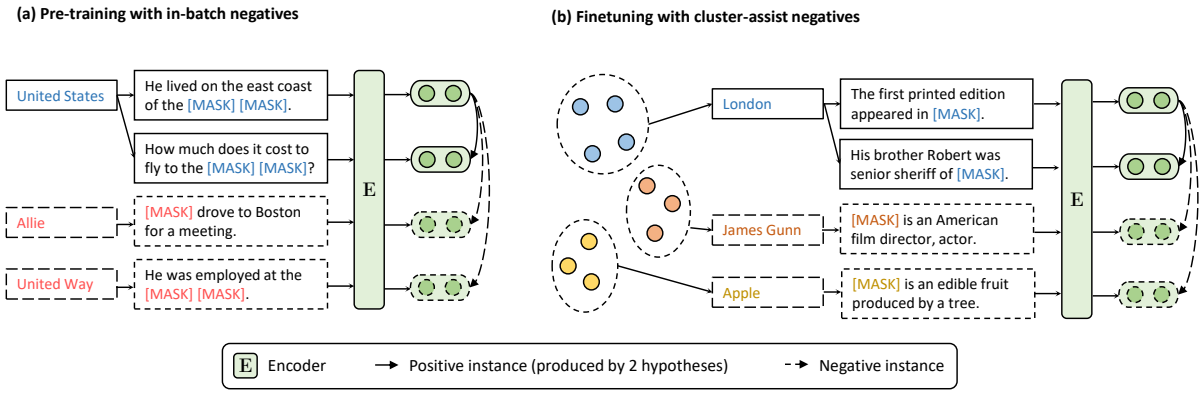


Figure 3.2. (a) Pre-training UCTopic on a large-scale dataset with positive instances from our two assumptions and in-batch negatives. (b) Finetuning UCTopic on a topic mining dataset with positive instances from our two assumptions and negatives from clustering.

sentations for both individual tokens and entire spans within sentences. Our designated phrase instance denoted as $x = (s, [l, r])$, encompasses a sentence s paired with a character-level span defined by $[l, r]$ boundaries. The encoder $\mathbf{E}$ processes the phrase x to yield the representation $\mathbf{h} = \mathbf{E}(x) = \mathbf{E}(s, [l, r])$. While LUKE can directly produce span representations, our findings suggest that its output may not effectively characterize phrases. Different from LUKE's methodology, which emphasizes entity prediction, our approach, denoted as UCTopic, emphasizes contrastive learning within phrase contexts. Consequently, representations derived from UCTopic are both context-sensitive and versatile across various domains.

3.4 Constructing Positive Instances

A challenge in contrastive learning revolves around the formulation of positive pairs (x, x^+). While antecedent literature [Wu et al., 2020, Meng et al., 2021] has leaned on augmentation stratagems such as word deletion, reordering, and paraphrasing, such techniques fall short in the realm of phrase representation learning. In this discourse, we harness the hypotheses introduced in Section 3.1 to generate positive instances apt for contrastive learning.

For clarity, let's delve into an illustrative example: As visualized in Figure 3.2 (a), the phrase `United States` is manifest in two distinct sentences: *He resided on the east coast of the*

United States" and *What are the expenses associated with traveling to the United States*". It is rationally anticipated that the representations of the phrase (`United States`) from these two statements resonate closely in semantic space. To steer the model towards harnessing context for deriving phrase semantics and concurrently obviate direct phrase mention comparisons in contrastive learning, we employ the `[MASK]` token to obscure the phrase mentions. These obfuscated sentences serve as the core positive instances. To harmonize the representational paradigms during both training and evaluation, within a positive pair, we probabilistically preserve one phrase mention intact, denoted by probability p.

In formal terms, assuming we are endowed with a phrase instance $x = (s, [l, r])$ and its accompanying positive counterpart $x^+ = (s', [l', r'])$ — where s symbolizes the sentence and $[l, r]$ demarcates the left and right confines of a phrase within s — we extract the phrase representations $\mathbf{h}$ and $\mathbf{h}^+$ via the encoder $\mathbf{E}$. Subsequently, in-batch negatives are harnessed for the pre-training phase. Thus, the primary training objective of UCTOPIC is delineated as follows:

3.5 Cluster-Assisted Contrastive Learning

Unlike pre-training on large-scale corpora, in-batch negatives often include instances semantically similar to the positives. To illustrate, consider a document with three distinct topics and a batch size of 32. Within such a batch, instances from the same topic are mistakenly treated as negatives, causing the contrastive learning process to receive ambiguous training signals and subsequently leading to reduced performance.

To mitigate the inaccuracies inherent in negatives and fine-tune phrase representations in line with document topics, we propose the cluster-assisted contrastive learning (CCL) approach. Basically, CCL leverages prior insights obtained from pre-trained representations combined with clustering to minimize negative noise. We initiate by understanding document topics using a clustering method based on pre-trained phrase representations from our model. These cluster centroids are then treated as topical representations for phrases. After evaluating the cosine

distance between phrase instances and the centroids, we allocate pseudo labels to the top t percent of instances proximate to these centroids. The label assigned to a particular phrase mention, denoted as p^m (where phrase mentions are sourced from sentence s such that $p^m = s[l:r]$), is resolved by the prevalent choice among instances $x_0^m, x_1^m, \ldots, x_n^m$ encompassing p^m, with n being the count of sentences with pseudo labels. Through this methodology, we procure preliminary insights on phrase mentions to bolster subsequent contrastive learning. As depicted in Figure 3.2 (b), three distinct phrase mentions—`London`, `James Gunn`, and `Apple`—originating from separate clusters are classified under varying topic categories.

Given a topic set $\mathscr{C}$ in our documents, we establish positive pairs $(x_{c_i}, x_{c_i}^+)$ for each topic $c_i \in \mathscr{C}$ using the method introduced in Section 3.4. For contrastive purposes, we extract phrases $p_{c_j}^m$ and instances $x_{c_j}^m$ from a different topic c_j to serve as negative instances $x_{c_j}^-$ during the contrastive learning process, ensuring that $c_j \in \mathscr{C} \wedge c_j \neq c_i$. As illustrated in Figure 3.2 (b), we forge positive pairs centered around the phrase `London`, and from two disparate clusters, we randomly extract negative instances based on the phrases `James Gunn` and `Apple`. With the guidance of pseudo labels, our methodology is adept at sidestepping instances semantically akin to `London`. The ultimate objective of the finetuning phase is:

$$l = -\log \frac{e^{\text{sim}(\mathbf{h}_{c_i}, \mathbf{h}_{c_i}^+)/\tau}}{e^{\text{sim}(\mathbf{h}_{c_i}, \mathbf{h}_{c_i}^+)/\tau} + \sum_{c_j \in \mathscr{C}} e^{\text{sim}(\mathbf{h}_{c_i}, \mathbf{h}_{c_j}^-)/\tau}}. \tag{3.2}$$

Regarding the masking approach during pre-training, we apply masking across all training instances. However, with a probability p, we leave both $x_{c_i}^+$ and $x_{c_j}^-$ unaltered.

The topic y of a phrase x is determined by calculating the cosine similarity between the phrase representation $\mathbf{h}$ and various topic representations $\tilde{\mathbf{h}}_{c_i}, c_i \in \mathscr{C}$, where $c_i \in \mathscr{C}$. The topic that is most similar to x is chosen as the phrase topic. Formally,

$$y = \text{argmax}_{c_i \in \mathscr{C}}(\text{sim}(\mathbf{h}, \tilde{\mathbf{h}}_{c_i})) \tag{3.3}$$

3.6 Experiments

We conduct our experiments for phrase representation evaluation from two aspects: (1) entity clustering and (2) topical phrase mining.

3.6.1 Entity Clustering

To evaluate the effectiveness of phrase representations, we first use UCTOPIC for entity clustering and then contrast it with other representation learning techniques.

Datasets. We perform entity clustering on four datasets that include annotated entities from various domains: general, review, and biomedical. These datasets are:

1. CoNLL2003 [Sang and Meulder, 2003] has 20,744 sentences taken from Reuters news articles. For our experiments, we consider Person, Location, and Organization entities. [1]

2. BC5CDR [Li et al., 2016a] is part of the BioCreative V CDR task corpus, encompassing 18,307 sentences from PubMed articles. This dataset contains 15,953 chemical entities and 13,318 disease entities.

3. MIT Movie (MIT-M) [Liu et al., 2013] includes 12,218 sentences, highlighting Title and Person entities.

4. W-NUT 2017 [Derczynski et al., 2017] is centered on identifying unique entities in emerging discussions. It comprises 5,690 sentences with six different entities. [2]

Baselines. To showcase the effectiveness of our pre-training approach combined with cluster-assisted contrastive learning (CCL), we evaluate it against baseline methods in two primary areas:

(1) Pre-trained token or phrase representations:

[1]The Misc category isn't evaluated since it doesn't correspond to a distinct semantic category.

[2]These are corporation, creative work, group, location, person, and product.

- **Glove** [Pennington et al., 2014]. These are pre-trained word embeddings created from 6B tokens with a dimensionality of 300. Phrase representations are derived by averaging these word embeddings.

- **BERT** [Devlin et al., 2019a]. Phrase representations can either be acquired by averaging token representations, known as BERT-Ave, or by using the CGExpan approach [Zhang et al., 2020d], where phrases are replaced with the `[MASK]` token. The representation of the `[MASK]` token is then used as the phrase embedding, termed BERT-MASK.

- **LUKE** [Yamada et al., 2020]. It is employed as a foundational model to demonstrate the potency of our pre-training and finetuning using contrastive learning.

- **DensePhrase** [Lee et al., 2021]. This is a supervised approach to pre-training phrase representations, mainly used for the question-answering task. We leverage the model provided by the authors to extract phrase representations.

- **Phrase-BERT** [Wang et al., 2021]. This approach yields context-independent phrase representations through pre-training. We utilize a model given by the authors and retrieve representations based on phrase mentions.

- **Ours w/o CCL**. This refers to the pre-trained phrase representations from our model but without leveraging the cluster-assisted contrastive finetuning.

(2) Fine-tuning techniques that build on the pre-trained representations of our model:

- **Classifier**. Using pseudo labels for guidance, we train an MLP layer to produce a classifier for phrase categories.

- **In-Batch Contrastive Learning**. This mirrors the contrastive learning used during pre-training, utilizing in-batch negatives.

- **Autoencoder**. This method has seen extensive use in prior neural topic and aspect extraction models [He et al., 2017, Iyyer et al., 2016, Tulkens and van Cranenburgh, 2020].

Table 3.1. Performance of entity clustering on four datasets from different domains. *Class.* represents using a classifier on pseudo labels. *Auto.* represents Autoencoder. The best results among all methods are bolded and the best results of pre-trained representations are underlined. *In-B.* represents contrastive learning with in-batch negatives.

Datasets	**CoNLL2003**		**BC5CDR**		**MIT-M**		**W-NUT2017**	
Metrics	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Pre-trained Representations								
Glove	0.528	0.166	0.587	0.026	0.880	0.434	0.368	0.188
BERT-Ave.	0.421	0.021	0.857	0.489	0.826	0.371	0.270	0.034
BERT-Mask	0.430	0.022	0.551	0.001	0.587	0.001	0.279	0.020
LUKE	0.590	0.281	0.794	0.411	0.831	0.432	0.434	0.205
DensePhrase	0.603	0.172	0.936	0.657	0.716	0.293	0.413	0.214
Phrase-BERT	0.643	0.297	0.918	0.617	<u>0.916</u>	<u>0.575</u>	0.452	0.241
Ours w/o CCL	<u>0.704</u>	<u>0.464</u>	<u>0.977</u>	<u>0.846</u>	0.845	0.439	<u>0.509</u>	<u>0.287</u>
Finetuning on Pre-trained UCTOPIC *Representations*								
Ours w/ Class.	0.703	0.458	0.972	0.827	0.738	0.323	0.482	0.283
Ours w/ In-B.	0.706	0.470	0.974	0.834	0.748	0.334	0.454	0.301
Ours w/ Auto.	0.717	0.492	0.979	0.857	0.858	0.458	0.402	0.282
UCTOPIC	**0.743**	**0.495**	**0.981**	**0.865**	**0.942**	**0.661**	**0.521**	**0.314**

We adopt the ABAE strategy [He et al., 2017] to implement our phrase-based autoencoder model.

Overall Performance.

Evaluation results for entity clustering are presented in Table 3.1. Across all datasets and metrics, our model has the best performance. Specifically, our method overperforms the state-of-the-art, Phrase-BERT, by an average of 38.2% NMI and outperforms our foundational model, LUKE, by 73.2% NMI.

Analyzing various pre-trained representations, our approach (Ours w/o CCL) outperforms other baselines on three out of four datasets, with MIT-M being the exception. This discrepancy is attributed to two factors:

1. The MIT-M dataset uses exclusively lowercase words, which differs from our pretraining dataset. This misalignment between training and testing leads to diminished performance.

Table 3.2. Ablation study on the input of phrase instances of W-NUT 2017. UCTOPIC here is pre-trained representations without CCL finetuning. Percentages in brackets are changes compared to Context+Mention.

Model	**UCTopic**		**LUKE**	
Metric	ACC	NMI	ACC	NMI
Context+Mention	0.44	0.29	0.39	0.21
Mention	0.32 (-27%)	0.15 (-48%)	0.28 (-28%)	0.10 (-52%)
Context	0.43 (-3%)	0.16 (-44%)	0.27 (-31%)	0.07 (-67%)

2. Sentences in MIT-M are characteristically shorter, averaging 10.16 words, especially when juxtaposed with other datasets like W-NUT2017, which averages 17.9 words. As a result, our model garners limited context from these abbreviated sentences.

Nevertheless, the performance degradation arising from these factors is alleviated by employing our CCL finetuning. Specifically, on the MIT-M dataset, our model posts superior results (0.661 NMI) compared to Phrase-BERT (0.575 NMI) post-CCL application.

Furthermore, in contrast to other finetuning techniques, our CCL finetuning enhances the quality of pre-trained phrase representations by focusing on dataset-specific attributes. The enhancement reaches as high as 50% NMI for the MIT-M dataset. The performance of Ours w/ Class. often trails that of our pre-trained model, suggesting that clustering-derived pseudo labels can introduce noise, undermining their utility as direct supervision for representation learning. Ours w/ In-B. exhibits behavior akin to Ours w/ Class., underscoring our rationale for choosing CCL over in-batch negatives. While autoencoders can augment pre-trained representations on three datasets, the gains remain modest, with performance even declining on W-NUT2017. Compared to other finetuning strategies, our CCL finetuning uniformly enhances pre-trained phrase representations across diverse domains.

Context or Mentions. To delve into the origins of our model phrase semantics, whether it is from phrase mentions or context, we execute an ablation study contrasting our method with

LUKE. To ensure the clustering results are not affected by recurrent phrase mentions, we use a singular phrase instance (that is, the sentence and the position of a phrase) for every phrase mention. As outlined in Table 3.2, we consider three types of inputs:

1. Context+Mention: This mirrors the input used in experiments shown in Table 3.1, incorporating the full sentence encompassing the phrase.

2. Mention: Only the phrase mentions serve as inputs for both models.

3. Context: Phrase mentions within sentences are obscured, so models solely derive information from the surrounding context.

From the results, it is evident that our model obtains more insight from context (0.43 ACC, 0.16 NMI) compared to mentions (0.32 ACC, 0.15 NMI). When compared to LUKE, our method shows greater resilience to the absence of phrase mentions. Specifically, when predicting based solely on context, our model performance dips modestly (-3% ACC and -44% NMI) relative to LUKE's steeper declines (-31% ACC and -67% NMI).

3.6.2 Topical Phrase Mining

Dataset. We perform topical phrase mining on three datasets including news, review, and computer science domains:

- **Gest**. This dataset comprises restaurant reviews sourced from Google Local [3]. We utilize 100K reviews, which translates to 143,969 sentences, for our topical phrase mining.

- **KP20k** [Meng et al., 2017] is an aggregation of titles and abstracts from computer science publications. For our experiments, we employ 500K sentences.

- **KPTimes** [Gallina et al., 2019] consists of news pieces from the New York Times (spanning 2006 to 2017) and an additional 10K articles from the Japan Times. Our topical phrase mining exploits 500K sentences from this dataset.

Table 3.3. The numbers of topics in three datasets.

Datasets	**Gest**	**KP20k**	**KPTimes**
# of topics	22	10	16

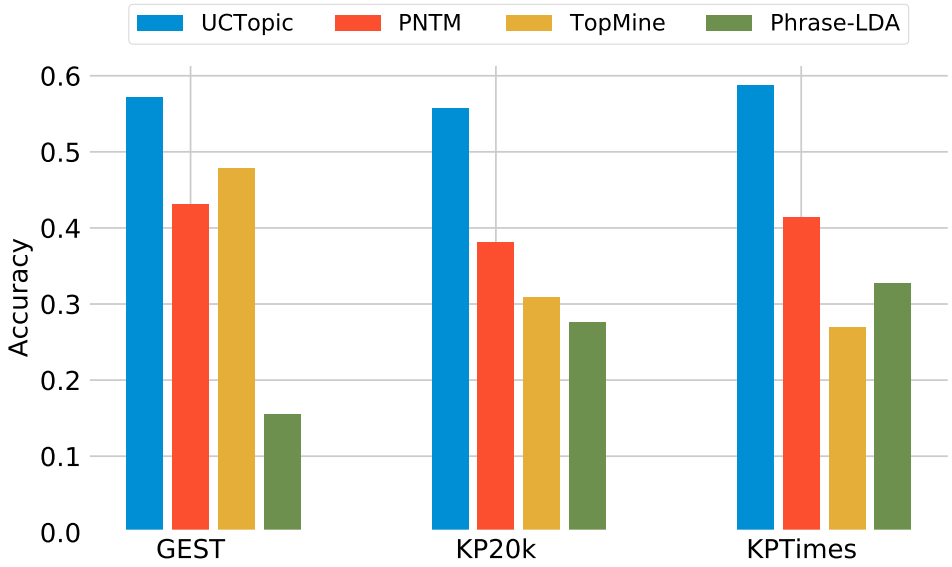


Figure 3.3. Results of phrase intrusion task.

The number of topics determined by Silhouette Coefficient is shown in Table 3.3.

Compared Baseline Methods. We compare our method with three topic model benchmarks:

- **Phrase-LDA** [Mimno, 2015]. This LDA variant integrates phrases by transforming them into unigrams. For instance, "city view" becomes "city_view".
- **TopMine** [El-Kishky et al., 2014]. This scalable approach first segments a document into phrases. Subsequently, it employs these phrases as guidelines to ensure consistent topic allocation for all constituent words.
- **PNTM** [Wang et al., 2021]. This contemporary topic model employs Phrase-BERT in conjunction with an autoencoder, aiming to recreate document representations. It is recognized as a state-of-the-art topic modeling approach.

[3]https://www.google.com/maps

Table 3.4. Number of coherent topics on Gest and KP20k.

	UCTopic	PNTM	TopMine	P-LDA
Gest	20	18	20	11
KP20k	10	9	9	4

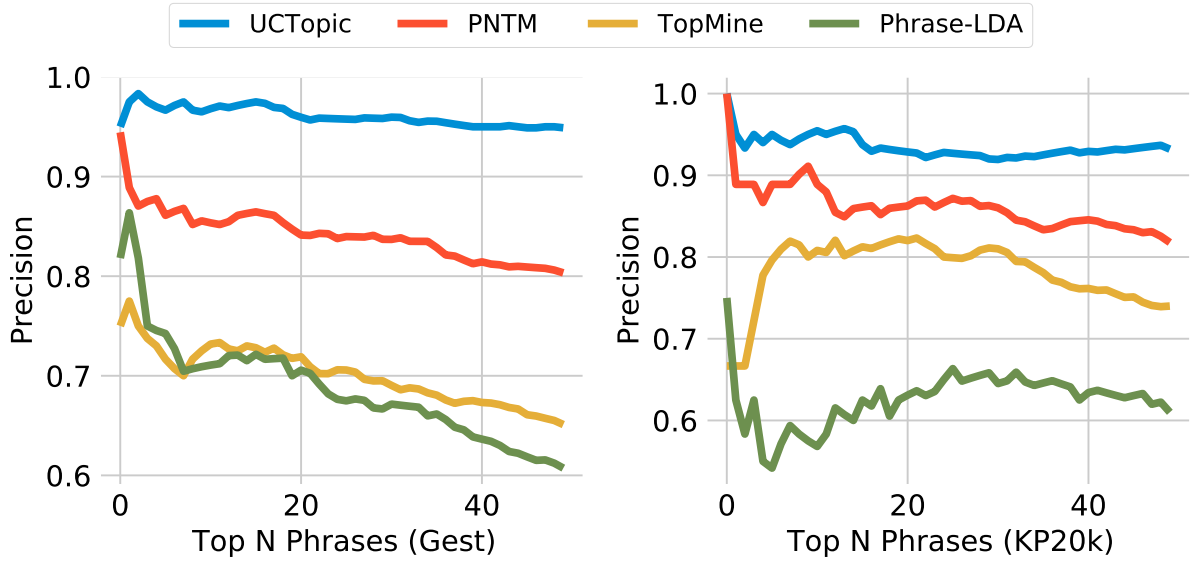


Figure 3.4. Results of top n precision.

Topical Phrase Evaluation. We assess the quality of topical phrases based on three dimensions:

1. *topical separation*;
2. *phrase coherence*;
3. *phrase informativeness and diversity*.

To assess *topical separation*, we undertake the **phrase intrusion** task, as outlined in prior studies [El-Kishky et al., 2014, Chang et al., 2009]. This task requires human participants to identify an 'intruder' phrase from a set of phrases. In our setup, every question presents 6 phrases: 5 are randomly selected from the top 50 phrases associated with a specific topic, while the remaining phrase is randomly picked from another topic's top 50 phrases. Participants are tasked with pinpointing the outlying phrase. For each method and dataset, we generate 50 such questions, culminating in 600 questions overall. All questions are shuffled. Given that

Table 3.5. Informativeness (tf-idf) and diversity (word-div.) of extracted topical phrases.

Datasets	**Gest**		**KP20k**	
Metrics	tf-idf	word-div.	tf-idf	word-div.
TopMine	**0.5379**	0.6101	0.2551	0.7288
PNTM	0.5152	0.5744	**0.3383**	0.6803
UCTopic	0.5186	**0.7486**	0.3311	**0.7600**

Table 3.6. Top topical phrases on Gest and KP20k and the minimum phrase frequency is 3.

Gest					**KP20k**	
Drinks		*Dishes*			*Programming*	
UCTopic	PNTM	UCTopic	PNTM	TopMine	UCTopic	TopMine
lager	drinks	cauliflower fried rice	great burger	mac cheese	markup language	software development
whisky	bar drink	chicken tortilla soup	great elk burger	ice cream	scripting language	software engineering
vodka	just drink	chicken burrito	great hamburger	potato salad	language construct	machine learning
whiskey	alcohol	fried calamari	good burger	french toast	java library	object oriented
rum	liquor	roast beef sandwich	good hamburger	chicken sandwich	programming structure	open source
own beer	booze	grill chicken sandwich	awesome steak	cream cheese	xml syntax	design process
ale	drink order	buffalo chicken sandwich	burger joint	fried chicken	module language	design implementation
craft cocktail	ok drink	pull pork sandwich	woody 's bbq	fried rice	programming framework	programming language
booze	alcoholic beverage	chicken biscuit	excellent burger	french fries	object-oriented language	source code
tap beer	beverage	tortilla soup	beef burger	bread pudding	python module	support vector machine

each question is uniquely generated, we engage 4 evaluators to respond, with each addressing approximately 150 questions. The outcome gauges the effectiveness of topical phrase separation. As depicted in Figure 3.3, UCTopic surpasses other benchmarks across the three datasets, indicating its prowess in discerning distinct topics within texts.

To assess *phrase coherence* within a given topic, we adopt the methodology from ABAE [He et al., 2017]. Annotators are tasked with determining whether the top 50 phrases of a topic are coherent, meaning the majority of phrases align with the same thematic topic. This evaluation involves 3 annotators reviewing four different models on the Gest and KP20k datasets. The count of topics deemed coherent is presented in Table 3.4. From the results, UCTopic, PNTM, and TopMine exhibit comparable tallies for coherent topics. However, Phrase-LDA lags behind these three. For topics identified as coherent, every top phrase is marked as accurate if it mirrors the respective topic. Aligning with ABAE's approach, we utilize *precision@n* for the evaluation. The outcomes, displayed in Figure 3.4, reveal that UCTopic consistently surpasses the other models, maintaining elevated precision even for larger values of n, while the precision

of rival models diminishes.

To evaluate *phrase informativeness and diversity*, we employ tf-idf and word diversity (word-div.) metrics on top topical phrases. Essentially, informative phrases should not be overly common within a corpus (e.g., "good food" in Gest). We utilize tf-idf to ascertain the "importance" of a phrase. To account for the variance in phrase lengths, we compute the average word tf-idf in a phrase to determine the phrase's tf-idf. Specifically, the equation is given by: $\text{tf-idf}(p,d) = \frac{1}{m}\sum_{1\leq i\leq m}\text{tf-idf}(w_i^p)$ where d symbolizes the document, and p represents the phrase. For our experiments, a document refers to a sentence within a review. Furthermore, it's desirable for our topical phrases to be varied within a topic rather than being repetitive or conveying the same meaning (e.g., "good food" versus "great food"). To assess the diversity of the leading phrases, we determine the proportion of unique words among all words in the phrase. To be precise, given a list of phrases $[p_1, p_2, \ldots, p_n]$, we segment these phrases into a word list represented as $\mathbf{w} = [w_1^{p_1}, w_2^{p_1}, \ldots, w_m^{p_n}]$. Let $\mathbf{w}'$ denote the set of distinct words within $\mathbf{w}$. The word diversity is then calculated by: $\frac{|\mathbf{w}'|}{|\mathbf{w}|}$ It's important to note that we restrict our evaluation to only the coherent topics identified in the *phrase coherence* section. Since Phrase-LDA has fewer coherent topics compared to the other models, our evaluation focuses primarily on the remaining three models.

We calculate the tf-idf and word-div. for the top 10 phrases and take the average value across topics to derive the final scores. The outcomes are presented in Table 3.5. Both PNTM and UCTopic have comparable tf-idf scores, attributable to their shared phrase lists sourced from spaCy. UCTopic identifies the most diverse set of phrases within a topic due to its more context-sensitive phrase representations. Conversely, since PNTM's representations are heavily influenced by phrase mentions, its extracted phrases often share the same words, leading to reduced diversity.

Case Study. We examine the top phrases from UCTopic, PNTM, and TopMine as shown in Table 3.6. The examples align with our user study and diversity assessment. While PNTM's phrases are coherent, they display less diversity than the others, with similar phrases such

as “drinks”, “bar drink”, and “just drink” from Gest being evident due to context-agnostic representations grouping alike mentions. TopMine’s phrases are diverse but sometimes lack coherence, as seen with “machine learning” and “support vector machine” in the programming topic. In comparison, UCTopic successfully extracts topical phrases from documents that are both coherent and diverse.

3.7 Conclusion

We introduce UCTopic, a contrastive learning framework designed to effectively learn phrase representations without any supervision. To enhance performance on topic mining datasets, we introduce cluster-assisted contrastive learning, which refines results by choosing negatives from specific clusters. This fine-tuning process optimizes our phrase representations according to topics in documents, further enhancing their quality. Our extensive experiments on entity clustering and topical phrase mining demonstrate that UCTopic significantly enhances phrase representations. Both objective metrics and a user study reveal that UCTopic successfully extracts topical phrases that are both coherent and diverse.

Chapter 3, in part, is a reprint of the material as it appears in “UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining” by Jiacheng Li, Jingbo Shang, Julian McAuley, which was published in *Annual Meeting of the Association for Computational Linguistics*, 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Justifying Recommendations Using Distantly Labeled Reviews and Fine-grained Aspects

In this chapter, we will introduce the explanation generation. Specifically, we present a work on topic-based controllable explanation generation. Recent studies have delved into generating reviews or 'tips' as explanations to validate why a recommendation aligns with user preferences. However, we find that current methods often fall short in generating relevant justifications in line with users' decision-making criteria. To tackle this recommendation justification challenge, we introduce novel datasets and methods. For data acquisition, we employ an 'extractive' technique to pinpoint review segments that reveal users' intentions. Using this strategy, we distantly label extensive review corpora, paving the way for creating large-scale personalized recommendation justification datasets. Regarding generation, we propose a reference-based Seq2Seq model incorporating aspect-planning for versatile aspect coverage.

4.1 Introduction

Previous research has focused on understanding user preferences and writing styles from crowd-sourced reviews [Dong et al., 2017, Ni and McAuley, 2018] to produce explanations via natural language. However, a significant segment of the review content, including 'tips', often doesn't significantly influence users' decision-making, as they might detail lengthy experiences

Table 4.1. In contrast to reviews and tips, we seek to automatically generate *recommendation justifications* that are more concise, concrete, and helpful for decision making. Examples of justifications from reviews, tips, and our annotated dataset are marked in bold.

Review examples:
I love this little stand! **The coconut mocha chiller and caramel macchiato are delicious.**
Wow what a special find. One of the most unique and special date nights my husband and I have had.
Tip examples:
Great food. Nice ambiance. Gnocchi were very good.
I can't get enough of this place.
Justification examples:
The food portions were huge.
Plain cheese quesadilla is very good and very cheap.

or general endorsements. Therefore, models that primarily learn from these reviews might overlook vital details that elucidate users' purchasing choices. Table 4.1 presents examples of reviews, tips, and optimal justifications. A newer avenue of research has delved into tip generation, wherein tips are summaries of reviews [Li et al., 2017a]. Although tips provide a more concise perspective and some may be apt for recommendation justifications, only a handful of platforms feature tips alongside reviews.

Producing *diverse* outputs is crucial in personalized content generation, especially in justification generation. Rather than consistently suggesting the most prevalent reasons, delivering diverse justifications tailored to individual user interests is more desirable. Recent studies indicate that integrating prior knowledge into generation systems can significantly enhance diversity. Such prior knowledge might encompass story-lines for story generation [Yao et al., 2019] or historical responses within dialogue systems [Weston et al., 2018].

Our objective is to produce compelling and varied justifications. Given the challenge of not having ground-truth data for ideal justifications, we introduce a method to pinpoint justifications within extensive review or tip collections. We derive specific aspects from these justifications and create user personas and item profiles composed of characteristic aspects. For enhanced generation quality and variety, we propose a reference-based Seq2Seq model equipped

with aspect-planning, which uses prior justifications for context and can craft justifications centered around diverse aspects.

4.2 Justification

In this section, we detail our approach to extracting top-quality justifications from user reviews. We aim to extract review segments suitable as justifications and subsequently construct a personalized justification dataset. Our approach is structured into three stages:

1. Marking a collection of review segments with binary labels to identify them as either 'good' or 'bad' justifications.

2. Using the labeled subset to train a classifier, which is then utilized to label the entirety of review segments, extracting suitable justifications for every user-item combination.

3. Implementing detailed aspect extraction on the identified justifications and formulating user personas along with item profiles.

Identifying Justifications From Reviews The initial step involves extracting text segments from reviews suitable for justifications. We define each segment as an Elementary Discourse Unit (EDU [Mann and Thompson, 1988]), which corresponds to a sequence of clauses. The model by Wang et al. [2018] is employed to derive EDUs from reviews. We examine the linguistic differences between recommendation justifications and reviews. Based on our analysis, we established two rules to eliminate segments likely unsuitable as justifications: (1) segments containing first-person or third-person pronouns, and (2) segments that are either too lengthy or too brief. Subsequently, two expert annotators evaluated 1,000 segments that passed our filters to assess if they qualified as 'good' justifications. This labeling was conducted iteratively, with ongoing feedback and discussions, to ensure consistency between the annotators.

Automatic Classification The subsequent step involved propagating labels throughout the entire review corpus. For this, we employed BERT [Devlin et al., 2019b] and fine-tuned it for our

Table 4.2. Examples of justifications with fine-grained aspects in our annotated dataset. The fine-grained aspects are italic and underlined.

Yelp
The *Tuna* is pretty amazing
Appetizers and *pasta* are excellent here
An excellent *selection* of both *sweet* and savory *crepes*
It was filled with delicious *food*, fantastic *music* and *dancing*
Amazon-Clothing
The *quality* of the *material* is great
Great *shirt*, especially for the *price*.
The *seams* and *stitching* are really nice
Fit the bill for a *Halloween costume*.

classification task. A [CLS] token was prefixed to every segment, and the final hidden state corresponding to this token was processed through a linear layer to obtain the binary prediction, with cross-entropy serving as the training loss. We fine-tuned the BERT classifier using the Train set and selected the best-performing model based on the Dev set. After three epochs, the BERT model achieved an F1-score of 0.80 on the Test set.

Fine-grained Aspect Extraction Fine-grained aspects refer to the specific properties of products that are present in user opinions. Using the approach suggested by Zhang et al. [2014a], we construct a sentiment lexicon that encompasses a collection of these aspects sourced from the entire dataset. Simple rules are employed to identify which aspects are present in a given justification. Table 4.2 showcases some samples from our dataset, with each instance featuring a justification written by a user about an item, accompanied by the corresponding fine-grained aspects referenced in the justification. It's crucial to mention that our annotations were limited to the Yelp dataset. We trained a classifier on this and subsequently applied it to both the Yelp and Amazon Clothing datasets. As evidenced by Table 4.2, the classifier exhibits commendable performance across both datasets.

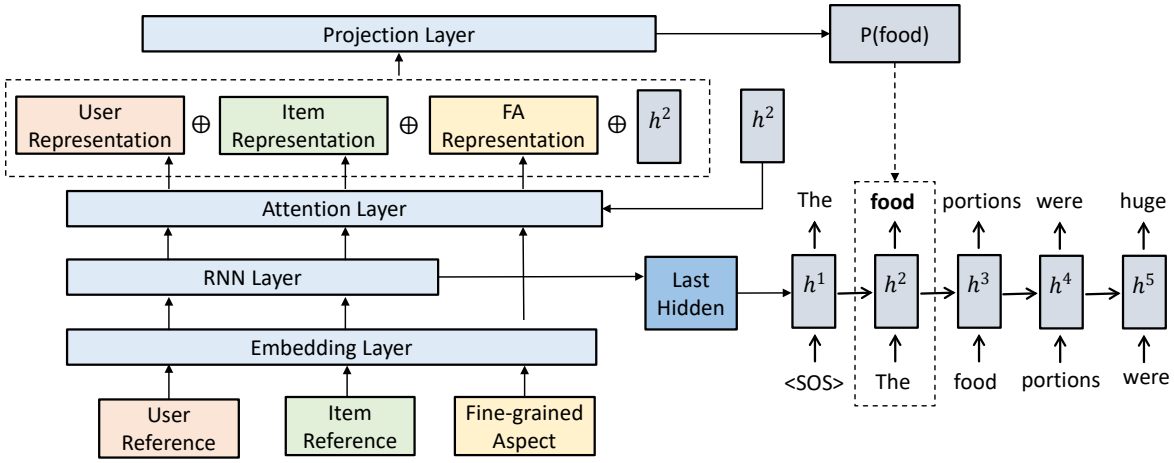


Figure 4.1. Structure of the reference-based Seq2Seq model with Aspect Planning

4.3 Reference-based Seq2Seq Model

Our foundational model is built upon the architecture of a Seq2Seq [Sutskever et al., 2014]. In our framework, termed 'Ref2Seq', the past justifications of users and items serve as references from which latent personalized features are derived. The architecture of our Reference-based Seq2Seq Model is illustrated in Figure 4.1. The model comprises two main components: (1) dual sequence encoders that deduce user and item latent features by referencing prior justifications; and (2) a sequence decoder that integrates these user and item representations to yield personalized justifications.

Sequence Encoders. The design of our user encoder and item encoder is identical, encompassing an embedding layer, a two-layer bi-directional GRU [Cho et al., 2014], and a subsequent projection layer. The input is a user (or item) reference D, which represents a collection of historical justifications. After being processed by the embedding layer, these justifications are relayed through the GRU, resulting in a sequence of hidden states $\mathbf{e} \in \mathbb{R}^{l_s \times l_r \times n}$:

$$\mathrm{E} = \mathrm{Embedding}(D), \mathbf{e} = \mathrm{GRU}(\mathrm{E}) = \overrightarrow{\mathrm{e}} + \overleftarrow{\mathrm{e}}, \tag{4.1}$$

where l_s represents the length of the sequence, n indicates the hidden size of the encoder

GRU, $\mathrm{E} \in \mathbb{R}^{l_s \times l_r \times n}$ signifies the embedded sequence representation, and $\overrightarrow{e}$ and $\overleftarrow{e}$ are the hidden vectors generated by the forward and backward GRU, respectively.

To combine information from various justifications, also referred to as 'references', the hidden states are subsequently processed through a linear layer:

$$\hat{\mathbf{e}} = W_e \cdot \mathbf{e} + \mathbf{b}_e, \tag{4.2}$$

where $\hat{\mathbf{e}} \in \mathbb{R}^{l_s \times n}$ is the final output of the encoder, and $W_e \in \mathbb{R}^{l_r}$, $\mathbf{b}_e \in \mathbb{R}$ are learned parameters.

Sequence Decoder. The decoder utilizes a two-layer GRU to forecast the target words beginning with a start token. The decoder's initial hidden state is derived from the combined final hidden states of both the user and item encoders. The hidden state at the t-th time-step is refreshed through the GRU based on the preceding hidden state and the input word, as follows:

$$\mathbf{h}_0 = \mathbf{e}^u_{l_s} + \mathbf{e}^i_{l_s}, \mathbf{h}_t = \mathbf{GRU}(w_t, \mathbf{h}_{t-1}), \tag{4.3}$$

where $\mathbf{e}^u_{l_s}$ and $\mathbf{e}^i_{l_s}$ represent the final hidden states from the user and item encoder outputs, respectively, denoted as $\hat{\mathbf{e}}_u$ and $\hat{\mathbf{e}}_i$.

To understand the connection between the reference and the generated output, we employ an attention fusion layer to aggregate the encoder outputs. For both user and item reference encoders, the attention vector is formulated as follows:

$$\begin{aligned} \mathbf{a}^1_t &= \sum_{j=1}^{l_s} \alpha^1_{tj} \mathbf{e}_j, \\ \alpha^1_{tj} &= \exp(\tanh({\mathbf{v}^1_\alpha}^\top (W^1_\alpha [\mathbf{e}_j; \mathbf{h}_t] + \mathbf{b}^1_\alpha)))/Z, \end{aligned} \tag{4.4}$$

where $\mathbf{a}^1_t \in \mathbb{R}^n$ represents the attention vector on the sequence encoder at time-step t. The term α^1_{tj} signifies the attention score between the encoder hidden state $\mathbf{e}_j$ and the decoder hidden state $\mathbf{h}_t$. Meanwhile, Z serves as a normalization factor.

Table 4.3. Statistics of our datasets.

Dataset	Train	Dev	Test	# Users	# Items	# Aspects
Yelp	1,219,962	115,907	115,907	115,907	51,948	2,041
Amazon Clothing	202,528	57,947	57,947	57,947	50,240	581

Aspect-Planning Generation. A challenge in generating justifications is enhancing controllability, specifically the ability to directly influence the generated content. Drawing inspiration from the 'plan-and-write' approach [Yao et al., 2019], we have augmented our base model into the Aspect-Planning Ref2Seq (AP-Ref2Seq). In this model, a fine-grained aspect is planned prior to generation. Rather than a stringent constraint, this aspect planning acts as an added layer of guidance, ensuring that the justification generation process is more controllable.

To generate a justification for a user u and item i, we initiate with a designated fine-grained aspect a. This aspect is then processed through the word embedding layer, producing the aspect embedding, denoted as $\mathrm{E_a}$. Subsequently, we determine the scores between the aspect's embedding and the decoder's hidden state using the following relationship:

$$\begin{aligned} \mathbf{a}_t^2 &= \alpha_t^2 \mathrm{E}_a, \\ \alpha_t^2 &= \exp(\tanh({\mathbf{v}_\alpha^2}^\top (W_\alpha^2 [\mathrm{E}_a; \mathbf{h}_t] + \mathbf{b}_\alpha^2)))/Z, \end{aligned} \tag{4.5}$$

where $\mathbf{a}_t^2 \in \mathbb{R}^n$ is an attention vector and α_t^2 is an attention score.

The attention vectors, specifically $\mathbf{a}_{ut}^1$ for user u, $\mathbf{a}_{it}^1$ for item i, and $\mathbf{a}_t^2$ for the fine-grained aspect a, are combined with the decoder's hidden state at time-step t. This amalgamation is then projected to derive the distribution P for the output word. The likelihood for word w at time-step t can be described as:

$$p(w_t) = \tanh(W_1 [\mathbf{h}_t; \mathbf{a}_{ut}^1; \mathbf{a}_{it}^1; \mathbf{a}_t^2] + \mathbf{b}_1), \tag{4.6}$$

where w_t represents the desired word at time-step t. Using the probability $p(w_t)$ for each time step t, the model is trained with a cross-entropy loss relative to the ground-truth sequence.

4.4 Experiments

Dataset We build two personalized justification datasets from existing review sources: Yelp and Amazon Clothing. [1] [2] We further refine the datasets by filtering out users with less than five justifications. For every user, two samples are randomly selected from their justifications to create the Dev and Test sets. The statistics of our two datasets are presented in Table 4.3.

Baselines For automatic evaluation, we include three baselines: Item-Rand, which selects a justification at random from an item's historical justifications; LexRank, an unsupervised approach frequently employed in text summarization [Erkan and Radev, 2004], which selects a justification as the summary from all historical justifications about an item and uses this as the justification for all users; and Attr2Seq [Dong et al., 2017], a Seq2Seq method that takes attributes, specifically user and item identity, as input.

While all models typically use beam search for generation, some recent studies suggest that the outputs from sampling methods are more diverse and apt for high-entropy tasks [Holtzman et al., 2019b]. Therefore, we also experiment with another decoding approach, 'Top-k sampling' [Radford et al., 2019], and introduce a variant of our model named Ref2Ref (Top-k). [3]

For human evaluation, we introduce two baselines: Ref2Seq (Review) and Ref2Seq (Tip). Both are identical to the Ref2Seq model but are trained using original review and tip data, respectively. Through these comparisons, we aim to demonstrate that models trained on our annotated dataset tend to produce text more aptly suited as justifications.

Overall Performance For automatic evaluation, we employ BLEU, Distinct-1, and Distinct-2 metrics [Li et al., 2015b] to evaluate the performance of our model. As indicated in Table 4.4, our reference-based models outperform in BLEU scores across both datasets, with the exception of BLEU-3 on Yelp. This suggests that Ref2Seq effectively harnesses user and item content to generate highly relevant content, in contrast to unpersonalized models like LexRank and

[1] https://www.yelp.com/dataset/challenge

[2] http://jmcauley.ucsd.edu/data/amazon

[3] For each time step, the next word is chosen from the top k most probable next tokens.

Table 4.4. Performance on Automatic Evaluation.

Dataset	Yelp				Amazon Clothing			
Model	BLEU-3	BLEU-4	Distinct-1	Distinct-2	BLEU-3	BLEU-4	Distinct-1	Distinct-2
Item-Rand	0.440	0.150	2.766	20.151	1.620	0.680	2.400	11.853
LexRank	2.290	0.920	1.738	8.509	3.480	2.250	2.407	14.956
Attr2seq	**7.890**	0.000	0.049	0.095	1.720	0.560	0.076	0.352
Ref2Seq	4.380	**2.450**	0.188	1.163	8.780	5.670	0.141	1.240
AP-Ref2Seq	3.390	1.830	0.326	2.094	**13.910**	**12.500**	0.557	3.661
Ref2Seq (Top-k)	1.630	0.700	0.818	11.927	3.960	2.130	0.697	**10.858**

Table 4.5. Performance on Human Evaluation, where R,I,D represents Relevance, Informativeness and Diversity, respectively.

Model	R	I	D
Ref2Seq (Review)	3.02	2.39	2.10
Ref2Seq (Tip)	3.25	2.35	2.34
Ref2Seq	3.87	3.13	2.96
Ref2Seq (Top-k)	**3.95**	**3.34**	3.39
ACMLM	3.23	3.29	**3.42**

personalized models like Attr2Seq which do not utilize historical justifications.

Conversely, it has been noted in recent studies that models with greater output diversity tend to score lower on overlap-centric metrics such as BLEU for open-domain generation tasks [Baheti et al., 2018, Gao et al., 2018]. We observe a similar trend in our personalized justification generation task. As illustrated in Table 4.4, sampling-based approaches like Ref2Seq (Top-k) has higher Distinct-1 and Distinct-2 values, yet its BLEU scores trail behind Seq2Seq models that deploy beam search.

Human Evaluation We conduct human evaluation focusing on three criteria: (1) *Relevance*, which assesses the pertinence of the generated output to an item; (2) *Informativeness*, which gauges the specificity and utility of the information in the generated justification for users; and (3) *Diversity*, which evaluates the uniqueness of the generated output in comparison to other justifications.

Our attention is centered on the Yelp dataset, from which we extract 100 generated samples for each of the five models, as depicted in Table 4.5. Human evaluators are tasked with

Table 4.6. Comparisons of the generated justifications from different models for three businesses on the Yelp dataset.

Model	Shake Shack	Teharu Sushi	MGM Grand Hotel
Ground Truth	The burger was good	The rolls are pretty great , typical rolls not that many specials	Room was very clean comfortable
LexRank	A great burger and fries.	Sushi ?	Great rooms.
Ref2Seq (Review)	i love trader joe 's , i love trader joe 's	the food was good and the service was great	i love this place ! the food is always good and the service is always great
Ref2Seq (Tip)	this place is awesome	love this place	come here
Ref2Seq	this place has some of the best burgers	the sushi is delicious	the room was nice
Ref2Seq (Top-k)	the fries are amazing	fresh and delicious sushi	open hotel for hours
ACMLM	breakfast sandwiches are overall very filling	overall fun experience with half price sushi	family style dinner , long time shopping trip to vegas, family dining , cheap lunch

assigning a score within the interval [1,5], with 1 being the lowest and 5 the highest, for each of these metrics. Every sample receives evaluations from a minimum of three annotators. The evaluation outcomes indicate that both Ref2Seq (Top-k) and ACMLM outscore other models in terms of Diversity and Informativeness.

Qualitative Analysis Table 4.6 reveals that models trained on reviews and tips often produce generic phrases like 'i love this place', which may lack detailed information essential for users making decisions. Conversely, models trained on justification datasets tend to incorporate specific details, referencing different aspects. LexRank typically creates relevant yet succinct content. In contrast, sampling-based models exhibit a capacity to generate a wider variety of content.

Balancing diversity and relevance in generation can be challenging. One method to address this challenge is to incorporate more constraints during the generation phase, an example being constrained Beam Search [Anderson et al., 2017]. In our study, we enhanced our baseline

Table 4.7. Generated justifications from AP-Ref2Seq. The planned aspects are randomly selected from users' personas.

Dataset	Aspects	Generated Output
Yelp	dining	the dining room is nice
	pastry	the pastries were pretty good
	chicken	the chicken fried rice is the best
	sandwich	the pulled pork sandwich is the best thing on the menu
Amazon-Clothing	product	great product , fast shippong
	price	design is nice , good price
	leather	comfortable leather sneakers . classic
	walking	sturdy , great city walking shoes

model, Ref2Seq, by integrating aspect-planning to steer the generation process. As depicted in Table 4.7, the majority of the planned aspects are discernible in the outputs produced by AP-Req2Seq.

4.5 Conclusion

In this chapter, we discuss the challenge of generating personalized justifications. We introduce an annotated dataset and craft a pipeline to distill justifications from extensive review datasets. We present Ref2Seq, which draws from historical justifications for improved generation and controls the generation with aspects (i.e., topics). Experimentally, Ref2Seq outperforms in BLEU scores. Through human assessments, we find that reference-oriented models garnered high relevance marks, whereas sampling approaches yield increased diversity and informativeness. We conclude that aspect planning serves as an effective strategy for guiding the generation of tailored and pertinent justifications.

Chapter 4, in part, is a reprint of the material as it appears in "Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects" by Jianmo Ni, Jiacheng Li, Julian McAuley, which was published in *Empirical Methods in Natural Language Processing*, 2019. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 5

UCEPIC: Unifying Aspect Planning and Lexical Constraints for Generating Explanations in Recommendation

In this chapter, we will introduce the explanation generation. Specifically, we present a work on keyphrase-based controllable explanation generation. Personalized natural language generation is important for explaining the alignment of recommendations with user preferences. While many models employ aspect planning to control the generation process, they often fall short of producing precise and relevant information, undermining the credibility of the explanations. We propose that integrating lexical constraints can rectify these shortcomings. We introduce a novel model, UCEPIC, which crafts premium personalized explanations for recommendations by unifying aspect planning and lexical constraints through an insertion-based generation approach. From a methodological perspective, to guarantee the quality and adaptability of text generation to diverse lexical constraints, we initiate by pre-training a universal text generator using our distinctive robust insertion process. Subsequently, to derive personalized explanations within this insertion-based generation context, we formulate a strategy that seamlessly merges aspect planning and personalized references into the insertion sequence. As a result, UCEPIC unifies aspect planning and lexical constraints, facilitating the generation of explanations for recommendations in varied scenarios. In contrast to earlier models that were solely governed by aspects, UCEPIC integrates explicit details from keyphrases, significantly enhancing the variety

Table 5.1. Comparison of previous explanation generators for recommendation in group (A), general lexically constrained generators in group (B), and our UCEPIC in group (C).

Group	Methods	Personalized generation	Aspect planning	Lexical constraints	Random keyphrases
(A)	ExpansionNet [Ni and McAuley, 2018]	✓	✓	✗	✗
	Ref2Seq [Ni et al., 2019a]	✓	✓	✗	✗
	PETER [Li et al., 2021b]	✓	✓	✗	✗
(B)	NMSTG [Welleck et al., 2019]	✗	✗	✓	✗
	POINTER [Zhang et al., 2020b]	✗	✗	✓	✗
	CBART [He, 2021]	✗	✗	✓	✓
(C)	Ours	✓	✓	✓	✓

and richness of the explanations on databases like RateBeer and Yelp.

5.1 Introduction

Providing explanations or justifications for recommendations in natural language has become increasingly popular in recent years [Li et al., 2021b, Ni and McAuley, 2018, Lu et al., 2018, Li et al., 2017b, 2020b, 2023, Ni et al., 2019a]. The objective is to present product details in a tailored manner, demonstrating the alignment of the recommendation with the user's preferences. For instance, given a user-item pair, a system might produce an explanation like "nice TV with 4K display and Dolby Atmos!". To generate such compelling, personalized explanations that are coherent and relevant, recent research has adopted aspect planning. This involves incorporating various aspects [Li et al., 2021b, Ni and McAuley, 2018, Li et al., 2023, Ni et al., 2019a] into the generation process, ensuring the produced explanations encompass these aspects and are therefore more attuned to both the product and user interests.

While these methods hold promise, they often falter in incorporating precise and detailed information into the explanations. Typically, aspects, like *screen* for a TV, shape the broad sentiment or theme of the text, leading to outputs like "good screen and audio!" Yet, many detailed product attributes that users might find valuable, such as "4K display and Dolby Atmos!", elude these generators. While some explanation generators aspire to craft rich, personalized

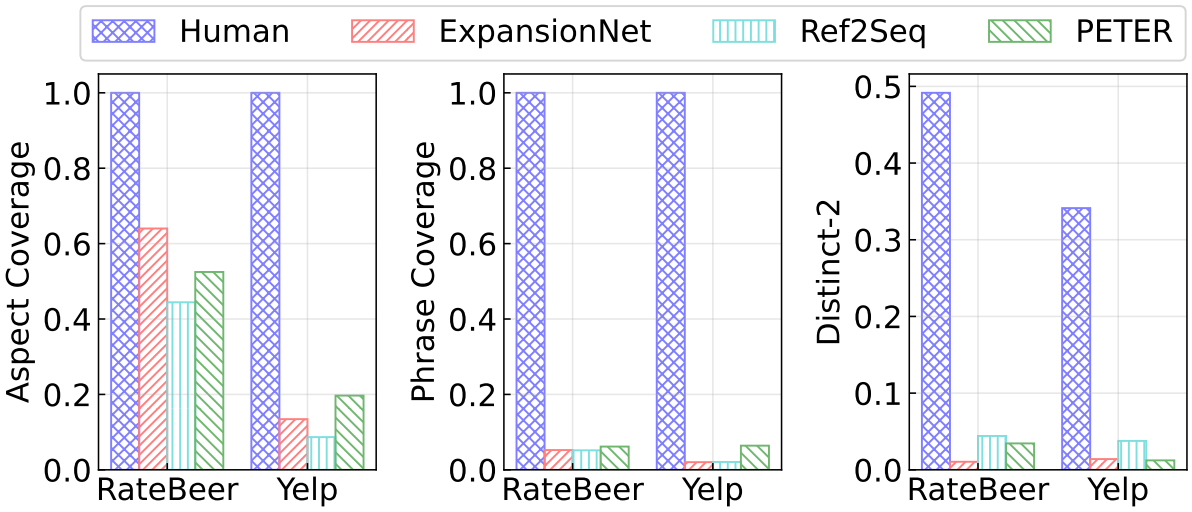


Figure 5.1. Preliminary experiments on the aspect coverage, phrase coverage, and Distinct-2 of generated explanations from previous models ExpansionNet [Ni and McAuley, 2018], Ref2Seq [Ni et al., 2019a] and PETER [Li et al., 2021b] on RateBeer and Yelp datasets.

explanations from user reviews [Li et al., 2021b, Ni and McAuley, 2018, Li et al., 2023, Ni et al., 2019a], our initial experiments indicate a notable absence of specific keyphrases from the training data in the generated content. As illustrated in Figure 5.1, outputs from prior methods often lack these unique keyphrases and exhibit diminished diversity compared to human-generated content. Relying solely on aspects leads to two major pitfalls: (1) The generation of overly generic statements, like "good screen!", which may not offer users varied or detailed explanations, and (2) The production of content with incorrect details, such as referencing a "2K screen" for a 4K TV, diminishing user trust.

For these challenges, we propose the incorporation of more strict constraints alongside aspects for recommendation explanations. Specifically, we can have a model that seamlessly integrates *lexical constraints* and *aspect planning*. By introducing lexical constraints, the model ensures the inclusion of specified keyphrases, like "Dolby Atmos", enriching the explanation's specificity and accuracy. These constraints can originate from diverse sources: *explanation systems* might select them based on item attributes; *vendors* could emphasize certain product features; or *users* might influence the generated explanations by adjusting their lexical preferences.

This approach could substantially elevate the quality, relevance, and diversity of explanations, outpacing methods that rely on aspect planning alone. Nonetheless, aspect planning still offers value, especially when no specific information is given, but a range of aspects must be addressed.

To address the challenge of seamlessly Unifying aspect-planning and lexical Constraints to enhance the Explanations in Recommendation, we introduce UCEPIC. Constructing UCEPIC poses several obstacles. Initially, most established explanation generation models, as highlighted in group (A) of Table 5.1, are not designed to accommodate lexical constraints. Predominantly built on auto-regressive generation platforms [Li et al., 2019a, Ni and McAuley, 2018, Li et al., 2020a, 2021a, Hua and Wang, 2019, Moryossef et al., 2019], these models operate on a "left-to-right" generation approach, making it challenging to ensure the inclusion of lexical constraints at any desired position. Concurrently, though insertion-based generation models, outlined in group (B) of Table 5.1, naturally incorporate lexical constraints into generated content, integrating personalization or aspects with their "encoder-decoder" architecture proves problematic. The presence of existing tokens tends to dominate the prediction of new tokens, causing the model to frequently churn out similar sentences while neglecting varied references [1] from encoders.

For the first challenge, UCEPIC adopts an insertion-based generation architecture, implementing *robust insertion pre-training* on a bi-directional transformer. This phase of robust pre-training equips UCEPIC with the foundational capability to produce text while effectively managing diverse lexical constraints. Drawing inspiration from Masked Language Modeling (MLM) [Devlin et al., 2019a], we devise an insertion procedure that incrementally embeds new tokens within sentences, ensuring UCEPIC remains adaptable to random lexical constraints.

For the second challenge, UCEPIC resorts to *personalized fine-tuning* to foster a sense of personalization and aspect awareness. To mitigate the tendency to "ignore references", our strategy involves treating references as potential insertion tokens for the generator. This encourages the model to embed new tokens that resonate with references. In terms of aspect

[1] In literature [Ni et al., 2019a], the term *references* denotes personalized user content like historical product reviews.

planning, we treat aspects as a distinct insertion phase, where tokens related to these aspects are initially crafted, laying the groundwork for subsequent generation. Finally, UCEPIC unifies lexical constraints, aspect planning, and personalized references into a singular insertion-based generation paradigm.

Generally, UCEPIC stands out as the pioneering explanation generation model that unifies aspect planning and lexical constraints. By doing so, UCEPIC markedly elevates the *relevance*, *coherence*, and *informativeness* of crafted explanations, setting it apart from current methodologies. The main contributions of our work can be encapsulated in the following key points:

- We illuminate the shortcomings of exclusively relying on aspect planning in contemporary explanation generation. Consequently, we advocate for the incorporation of lexical constraints to bolster explanation generation.
- We unveil UCEPIC, which is equipped with robust insertion pre-training and personalized fine-tuning. This design adeptly amalgamates aspect planning, lexical constraints, and references within an insertion-based generation paradigm.
- Through rigorous experiments on two distinct datasets, we validate the prowess of UCEPIC. Both objective metrics and discerning human evaluations attest to UCEPIC's proficiency in substantially enhancing the diversity, relevance, coherence, and depth of generated explanations.

5.2 Overview

Aspect planning and lexical constraints serve as important elements in explanation generation. Given a user persona denoted as R^u and an item profile represented by R^i pertaining to user u and item i, respectively, as the reference data, a model utilizing aspect planning will yield an explanation E^{ui} corresponding to a specific aspect A^{ui}. Crucially, this explanation is

Table 5.2. Notation of UCEPIC.

Notation	Description
R^u, R^i	historical review profile of user u and item i.
E^{ui}	generated explanation when item i is recommended to user u.
A^{ui}	aspects controlling explanation generation for item i and user u.
C^{ui}	lexical constraints (e.g., keywords) controlling explanation generation for item i and user u.
S^k, $\hat{S}^k$	text sequence of the k-th stage generation. S^k is training data and $\hat{S}^k$ is model prediction.
$I^{k,k-1}$, $\hat{I}^{k,k-1}$	intermediate sequence between S^{k-1} and S^k. (training data and model prediction)
$J^{k,k-1}$, $\hat{J}^{k,k-1}$	insertion number sequence between S^{k-1} and S^k. (training data and model prediction)
$\mathbf{D}$	a bi-directional transformer for encoding.
$\mathbf{H}_{MI}$	a linear projection layer for insertion numbers.
$\mathbf{H}_{TP}$	a multilayer perceptron with activation function for token prediction.

not mandated to incorporate any distinct words or phrases. On the other hand, when leveraging lexical constraints, the generation process is more stringent. Given a set of lexical constraints, which can be in the form of phrases or specific keywords, represented as $\mathrm{C}^{ui} = \{c_1, c_2, \ldots, c_m\}$, the model is obligated to produce an explanation $\mathrm{E}^{ui} = (w_1, w_2, \ldots, w_n)$ that seamlessly integrates every single given lexical constraint, c_i, implying $c_i = (w_j, \ldots, w_k)$. These lexical constraints can be sourced from various stakeholders: the users, the businesses, or even inherent attributes of items as suggested by personalized recommendation systems. UCEPIC is ingeniously designed to harmoniously blend both these constraints, offering dual operational modes: one that generates explanations based on aspect planning, and another that leans into lexical constraints. It's essential to note that our focus remains on the intricacies of the explanation generation technique, presupposing that the aspects and lexical constraints are pre-determined. A detailed summary of the notations we've employed can be found in Table 5.2.

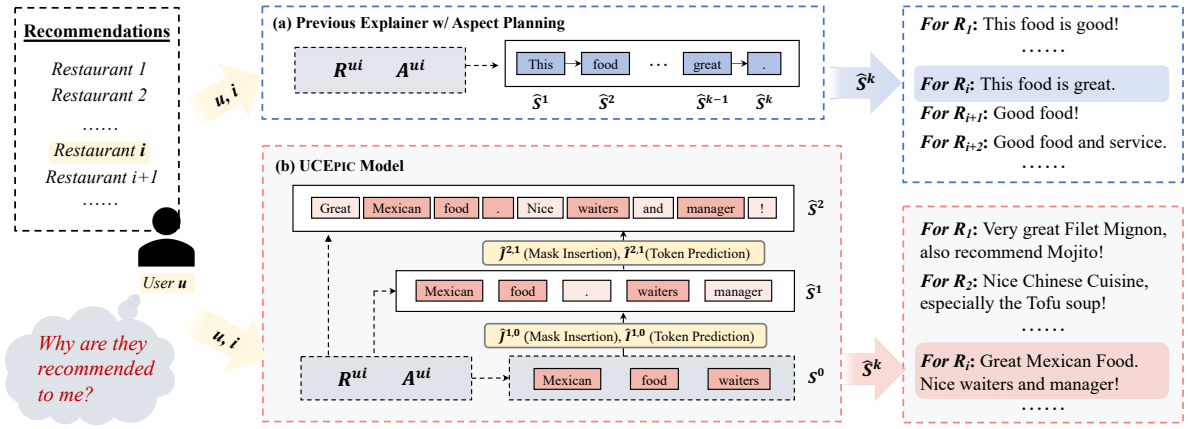


Figure 5.2. Overview of generating explanations for a given user and recommended items using (a) an aspect-planning autoregressive generation model; using (b) our UCEPIC that unifies aspect-planning and lexical constraints.

5.3 Robust Insertion

Motivation Previous methods for generating explanations [Ni et al., 2019a, Li et al., 2021b] typically employ auto-regressive generation techniques conditioned on personalized inputs such as personalized references and aspects. In Figure 5.2 (a), you can see that the auto-regressive process generates words in a "left-to-right" fashion, making it challenging to incorporate lexical constraints during the generation process. However, in insertion-based generation, as depicted in Figure 5.2 (b), where new tokens are progressively inserted based on existing words, it becomes easier to incorporate lexical constraints by treating constraints as the initial stage of insertion.

Formulation The process of generating text through insertion can be broken down into a series of stages, which we'll denote as $S = S^0, S^1, \ldots, S^{K-1}, S^K$. Here, S^0 represents the initial stage where we establish lexical constraints, while S^K represents the final text that we generate. At each step $k \in 1, \ldots, K$, S^{k-1} is a subset of S^k, and the generation process continues until no more tokens are added to S^K. During the training process, we use all sentences to create training pairs that essentially reverse the insertion-based generation process. Specifically, we form pairs of text sequences that correspond to adjacent stages, such as (S^{k-1}, S^k), to recreate the process in reverse. Each explanation E^{ui} in our training data is divided into a series of consecutive

Table 5.3. Data construction examples.

Data	Example
S^K (sentence)	<s>Good tacos. Love the crispy citrus + tropical fruits flavor. </s>
$I^{K,K-1}$	<s>[MASK] tacos. Love the [MASK] [MASK] + tropical fruits flavor. </s>
$J^{K,K-1}$	[1 0 0 0 2 0 0 0 0 0 0]
S^{K-1}	<s>tacos. Love the + tropical fruits flavor. </s>
...	...
S^0 (lexical constraints)	<s>tropical fruits flavor </s>

pairs: $(S^0, S^1), (S^1, S^2), \ldots, (S^{K-1}, S^K)$. In this construction of training data, the final stage S^K corresponds to the explanation text E^{ui}.

Data Construction. To obtain the previous stage S^{k-1} from a given sequence stage S^k, we employ two operations: masking and deletion. Specifically, we randomly mask tokens in the sequence with a probability of p, similar to the Masked Language Model (MLM) approach, resulting in an intermediate sequence denoted as $I^{k,k-1}$. Then, we delete the [MASK] tokens from this intermediate sequence $I^{k,k-1}$ to derive the stage S^{k-1}. The number of deleted [MASK] tokens after each token in $I^{k,k-1}$ is recorded as an insertion number sequence $J^{k,k-1}$. Each training instance comprises four sequences: $(S^{k-1}, I^{k,k-1}, J^{k,k-1}, S^k)$. This data construction process is illustrated with a simple example in Table Table 5.3. Since we delete $T * \mathrm{p}$ tokens in sequence S^k, where T is the length of S^k, the average number of deletions is $\log_{\frac{1}{1-\mathrm{p}}} T$. Models trained on this data can effectively leverage knowledge from BERT-like models that employ a similar pre-training process involving masked word prediction.

The process of insertion generation reverses this data construction. When making insertion predictions from $\hat{S}^{k-1}$ to $\hat{S}^k$, the model executes two operations: mask insertion and token prediction. Initially, UCEPIC inserts [MASK] tokens between any two existing tokens in $\hat{S}^{k-1}$ based on the insertion predictions $\hat{J}^{k,k-1}$. Subsequently, with the aid of a language modeling head, the model predicts the masked tokens within $\hat{I}^{k,k-1}$ and restores them to words, thereby obtaining the reconstructed $\hat{S}^k$.

Modules UCEPIC utilizes a bi-directional Transformer architecture, featuring two distinct pre-

diction heads for mask insertion and token prediction tasks. This architecture closely resembles the one employed in RoBERTa [Liu et al., 2019b]. Within this model, the bi-directional Transformer, denoted as **D**, is responsible for making predictions for both mask insertion numbers and word tokens. Two separate prediction heads are employed: $\mathbf{H}_{MI}$ for mask insertion and $\mathbf{H}_{TP}$ for token prediction. The $\mathbf{H}_{TP}$ head is designed as a multilayer perceptron (MLP) and employs the Gaussian Error Linear Unit (GeLU) activation function [Hendrycks and Gimpel, 2016]. On the other hand, the $\mathbf{H}_{MI}$ head is implemented as a linear projection layer. The final predictions for mask insertion numbers and word tokens are computed using these respective heads.

$$y_{MI} = \mathbf{H}_{MI}(\mathbf{D}(\hat{S}^{k-1})),\ \hat{J}^{k,k-1} = \text{argmax}(y_{MI}) \tag{5.1}$$

$$y_{TP} = \mathbf{H}_{TP}(\mathbf{D}(\hat{I}^{k,k-1})),\ \hat{S}^{k} = \text{argmax}(y_{TP}) \tag{5.2}$$

where $y_{MI} \in \mathbb{R}^{l_s \times d_{ins}}$ and $y_{TP} \in \mathbb{R}^{l_I \times d_{vocab}}$, l_s and l_I are the length of $\hat{S}^{k-1}$ and $\hat{I}^{k,k-1}$ respectively, d_{ins} is the maximum number of insertions and d_{vocab} is the size of vocabulary. $\hat{I}^{k,k-1}$ is obtained by inserting `[MASK]` tokens into $\hat{S}^{k-1}$ according to $\hat{J}^{k,k-1}$.

To tackle the complexity of the random insertion process, we adopt a two-step approach for pre-training the UCEPIC model before personalization. In the initial pre-training phase, we use a robust insertion method for general text generation, which doesn't involve personalization. This pre-trained model is capable of generating sentences based on randomly provided lexical constraints.

5.4 Personalized References and Aspect Planning

Motivation To incorporate personalized references and aspects into the model, one straightforward approach is to introduce another text and aspect encoder and condition the insertion generation on this encoder, akin to a sequence-to-sequence model [Sutskever et al., 2014]. However, it has been observed that using a pre-trained insertion model with an additional encoder tends to produce similar sentences with different personalized references and aspects. This phe-

nomenon occurs because the pre-trained insertion model heavily relies on the lexical constraints or existing tokens in text sequences as a strong signal to determine new inserted tokens. Even when the encoder provides personalized features, the model often overfits to the features derived from the existing tokens. In the absence of distinct lexical tokens providing different starting points, the generated sentences tend to be quite similar.

Formulation To enhance the model's ability to learn personalization, we propose treating references and aspects as unique existing tokens during the insertion process. In this approach, we introduce a training stage denoted as S_+^k that incorporates references and aspects as follows:

$$\begin{aligned} S_+^k &= [\mathrm{R}^{ui}, \mathrm{A}^{ui}, S^k] \\ &= [w_0^r, \ldots, w_{|\mathrm{R}^{ui}|}^r, w_0^a, \ldots, w_{|\mathrm{A}^{ui}|}^a, w_0, \ldots, w_{|S^k|}] \end{aligned} \tag{5.3}$$

where R^{ui} and A^{ui} represent personalized references and aspects, with w^r, w^a, and w representing the tokens or aspect identifiers in references, aspects, and the insertion stage tokens, respectively. To ensure consistency in token positions, which are crucial for the insertion-based generation process, we assign position IDs starting from 0 in the Transformer model for R^{ui}, A^{ui}, and S^k during both pre-training and fine-tuning stages. Furthermore, we create an insertion number sequence denoted as $J_+^{k,k-1}$, which consists of zero vectors corresponding to the lengths of R^{ui} and A^{ui}, followed by the insertion numbers from $J^{k,k-1}$. This is done because we do not insert any tokens into the references and aspects. Similarly, we construct an intermediate training stage denoted as $I_+^{k,k-1}$, which incorporates the personalized references R^{ui}, aspects A^{ui}, and the insertion tokens $I^{k,k-1}$. These modifications facilitate the seamless integration of references and aspects into the insertion-based generation process while preserving the overall token positioning consistency.

Modules We utilize the bi-directional Transformer **D** to encode $\hat{S}_+^k$ and $\hat{I}_+^{k,k-1}$, extracting the

insertion numbers denoted as y_{MI} and predicting the tokens as y_{TP} as outlined below:

$$[\mathrm{O}_S^{R^{ui}}, \mathrm{O}_S^{A^{ui}}, \mathrm{O}^{S^k}] = \mathbf{D}(\hat{S}_+^k) \tag{5.4}$$

$$[\mathrm{O}_I^{R^{ui}}, \mathrm{O}_I^{A^{ui}}, \mathrm{O}^{I^{k,k-1}}] = \mathbf{D}(\hat{I}_+^{k,k-1}) \tag{5.5}$$

$$y_{MI} = \mathbf{H}_{MI}(\mathrm{O}^{S^k}) \tag{5.6}$$

$$y_{TP} = \mathbf{H}_{TP}(\mathrm{O}^{I^{k,k-1}}) \tag{5.7}$$

Similar to Equation (5.1) and Equation (5.2), we can obtain $\hat{J}^{k,k-1}$ and $\hat{S}^k$ through the argmax operation. Since personalized references and aspects are treated as unique existing tokens, our model directly incorporates token-level information as generation conditions, resulting in diverse explanations.

Recall that existing text sequences serve as strong signals for token prediction. To enhance the generation of aspects and improve aspect-planning, we introduce two distinct starting stages: S_{+a}^0 for aspects and S_{+l}^0 for lexical constraints. In particular, we aim to generate aspect-related tokens from the starting stage, where there are no existing tokens, based on the provided aspects and personalized references. Therefore, the aspect starting stage is defined as $S_{+a}^0 = [\mathrm{R}^{ui}, \mathrm{A}^{ui}]$, while the lexical constraint starting stage is $S_{+l}^0 = [\mathrm{R}^{ui}, \mathrm{A}^{pad}, \mathrm{C}^{ui}]$, where A^{pad} represents a special aspect used for lexical constraints. During training, we sample S_{+a}^0 with a probability of p to ensure effective learning of aspect-related generation, a feature that is not present in the pre-training phase.

Model Training The training procedure for UCEPIC aims to learn the reverse process of data generation. We are provided with stage pairs (S_+^{k-1}, S_+^k) and training instances of the form $(S_+^{k-1}, I_+^{k,k-1}, J_+^{k,k-1}, S_+^k)$ obtained during the pre-processing step [2]. Our objective during training

[2]For fine-tuning that involves personalized references and aspects, we train the model using stage pairs (S_+^{k-1}, S_+^k) and training instances of the form $(S_+^{k-1}, I_+^{k,k-1}, J_+^{k,k-1}, S_+^k)$

is to optimize the following:

$$
\begin{aligned}
\mathscr{L} &= -\log p(S^k|S^{k-1}) \\
&= -\log \underbrace{p(S^k, J^{k,k-1}|S^{k-1})}_{\text{Unique } J \text{ assumption}} \\
&= -\log p(S^k|J^{k,k-1}, S^{k-1}) p(J^{k,k-1}|S^{k-1}) \\
&= -\log \underbrace{p(S^k|I^{k,k-1})}_{\text{Token prediction}} \underbrace{p(J^{k,k-1}|S^{k-1})}_{\text{Mask insertion}}, \\
&\text{where } I^{k,k-1} = \text{MaskInsert}(J^{k,k-1}, S^{k-1})
\end{aligned}
\tag{5.8}
$$

where MaskInsert represents the operation of inserting mask tokens. We operate under a reasonable assumption that $J_+^{k,k-1}$ is unique given the combination of (S_+^k, S_+^{k-1}). This assumption typically holds true unless there are specific cases where multiple $J_+^{k,k-1}$ could be valid (for instance, when deciding which "moving" word to mask in a phrase like "a moving moving moving van"). The intermediate sequence $I_+^{k,k-1}$, as per its definition, is equivalent to the combination of $(J_+^{k,k-1}, S_+^{k-1})$. In Equation (5.8), we simultaneously learn two aspects: (1) The likelihood of mask insertion numbers for each token, which is handled by our model with $\mathbf{H}_{MI}$. (2) The likelihood of word tokens to replace the masked tokens, which is addressed by our model with $\mathbf{H}_{TP}$.

Similar to the training of BERT [Devlin et al., 2019a], we focus our optimization efforts exclusively on the masked tokens within the token prediction task. We adopt a strategy where the tokens selected for masking have a 10% probability of remaining unchanged and a 10% probability of being randomly replaced by another token from the vocabulary. In the case of mask insertion number prediction, most of the numbers within $J_+^{k,k-1}$ are typically set to 0, as we don't insert any tokens between existing tokens in most scenarios. To strike a balance in the insertion numbers, we introduce randomness by probabilistically masking the 0 values in $J_+^{k,k-1}$ with a probability denoted as q. Note that since our mask prediction task bears similarities to masked language models, we can naturally initialize UCEPIC using pre-trained weights from

Table 5.4. Statistics of datasets

Dataset	Train	Dev	Test	#Users	#Items	#Aspects
RateBeer	16,839	1,473	912	4,385	6,183	8
Yelp	252,087	37,662	12,426	235,794	22,412	59

RoBERTa [Liu et al., 2019b] to leverage prior knowledge.

Inference During the inference phase, we initiate the process with either the given aspects A^{ui} or the lexical constraint C^{ui} to construct the starting stage, denoted as S^0_{+a} or S^0_{+l}, respectively. We then repeatedly predict $\{\hat{S}^1_+, \ldots, \hat{S}^K_+\}$ until either no additional tokens are generated or the maximum stage limit is reached. The final generated explanation, $\hat{S}^K$, is derived from $\hat{S}^K_+$ by removing R^{ui} and A^{ui}. To elaborate on the inference process from the $\hat{S}^{k-1}_+$ stage to the $\hat{S}^k_+$ stage:

1. Given $\hat{S}^{k-1}_+$, our model employs $\mathbf{H}_{MI}$ to predict the insertion number sequence $\hat{I}^{k,k-1}_+$. It's worth noting that for phrases provided in S^0_{+l}, we set the predicted insertion number as 0 to prevent any modification.

2. With $\hat{I}^{k,k-1}_+$ obtained from MaskInsert$(\hat{I}^{k,k-1}_+, \hat{S}^{k-1}_+)$, our model utilizes $\mathbf{H}_{TP}$ to predict $\hat{S}^k_+$ using a specific decoding strategy, such as greedy search or top-K sampling.

3. Given $\hat{S}^k_+$, our model assesses whether it meets the termination criteria, which can be defined by a maximum number of iterations or when no new tokens are inserted into $\hat{S}^k_+$.

This process is repeated iteratively to generate the final explanation, ensuring that the termination conditions are met.

5.5 Experiments

Datasets For the pre-training phase, we utilize the English Wikipedia dataset, which consists of approximately 11.6 million sentences, to train our model for robust insertion. To ensure a fair comparison with baseline models pre-trained on general corpora, we employ Wikipedia as the pre-training dataset. In the fine-tuning stage, we switch to Yelp and RateBeer datasets,

sourced from the respective URLs[3] and the RateBeer dataset [McAuley and Leskovec, 2013]. To maintain consistency, we filter reviews with a length exceeding 64 tokens. For each user, in line with previous work [Ni et al., 2019a], we randomly select two samples from their entire set of reviews to create the development and test sets. To extract lexical constraints and aspects for the purposes of lexical constraint and aspect planning, we employ an unsupervised aspect extraction tool [Li et al., 2022]. The number of aspects for each dataset is automatically determined by this tool. Aspects offer a high-level representation of the generated explanations, and it's important to note that typically, the number of aspects is considerably smaller than the number of lexical constraints.

Baselines We assess the effectiveness of our model through two groups of baseline models for automatic evaluation, focusing on both aspect planning and lexical constraints.

The first group comprises existing text generation models for recommendation with *aspect planning*, including:

1. **ExpansionNet** [Ni and McAuley, 2018], which generates reviews while conditioning on different aspects extracted from a given review title or summary.

2. **Ref2Seq** [Ni et al., 2019a], a Seq2Seq model that incorporates contextual information from reviews and utilizes fine-grained aspects to control explanation generation.

3. **PETER** [Li et al., 2021b], a Transformer-based model that leverages user and item IDs along with provided phrases to predict words in target explanations. This baseline represents a state-of-the-art model for explainable recommendations.

We compare these baselines under both aspect planning and lexical constraints scenarios. Specifically, we input lexical constraints (i.e., keyphrases) into the models and expect them to incorporate these keyphrases into the generated text.

The second group encompasses general natural language generation models with a focus on *lexical constraints*, including:

[3]https://www.yelp.com/dataset

1. **NMSTG** [Welleck et al., 2019], which employs a tree-based text generation approach. Given lexical constraints in the form of a prefix tree, the model generates words both to the left and right, resulting in a binary tree structure.

2. **POINTER** [Zhang et al., 2020b], an insertion-based generation method pre-trained on constructed data using dynamic programming. Our model is trained based on a pre-trained model released by the authors.

3. **CBART** [He, 2021], which utilizes the pre-trained BART Lewis et al. [2020] and instructs the decoder to insert and replace tokens based on guidance from the encoder.

The second group of baselines lacks the capability to integrate aspects or personalized information as references into their text generation process. These models are trained and generate text exclusively based on the provided lexical constraints, without taking into account aspects or personalized details.

Metrics We assess the generated sentences from two key perspectives: generation quality and diversity. To evaluate generation quality, we adopt n-gram-based metrics, including BLEU (B-1 and B-2) [Papineni et al., 2002a], METEOR (M) [Banerjee and Lavie, 2005], and ROUGE-L (R-L) [Lin, 2004]. These metrics quantify the similarity between the generated text and human-written reference sentences. In terms of generation diversity, we employ Distinct (D-1 and D-2) [Li et al., 2016b], which measures the diversity of generated text by considering the distinctiveness of n-grams. Additionally, we introduce BERT-score (BS) [Zhang et al., 2020a] as a semantic metric that assesses the quality of generated text in terms of its semantic similarity to reference sentences, rather than relying solely on n-gram matching. This comprehensive evaluation framework allows us to gauge both the quality and diversity of the generated sentences effectively.

Overall Performance In Table 5.5, we present the evaluation results for various text generation methods. In terms of aspect-planning generation, UCEPIC achieves performance on par with the state-of-the-art model PETER. Specifically, while PETER outperforms our model in metrics like

Table 5.5. Performance comparison of the explanation generation models (ExpansionNet, Ref2Seq, PETER), lexically constrained generation models (NMSTG, POINTER, CBART) and UCEPIC. All values are in percentage (%). We underline the highest scores of aspect-planning generation results and the highest scores of lexically constrained generation are **bold**.

	RateBeer							Yelp						
Models	B-1	B-2	D-1	D-2	M	R	BS	B-1	B-2	D-1	D-2	M	R	BS
Human-Oracle	–	–	8.30	49.16	–	–	–	–	–	3.8	34.1	–	–	–
Aspect-planning generation														
ExpansionNet	8.96	1.79	0.20	1.05	16.30	10.13	75.58	4.92	0.47	0.18	1.40	7.78	5.42	76.27
Ref2Seq	17.15	4.17	0.95	4.41	16.66	15.66	80.76	8.34	0.98	0.46	3.77	7.58	11.19	82.66
PETER	25.25	5.35	0.74	3.44	19.19	20.34	84.03	14.26	2.25	0.26	1.23	12.25	14.75	82.55
UCEPIC	27.42	2.89	4.49	29.23	19.54	15.48	83.53	8.03	0.72	1.89	14.75	8.10	11.58	83.53
Lexically constrained generation														
ExpansionNet	5.41	0.49	0.97	4.91	6.09	5.55	76.14	1.49	0.08	0.40	1.90	2.19	1.93	73.68
Ref2Seq	17.94	4.50	1.09	5.49	17.03	15.17	83.72	6.38	0.77	0.51	3.64	7.02	10.58	82.88
PETER	15.03	2.46	2.04	11.40	9.49	13.27	79.08	7.59	1.32	1.52	8.70	7.64	12.24	80.89
NMSTG	22.82	2.30	6.02	50.39	15.17	15.35	82.31	13.67	0.77	**4.57**	**57.02**	9.64	11.13	80.80
POINTER	6.00	0.31	**11.24**	**56.02**	7.41	11.21	81.80	1.50	0.06	5.49	29.76	3.24	5.23	80.85
CBART	2.49	0.54	8.49	34.74	8.45	13.84	83.30	2.19	0.60	5.32	26.79	9.41	15.00	84.08
UCEPIC	**27.97**	**5.09**	5.24	32.04	**19.90**	**17.05**	**84.03**	**13.77**	**3.06**	2.85	20.39	**14.45**	**16.92**	**84.55**

B-2 and ROUGE-L, UCEPIC excels in generating significantly more diverse text. This difference can be attributed to the nature of auto-regressive generation models like PETER, which tend to produce text with higher n-gram metric scores compared to insertion-based generation models like UCEPIC. Auto-regressive models generate each new token based solely on the preceding tokens, whereas UCEPIC considers tokens in both directions. Despite this intrinsic difference, UCEPIC still achieves comparable scores in B-1, Meteor, and BERT metrics when compared to PETER.

However, when operating under lexical constraints, the results of existing explanation generation models tend to be lower compared to aspect-planning generation. This indicates that current explanation generation models face challenges in incorporating specific information, such as keyphrases, into explanations. While current lexically constrained generation methods produce text with high diversity, they often insert tokens that are less relevant to users and items. As a result, the generated text may lack coherence, leading to lower n-gram metric

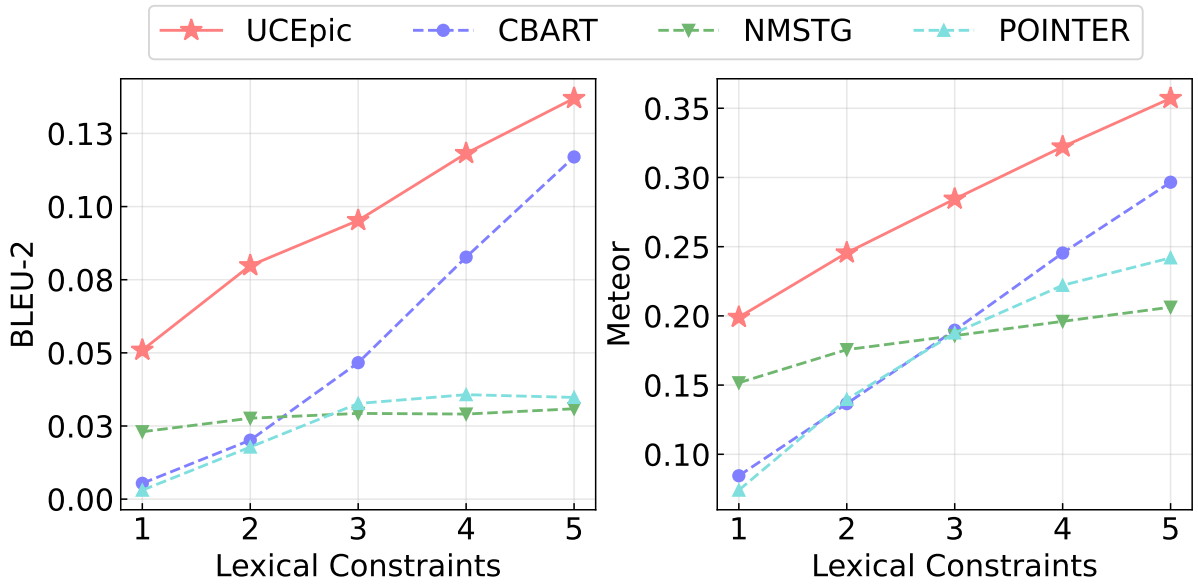


Figure 5.3. Performance (i.e., B-2 and Meteor) of lexically constrained generation models on RateBeer data with different numbers of keyphrases.

scores compared to UCEPIC. This discrepancy arises because these methods are unable to incorporate user personas and item profiles from references, which are crucial for explainable recommendation. In contrast, UCEPIC adeptly includes keyphrases in explanations and learns user-item information from references. Therefore, our model outperforms existing explanation generation models and lexically constrained generation models by a substantial margin.

Number of Lexical Constraints Figure 5.3 illustrates the performance of lexically constrained generation models across varying numbers of keyphrases. In general, UCEPIC consistently outperforms other models across different numbers of lexical constraints. Notably, NMSTG and POINTER do not exhibit significant improvements as the number of keyphrases increases, primarily because they cannot handle random keywords, and the provided phrases are often split into individual words. The performance gap between UCEPIC and CBART widens as the number of keyphrases decreases. CBART struggles to generate explanations with limited keywords, lacking sufficient information. In contrast, UCEPIC mitigates this issue by incorporating user personas and item profiles from references. These results indicate that existing lexically constrained generation models are ill-suited for explanation generation with lexical constraints.

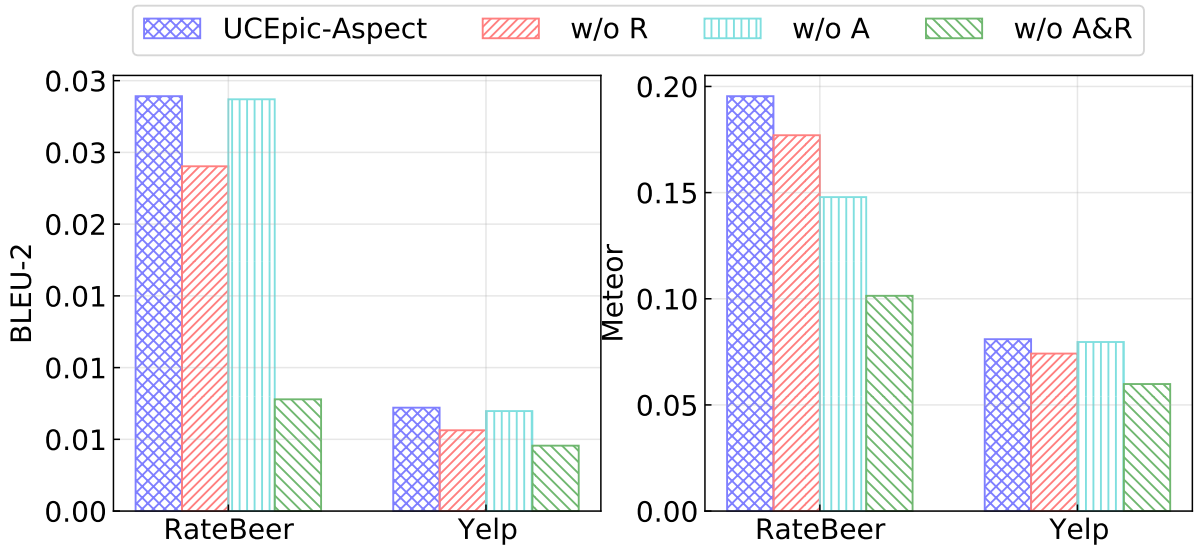


Figure 5.4. Ablation study on aspects and references.

Ablation Study To validate the effectiveness of our comprehensive approach and underscore the importance of aspects and references in explanation generation, we conducted an ablation study on two datasets, and the results are presented in Figure 5.4. We trained our model and generated explanations under three conditions: without aspects (w/o A), without references (w/o R), and without both aspects and references (w/o A&R). The results reveal the following insights:

1. Excluding aspects from the model (w/o A) leads to a decrease in BLEU-2 and Meteor scores. This decline suggests that aspects play a crucial role in guiding the semantics of explanations, contributing to improved generation quality.

2. When references are omitted (w/o R), the model tends to generate similar sentences, often containing high-frequency words from the training data. This results in a notable drop in performance, emphasizing the significance of references in generating diverse and contextually relevant explanations.

3. The most significant performance drop occurs when both references and aspects are absent from the model (w/o A&R). This underscores the effectiveness of our unified approach for incorporating references and aspects, as it provides vital user-item information for

Table 5.6. UCEPIC with different constraints on Yelp dataset. L denotes lexical constraints.

Constraints	B-1	B-2	D-1	D-2	M	R	BS
Aspect	8.03	0.72	1.89	14.75	8.09	11.58	83.53
L-Extract	13.77	3.06	2.85	20.39	14.45	16.92	84.55
L-Frequent	10.05	0.87	2.02	15.88	9.14	12.23	83.73
L-Random	9.81	0.79	3.00	21.04	8.73	11.61	83.50
Aspect & L	13.12	3.01	2.89	20.34	14.41	16.94	84.56

explanation generation.

In summary, our ablation study highlights the important role of aspects and references in the explanation generation process and underscores the effectiveness of our comprehensive approach that integrates both aspects and references into the model.

Kind of Constraints We conducted a performance analysis of UCEPIC using various types of constraints on the Yelp dataset, and the results are presented in Table 5.6. The settings for Aspect and L-Extract remain consistent with UCEPIC under aspect-planning and lexical constraints, as described in Table 5.5. In addition to these constraints, we investigated three other types:

1. L-Frequent. This constraint involves using the most frequent noun phrase of an item as the lexical constraint.

2. L-Random. For this constraint, we randomly selected the lexical constraint from all noun phrases associated with an item.

3. Aspect & L. This method combines both aspect-planning and lexical constraints, as previously described in Table 5.5, allowing the simultaneous use of both types of constraints.

The analysis of the results reveals the following key observations:

1. L-Extract and Aspect & L exhibit similar performance, implying that the presence of lexical constraints exerts strong constraints on the generation process, limiting the degree of controllability afforded by aspect planning.

Table 5.7. Generated explanations from Yelp dataset. Lexical constraints (phrases) are highlighted in explanations.

Phrases	**pepper chicken**	**north shore, meat**
Human	Food was great. The pepper chicken is the best. This place is neat and clean. The staff are sweet. I recomend them to anyone!!	Great Italian food on the north shore! Menu changes daily based on the ingredients they can get locally. Everything is organic and made "clean". There is no freezer on the property, so you know the meat was caught or prepared that day. The chef is also from Italy! I highly recommend!
Ref2Seq	best restaurant in town ! ! !	what a good place to eat in the middle of the area. the food was good and the service was good.
PETER	This place is great! I love the food and the service is always great. I love the chicken and the chicken fried rice. I love this place.	The food was good, but the service was terrible. The kitchen was not very busy and the kitchen was not busy. The kitchen was very busy and the kitchen was not busy.
POINTER	pepper sauce chicken !	one of the best restaurants in the north as far as i love the south shore. great meat ! !
CBART	Great spicy pepper buffalo wings and chicken wings.	Best pizza on the north shore ever! Meatloaf is to die for, especially with meat lovers.
UCEPIC	Great Chinese restaurant, really great food! The customer service are amazing! Everything is delicious and delicious! I think this local red hot pepper chicken is the best.	I had the best Italian north shore food. The service is great, meat that is fresh and delicious. Highly recommend!

2. Generation under lexical constraints consistently outperforms aspect-planning generation.

3. Different lexical constraint selection methods (i.e., L-Extract, L-Frequent, L-Random) yield significant variations in generation performance. This underscores the potential for further exploration and experimentation in the domain of lexical constraint selection in future research.

Human Evaluation We conducted a human evaluation of the generated explanations. Here is how the evaluation was carried out:

1. We randomly selected 500 ground-truth explanations from the Yelp dataset.

2. We collected corresponding generated explanations from PETER-aspect, POINTER, CBART, and UCEPIC for each ground-truth explanation.

3. Annotators were asked to choose the best explanation among those generated by PETER, POINTER, CBART, and UCEPIC based on different aspects, including relevance,

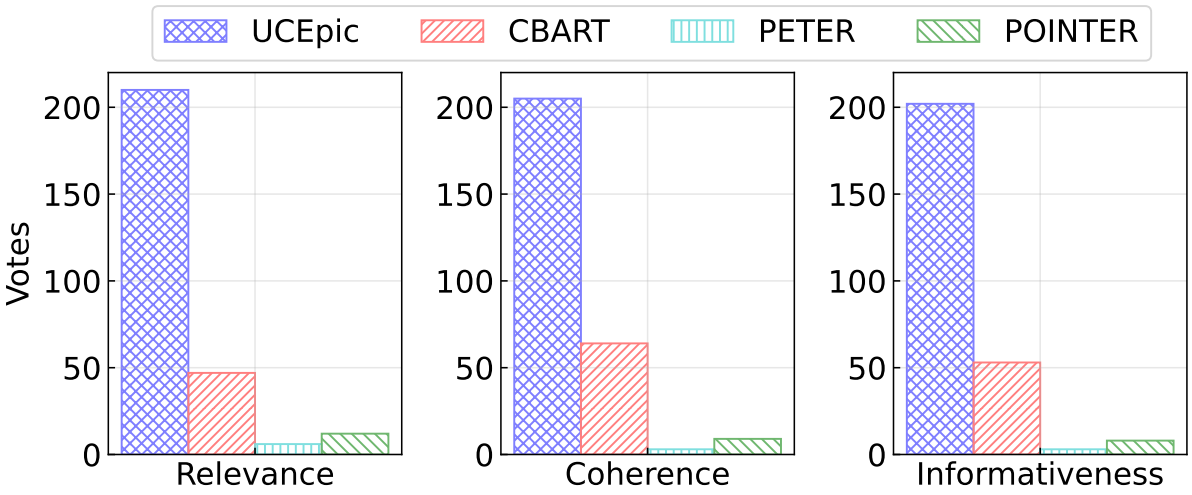


Figure 5.5. Human evaluation on explanation quality.

coherence, and informativeness.

We defined these aspects as follows:

- Relevance: This evaluates whether the details in the generated explanation are consistent and pertinent to the ground-truth explanations.
- Coherence: This assesses whether the sentences in the generated explanation are logical and fluently presented.
- Informativeness: This gauges whether the generated explanation contains specific information rather than vague descriptions.

The results of the evaluation are depicted in Figure 5.5, showing that UCEPIC outperforms other methods across all aspects, especially in terms of relevance and informativeness. Notably, lexically constrained generation methods (UCEPIC and CBART) significantly enhance the quality of explanations, as they enable the inclusion of specific product information in the explanations through lexical constraints. Conversely, POINTER does not benefit from lexical constraints and struggles with random keyphrases, resulting in explanations that do not see improvements from the inclusion of lexical constraints.

Case Study We compared the generated explanations from various existing explanation generation models, including Ref2Seq and PETER, as well as lexically constrained generation models like POINTER and CBART, with those from UCEPIC. The results are summarized in Table 5.7, and the following observations can be made:

- Ref2Seq and PETER typically produce generic sentences that lack specificity and informativeness. These models struggle to incorporate specific item information through traditional auto-regressive generation.
- POINTER and CBART manage to include the given phrases (e.g., "pepper chicken") in their generated explanations. However, they fail to learn information from references, resulting in some inaccuracies (e.g., "pepper sauce chicken," "chicken wings") that can potentially mislead users.
- In contrast, UCEPIC consistently generates coherent and informative explanations that include specific item attributes and maintain high relevance to the recommended items.

Overall, UCEPIC stands out for its ability to provide relevant, coherent, and informative explanations that capture specific item details effectively.

5.6 Conclusion

We introduce the concept of incorporating lexical constraints into explanation generation to enhance the informativeness and diversity of generated reviews by including specific details. To address this, we present UCEPIC, explanation generation model that combines both aspect planning and lexical constraints within an insertion-based generation framework. We conducted extensive experiments using the RateBeer and Yelp datasets, and our results demonstrate that UCEPIC outperforms existing explanation generation models and lexically constrained generation models. Human evaluation and a case study further confirm that UCEPIC generates

coherent, informative explanations that maintain a high level of relevance to the recommended item.

Chapter 5, in part, is a reprint of the material as it appears in "UCEpic: Unifying Aspect Planning and Lexical Constraints for Generating Explanations in Recommendation" by Jiacheng Li, Zhankui He, Jingbo Shang, Julian McAuley, which was published in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 6

Personalized Showcases: Generating Multi-Modal Explanations for Recommendations

In this chapter, we introduce how to produce multi-modal explanations (including text and images) as showcases for recommendations. Current recommendation explanation models primarily generate text-based explanations, which often lack diversity in their content. We introduce a novel task called "personalized showcases" to enhance explanations by combining textual and visual information to justify recommendations. In this task, we begin by curating a personalized set of images that align with a user's interests in a recommended item. Subsequently, we generate natural language explanations that correspond to the selected images. To facilitate this, we have compiled a large-scale dataset from Google Maps and created a high-quality subset to generate multi-modal explanations. Our approach involves a personalized multi-modal framework that leverages contrastive learning to generate diverse and visually-aligned explanations. Our experiments demonstrate that our framework benefits from incorporating different modalities as inputs and is capable of producing explanations that are more diverse and expressive than previous methods, as evidenced by various evaluation metrics.

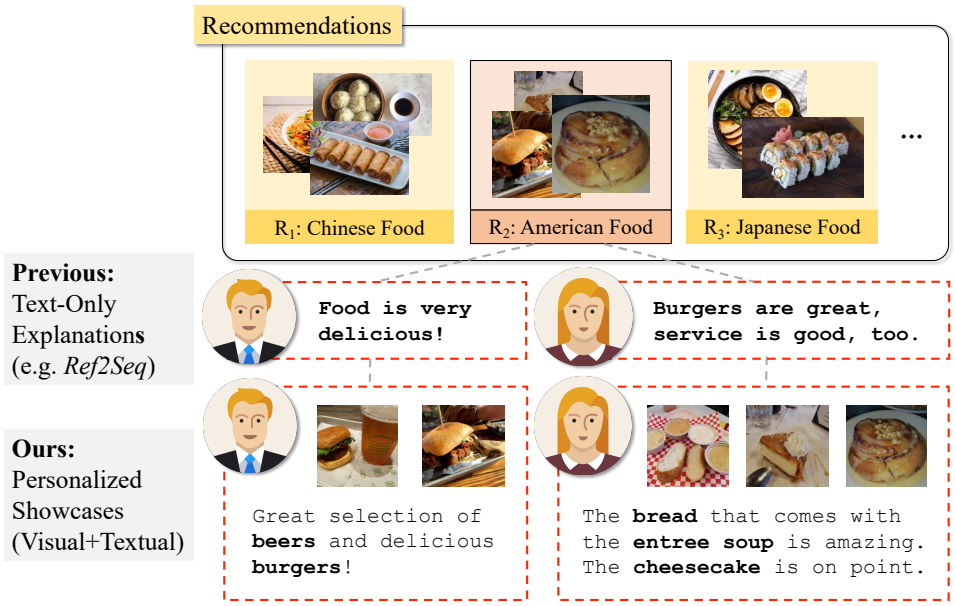


Figure 6.1. Illustration of previous text-only explanation and our personalized showcases for recommendations. Given a recommended item or business: (1) Text-only Explanation models only use historical textual reviews from user and item sides to generate textual explanations. (2) We propose a personalized showcases task to enrich the personalized explanations with multi-modal (visual and textual) information, which can largely improve the informativeness and diversity of generated explanations.

6.1 Introduction

The potential of personalized explanation generation models to enhance the clarity and trustworthiness of recommendations has been acknowledged in past research [Dong et al., 2017, Chen et al., 2019, Baheti et al., 2018, Zang and Wan, 2017], which typically focused on producing explanations from users' past reviews, tips [Li et al., 2017a], or justifications [Ni et al., 2019b]. Despite this, such methods often fall short in terms of explanation variety, mainly due to the prevalence of generic sentences and a lack of specific, supportive information, such as images, in the generation process. To address these limitations, we introduce a novel task called "personalized showcases" as detailed in Figure 6.1. This task involves explaining recommendations through both images and text, targeting the integration of visual elements that align with a user's preferences and crafting textual explanations to match.

Creating a dataset is the initial step for this task. Traditional review datasets like those from Amazon [Ni et al., 2019b] and Yelp [1] do not fit well with the requirements of this task. Therefore, we assemble a new, extensive multi-modal dataset, referred to as GEST, from the Google Local [2] Restaurants, which includes both review texts and corresponding images. To enhance the dataset for personalized showcases, we meticulously annotated a small segment to identify image-sentence pairs that correlate closely. Utilizing these annotations, we trained a classifier with the help of CLIP [Radford et al., 2021a] to select visually coherent explanations throughout the dataset. The curated images and associated textual explanations provided by users are then employed as a learning foundation for the personalized showcases.

In the realm of this innovative task, we have devised a novel multi-modal explanation framework. This framework initiates the process by selecting a subset of images from a business's historical photo collection that aligns with the user's discernible interests. It then leverages these selected images alongside the user's profile data, such as previous reviews, as input to a multi-modal decoder designed to generate textual explanations. Notwithstanding, the creation of expressive, varied, and captivating textual content that resonates with the user's interests presents a significant technical challenge. The input complexity, encompassing multiple images and historical reviews, demands advanced capabilities for information extraction and the integration of multiple modalities. Moreover, the need for coherent alignment between the visual content and the accompanying textual explanations elevates the complexity of this task. Additionally, conventional encoder-decoder architectures, often employing cross-entropy loss coupled with teacher forcing, are prone to produce monotonous and repetitive sentences, a phenomenon frequently observed in the training data (e.g., "food is great") [Holtzman et al., 2019a].

To address these challenges, we introduce the **P**ersonalized **C**ross-Modal **C**ontrastive **L**earning (PC^2L) framework, which utilizes contrastive learning to differentiate between input modalities and the resultant sequences. While contrastive learning has garnered interest for

[1] https://www.yelp.com/dataset
[2] https://www.google.com/maps

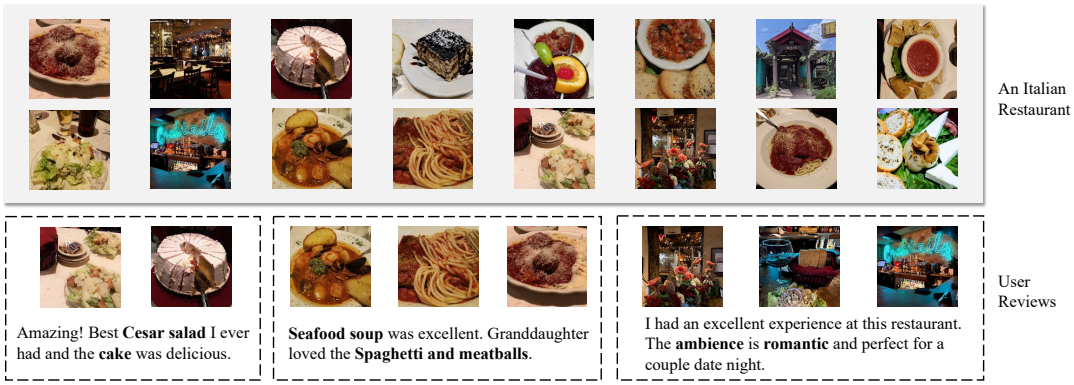


Figure 6.2. Example of business and user reviews in GEST. For a business (e.g., an Italian restaurant), GEST contains historical reviews and images from different users.

its efficacy in self-supervised representation learning [Oord et al., 2018, Chen et al., 2020a], the conventional approach of employing negative samples within a mini-batch has shown to be less than ideal [Lee et al., 2020], as the randomly selected embeddings are often easily distinguishable within the latent space. To optimize this, we develop a cross-modal contrastive loss that promotes congruence between images and their textual explanations by incorporating 'hard' negative samples, crafted through the strategic replacement of entities in the output. Taking inspiration from patterns indicating that users with analogous review histories tend to exhibit parallel interests, we further refine this model with a personalized contrastive loss that adjusts the weighting of negative samples based on the degree of similarity in their review history. Comprehensive evaluations, both automatic and via human assessment, indicate that our model outperforms existing benchmarks, delivering more engaging, varied, and visually coherent explanations.

6.2 Task Definition

In the context of the "personalized showcases" task, our objective is to furnish tailored textual and visual explanations to elucidate recommendations made to users. To formally define this task, we consider a user, denoted as $u \in U$, and a business (item), denoted as $b \in B$, where U and B represent the sets of users and businesses, respectively. In this task, we aim to provide a

set of textual explanations, denoted as $S = \{s_1, s_2, \ldots, s_m\}$, and visual explanations, denoted as $I = \{i_1, i_2, \ldots, i_n\}$. Here, s and i represent sentences and images in the explanations, respectively. The textual and visual explanations in S and I are matched with each other and are personalized to clarify why business b is being recommended to user u.

To facilitate a deeper understanding of the relationship between textual and visual explanations and to establish baselines for future research, we decompose the task into two distinct steps, as illustrated in Figure 6.3:

1. **Selecting Relevant Images**: The first step involves selecting a set of images, denoted as I, from a candidate set of images associated with business b, represented as $I_b = \{i_1^b, i_2^b, \ldots, i_{|I_b|}^b\}$. These selected images should be aligned with the user's interests, determined by their profile, which includes historical reviews $X_u = \{x_1^u, x_2^u, \ldots, x_K^u\}$ and images $I_u = \{i_1^u, i_2^u, \ldots, i_n^u\}$.

2. **Generating Textual Explanations**: Once the relevant images I have been selected, we utilize the user's historical reviews X_u and the chosen images to generate visually-informed textual explanations, denoted as S.

Given a user u, a business b, and a candidate image set I_b, we first determine a set of images I that align with the user's interests, based on their historical reviews X_u and images I_u. Subsequently, we employ the user's historical reviews X_u and the selected images I to generate visually-informed textual explanations S. This decomposition of the task allows for a more detailed analysis of the interplay between textual and visual elements in the explanation generation process.

Our approach considers several key aspects in the context of the "personalized showcases" task:

1. **Accuracy**: We prioritize the accurate prediction of target images, which are images associated with the ground-truth review, from a pool of business image candidates. Additionally, the generated textual explanations should exhibit relevance to the specific business being

Table 6.1. Data statistics for GEST. Avg. R. Len. denotes average review length and #Bus. denotes the number of Businesses. -raw denotes raw GEST. -s1 denotes GEST data for the first step, and -s2 denotes GEST data for the second step of our proposed personalized showcases framework.

Dataset	#Image	#Review	#User	#Bus.	Avg. R. Len.
GEST-raw	4,435,565	1,771,160	1,010,511	65,113	36.26
GEST-s1	1,722,296	370,563	119,086	48,330	45.48
GEST-s2	203,433	108,888	36,996	30,831	24.32

recommended. This ensures that the explanations effectively connect the chosen images with the business in question.

2. **Diversity**: We aim to ensure diversity in both the selection of images and the textual explanations. In terms of image selection, our goal is to choose a diverse set of images that provide a comprehensive view of the business. For instance, in the case of a restaurant, this may involve including images of various dishes and the restaurant's ambiance. In the realm of textual explanations, diversity is sought to make the generated text more expressive and informative.

3. **Alignment**: In contrast to previous explanation or review generation tasks that rely solely on historical reviews or aspects as inputs, our approach operates in a visually-aware setting. As such, the generated explanations in this new task are expected to accurately describe the content within the selected images. This entails covering essential elements such as naming the dishes and describing the environment, effectively aligning the textual explanations with the visual content.

These considerations collectively contribute to the effectiveness and informativeness of our approach in generating personalized showcases, enriching the user experience in understanding recommendations through both text and images.

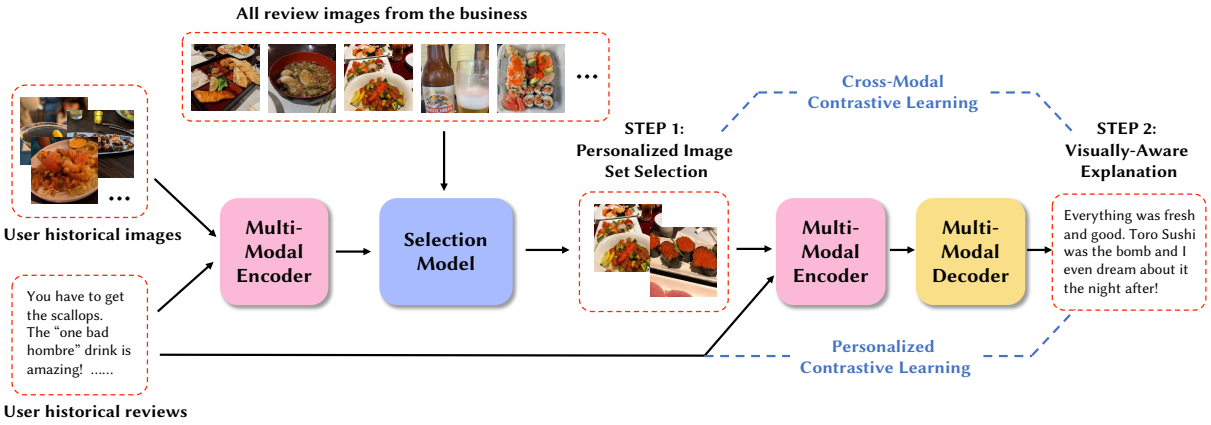


Figure 6.3. Illustration of our *personalized showcases* framework for the given business. We take user historical images and textual reviews as inputs. First, we select an image set that is most relevant to a user's interest. Then we generate natural language explanations accordingly with a multi-modal decoder. A cross-modal contrastive loss and a personalized contrastive loss are applied between each input modality and the explanations. Last, the selected images and generated textual explanations will be organized as multi-modal explanations to users.

6.3 Dataset

We have aggregated a collection of reviews accompanied by images from *Google Local*. The raw data set, referred to as GEST-raw and detailed in Table 6.1, comprises 1,771,160 reviews across 1,010,511 users and 65,113 businesses. Each review is associated with at least one image, cumulating in a total of 4,435,565 image URLs within the dataset.

This dataset has been partitioned into two distinct subsets for analysis: (1) GEST-s1, which facilitates the personalized selection of image sets, and (2) GEST-s2, which is used for the generation of visually-aware textual explanations. The statistics for these processed subsets are documented in table 6.1.

To distinguish our GEST from pre-existing review datasets and to underscore the value of the *personalized showcases*, we introduce a CLIP-based dissimilarity measure, computed at three granularity levels to evaluate the diversity of user-generated images for each business. This methodology is used to contrast the visual diversity within our GEST against two notable review datasets: Amazon Reviews [McAuley et al., 2015, Ni et al., 2019b] and Yelp.

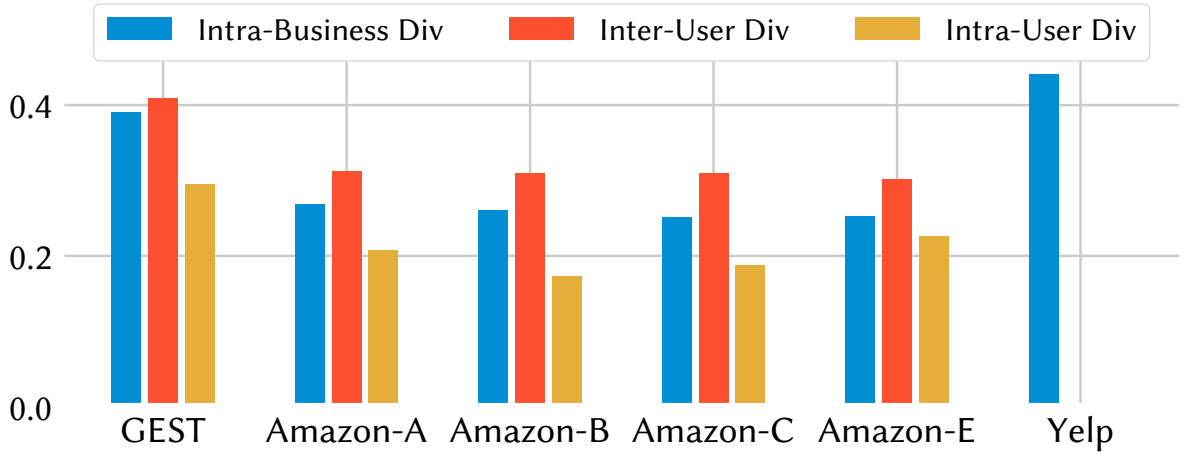


Figure 6.4. Visual Diversity Comparison. A, B, C, E in Amazon denote different categories of amazon review datasets, which are uniformly sampled from *All*, *Beauty*, *Clothing* and *Electronics*, respectively. Intra-/Inter- User Diversity for Yelp dataset is unavailable since Yelp images lack user information.

Consistent with the approach of Radford et al. [2021a], Zhu et al. [2021], we employ the cosine similarity metric, as derived from a pre-trained CLIP model, to define the dissimilarity between two images i_m and i_n as $dis(i_m, i_n) = 1 - sim(i_m, i_n)$. We then operationalize visual diversity across three dimensions: *Intra-Business Div*, *Inter-User Div*, and *Intra-User Div*. Higher values indicate a greater degree of visual diversity.

Our examination of visual diversities encompasses not only our GEST but also extends to Amazon Reviews, considering both the aggregate of all categories (*All* (A)) and specific subcategories such as *Beauty* (B), *Clothing* (C), and *Electronics* (E). For Amazon, we treat each item page akin to a 'business' and analyze 5,000 such entities that feature more than one user-uploaded image. It is pertinent to note that due to the absence of user metadata for images in the Yelp dataset, user-level diversity metrics are not computable. The insights from our comparative analysis are encapsulated in Figure 6.4.

- **Dataset Diversity Metrics:** Visual inspection of Figure 6.4 reveals that within the GEST and Amazon datasets, *Inter-User Div* registers as the highest, while *Intra-User Div* scores the lowest. This suggests that even when considering the same business or item, individual

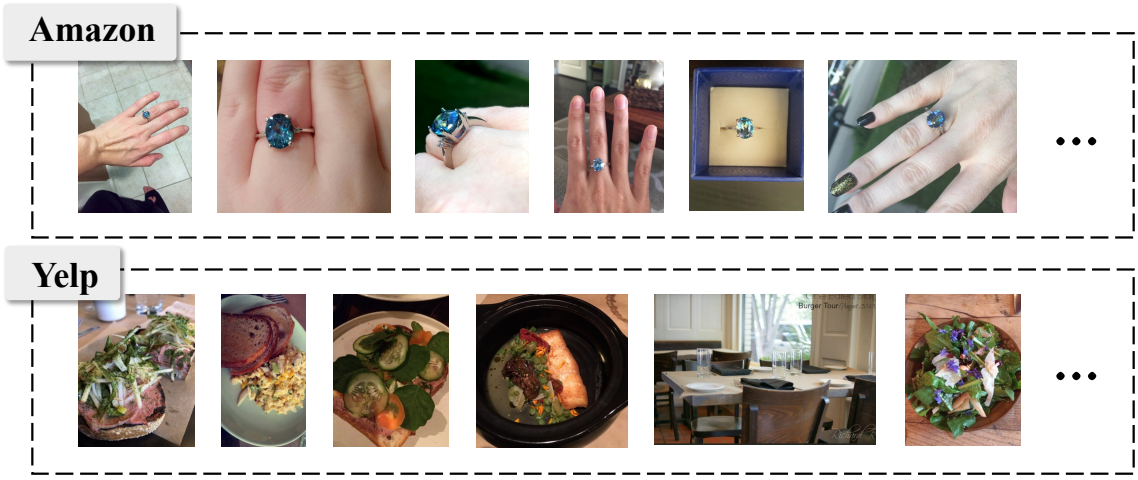


Figure 6.5. Example of user-generated images from Amazon from an item page and for Yelp from a business. Amazon images mainly focus on a single item and Yelp images for a business are diverse (yet the current public Yelp dataset has no user-image interactions).

users tend to highlight distinct visual content.

- **Comparative Analysis of GEST and Amazon:** As illustrated in Figure 6.4, all three visual diversity indices for Amazon are significantly lower than those for GEST. This discrepancy may be attributed to divergent user interaction patterns on the two platforms. For instance, as depicted in Figure 6.5, images uploaded by users on Amazon predominantly concentrate on the purchased product, with variations mainly in the details they wish to showcase. Typically, these images feature the product as the sole subject, thereby constraining the scope of visual diversity. Conversely, GEST, with samples in Figure 6.2, permits users to post reviews on restaurants, offering a canvas for a broader spectrum of content ranging from different items to various angles and perspectives. Hence, leveraging GEST is likely to yield more informative *personalized showcases* tailored to distinct user profiles, as opposed to using Amazon's dataset.

- **Comparison Between GEST and Yelp:** The imagery on Yelp is noted for its high quality, as exemplified in Figure 6.5, and the *Intra-Business Diversity* metric stands at 0.44, which surpasses that of GEST at 0.39. The visual content in Yelp exhibits a resemblance to that

found in GEST. Nonetheless, Yelp's dataset does not align with our task requirements due to the absence of linked user information.

The review corpus frequently comprises text that is tangentially related to accompanying images, rendering it suboptimal for direct usage as explanations. Consequently, we curated an explanation-specific dataset derived from GEST-raw by refining sentences within reviews to ensure they correlate with the corresponding images, thus constituting viable explanations.

To execute this, three evaluators were recruited to assess a randomly selected subset of 1,000 reviews from the aggregate dataset to ascertain their suitability as "good" explanations. This assessment was conducted through an iterative process, incorporating periodic feedback and discussions to synchronize the evaluators' assessment criteria. The subset comprised 9,930 image-sentence pairs, which were subsequently partitioned into training, validation, and test sets in an 8:1:1 ratio.

Subsequent to the annotation process, a binary classifier model Φ was trained using the labeled image-sentence pairs. This involved encoding the sentence and image to obtain their embeddings through the CLIP framework. The resultant embeddings were concatenated and input into a dense neural layer. A hyper-parameter optimization procedure determined that a classification threshold of 0.5 on the predicted probability yielded optimal results, achieving an Area Under the Receiver Operating Characteristic Curve (AUC) of 0.97 and an F-1 score of 0.71 in the test dataset.

This trained model was then employed to sift through the entirety of the review content to distill explanations, with GEST-s2 encapsulating the statistics of the filtered dataset as detailed in Table 6.1.

6.4 Methodology

In this section, we present the architecture of our framework dedicated to generating personalized showcases. The schematic (Figure 6.3) delineates the initial phase of personalized

image set curation followed by the module responsible for the generation of visually-informed textual explanations. Subsequently, we elaborate on our bespoke approach to personalized cross-modal contrastive learning.

6.4.1 Personalized Image Set Selection

The first step is to select an image set as a visual explanation that is relevant to a user's interests and is diverse. We formulate this selection step as a diverse recommendation with multi-modal inputs.

Multi-Modal Encoding Framework The encoding of user-generated textual and visual content is typically achieved using various pre-trained deep learning architectures such as ResNet He et al. [2016], ViT Dosovitskiy et al. [2021], and BERT Devlin et al. [2019a]. For our purposes, we employ CLIP [Radford et al., 2021b], which is at the forefront of pre-trained cross-modal retrieval models, to encode both the visual and textual profiles of users. CLIP processes raw images into image feature representations and textual as well as visual user profiles into corresponding user profile features.

Image Subset Selection Mechanism The task of selecting an optimal subset of images is conducted via a Determinantal Point Process (DPP) methodology [Kulesza and Taskar, 2012], which has found recent applications in various tasks requiring diverse recommendations [Wilhelm et al., 2018, Bai et al., 2019]. In contrast to other algorithms that cater to recommendations of individual items, the DPP method is aptly geared for the selection of multiple images. For a given user u and business b, we employ the approach delineated in [Wilhelm et al., 2018] to forecast the set of images $\hat{I}_{u,b}$ as follows:

$$\hat{I}_{u,b} = \mathrm{DPP}(I_b, u), \tag{6.1}$$

where I_b denotes the collection of images associated with business b. In our approach, we determine the relevance between user and image using features derived from the user's profile

and image characteristics as encoded by CLIP. Further specifics regarding the DPP model are elaborated in [Wilhelm et al., 2018].

6.4.2 Visually-Aware Explanation Generation

Subsequent to the selection of a pertinent image set, the objective shifts to crafting tailored explanations that combine the chosen image set with the user's historical review data, utilizing the curated explanation dataset GEST-s2. To this end, we architect a multi-modal encoder-decoder framework leveraging the GPT-2 model [Radford et al., 2019] as the foundational structure.

Multi-Modal Encoder For a collection of historical reviews $X = \{x_1, x_2, \ldots, x_K\}$ associated with a user u (subscript u omitted for brevity), we employ the textual encoding component of CLIP to derive the review feature set $R = \{r_1, r_2, \ldots, r_K\}$. Analogously, we process the input image set $I = \{i_1, i_2, \ldots, i_n\}$ through a pretrained ResNet [He et al., 2016] to obtain visual feature vectors $V = \{v_1, v_2, \ldots, v_n\}$. These features are subsequently mapped into a common latent space as follows:

$$Z_i^V = W^V v_i, Z_i^R = W^R r_i, \tag{6.2}$$

where W^V and W^R denote the trainable projection matrices. We then apply a multi-modal attention (MMA) mechanism, constructed with a series of stacked self-attention layers [Vaswani et al., 2017], to process the projected features:

$$[H^V; H^R] = \text{MMA}([Z^V; Z^R]), \tag{6.3}$$

where H_i^V and H_i^R embody the integrated features from both textual and visual modalities, and $[;]$ represents the concatenation operation. This architecture is designed to accommodate inputs of varying lengths from each modality and to facilitate modality cross-communication via co-attention mechanisms.

Multi-Modal Decoder. Drawing on the transformative impact of advanced pre-trained language models, we incorporate GPT-2 as the basis for our explanation generation decoder. To effectively

harness the extensive linguistic proficiency encoded within GPT-2, we integrate an encoder-decoder attention module, following the structural paradigm detailed in [Chen et al., 2021].

Within this multi-modal GPT-2 architecture, the decoder synthesizes the target explanatory sequence $Y = \{y_1, y_2, ..., y_L\}$ through an iterative decoding process at each timestep t, which is mathematically depicted as:

$$\hat{y}_t = \text{Decoder}([H^V; H^R], y_1, \ldots, y_{t-1}). \tag{6.4}$$

To optimize the generation of explanations, we apply a cross-entropy (CE) loss function aiming to enhance the conditional log likelihood $\log p_\theta(Y|X,I)$ across a set of N training samples denoted by $(X^{(i)}, I^{(i)}, Y^{(i)})_{i=1}^N$:

$$\mathscr{L}_{CE} = -\sum_{i=1}^{N} \log p_\theta(Y^{(i)}|X^{(i)}, I^{(i)}). \tag{6.5}$$

For the training phase, authentic images associated with the user are utilized, whereas for inference, the dataset employs images delineated by our image-selection model.

6.4.3 Personalized Cross-Modal Contrastive Learning

Contrary to image captioning tasks that demand succinct image descriptions, our endeavor engages multiple images as stimuli to evoke personal narratives and judgments. To foster the generation of explanations that are eloquent, varied, and visually coherent, we introduce a **Personalized Cross-Modal Contrastive Learning** (PC^2L) framework. This framework begins by mapping the latent representations of images, historical reviews, and the target explanations into a shared latent space:

$$\tilde{H}^V = \phi_V(H^V),\ \tilde{H}^R = \phi_R(H^R),\ \tilde{H}^Y = \phi_Y(H^Y) \tag{6.6}$$

Here, ϕ_V, ϕ_R, and ϕ_Y represent transformation functions composed of dual-layer fully connected neural networks with ReLU activation functions [Nair and Hinton, 2010], coupled with an average pooling operation across the hidden states H^V, H^R, and H^Y derived from the terminal self-attention layers.

For standard contrastive learning employing the InfoNCE loss [Oord et al., 2018, Chen et al., 2020a], the objective is to augment the affinity between the source modality and the target sequence while concurrently diminishing the affinity with dissimilar (negative) pairings, shown as follows:

$$\mathscr{L}_{CL} = -\sum_{i=1}^{N} \log \frac{\exp(s_{i,i}^{X,Y})}{\exp(s_{i,i}^{X,Y}) + \sum_{j \in K} \exp(s_{i,j}^{X,Y})}, \tag{6.7}$$

where $s_{i,j}^{X,Y} = sim(\tilde{H}_{(i)}^{X}, \tilde{H}_{(j)}^{Y})/\tau$ signifies the cosine similarity between two vectors normalized by a temperature parameter τ, with (i) and (j) indexing samples within a mini-batch, and K representing the set of negative samples for the i-th sample.

One complexity in this domain is the necessity for the model to articulate various elements or aspects within an ensemble of images. To anchor the visual elements in the multiple-image features with the generated textual output robustly, we propose an innovative cross-modal contrastive loss. Specifically, for a given target explanation $Y = \{y_1, y_2, ..., y_L\}$, we create a challenging negative sample $Y^{ent} = \{y'_{ent1}, y_2, ...y'_{ent2}, ...y_L\}$ by substituting entities in the explanation with different entities encountered in the corpus—transforming, for instance, "I like the sushi" into "I like the burger". This practice compels the model to differentiate between the correct sequence and those with incongruent entities relative to the images during training. The adversative representation of Y^{ent} is incorporated as an additional negative sample *ent* in the formulation of the cross-modal contrastive loss:

$$\mathscr{L}_{CCL} = -\sum_{i=1}^{N} \log \frac{\exp(s_{i,i}^{V,Y})}{\exp(s_{i,i}^{V,Y}) + \sum_{j \in K \cup ent} \exp(s_{i,j}^{V,Y})}, \tag{6.8}$$

Concurrently, to augment the individualized nature of the explanation synthesis, we

modulate the weighting of negative pairs contingent upon user-specific traits. The underpinning premise is that users with more pronounced individual traits are prone to proffer divergent explanations. Propelled by this insight, we introduce a personalized variant of the contrastive loss function:

$$\mathscr{L}_{PCL} = -\sum_{i=1}^{N} \log \frac{\exp(s_{i,i}^{R,Y})}{\exp(s_{i,i}^{R,Y}) + f(i,j) \sum\limits_{j \in K} \exp(s_{i,j}^{R,Y})}, \tag{6.9}$$

wherein the negative pairings within a mini-batch are dynamically re-scaled leveraging a user personality affinity function f. Within our proposed architecture, user attributes are encapsulated through their antecedent reviews. Specifically, the function f is formalized as:

$$f(i,j) = \alpha^{(1-sim(\tilde{R}_{(i)}, \tilde{R}_{(j)}))}, \tag{6.10}$$

whereby the weights of negative pairs with analogous historical interactions are diminished, whereas those with disparate histories are amplified. The variable α (with $\alpha > 1$) operates as a scaling hyperparameter for the negative samples, and *sim* denotes the cosine similarity, with $\tilde{R}_{(i)}$ and $\tilde{R}_{(j)}$ representing the averaged features from the historical reviews of two distinct users.

The optimization of the model is conducted through a composite loss function, encompassing both the cross-entropy loss and the dual contrastive losses:

$$\mathscr{L}_{total} = \mathscr{L}_{CE} + \lambda_1 \mathscr{L}_{CCL} + \lambda_2 \mathscr{L}_{PCL}, \tag{6.11}$$

where λ_1 and λ_2 serve as hyperparameters that calibrate the contribution of each contrastive loss to the overall optimization process.

6.4.4 Visual Grounding Metric

The objective of our model is to articulate explanations that aptly characterize the imagery within a specified collection of visuals. Conventional n-gram based evaluation frameworks such as BLEU, while formulated for the diagnostic assessment of machine translation outputs, lack

the acumen to appraise textual quality adequately. Their sensitivity is restricted to lexical discrepancies, thus they do not confer merits for semantic or syntactic divergences between the predicted outputs and reference texts [Reiter, 2018, Zhang et al., 2019, Sellam et al., 2020]. For a robust appraisal of the congruence between visual inputs and textual explanations, we propose an automated evaluation metric: CLIP-ALIGN, predicated on the findings of [Radford et al., 2021a].

Upon acquiring a set of images $I = \{i_1, i_2, ..., i_n\}$ and an assemblage of statements from the produced text $S = \{s_1, s_2, ..., s_m\}$, we employ CLIP to derive the embeddings for all images and sentences. The metric is calculated as follows:

$$\text{CLIP-ALIGN} = \frac{1}{n}\sum_{i=1}^{n} \max(\{cs_{1,i}, ..., cs_{m,i}\}) \tag{6.12}$$

where $cs_{i,j}$ represents the confidence score yielded by the CLIP-augmented classifier Φ, which is honed on our annotated dataset. Substituting $cs_{i,j}$ with the cosine similarity between embeddings of images and sentences, we attain an alternative metric, CLIP-SCORE, analogous to [Hessel et al., 2021].

In contrast to erstwhile CLIP-oriented metrics [Hessel et al., 2021, Zhu et al., 2021], CLIP-ALIGN is particularly attuned to the precision and the fidelity of the association between the depicted entities within the sentences and their corresponding images (for instance, “food is great” and “burger is great” would accrue commensurate elevated scores with an identical burger image when evaluated using CLIP-SCORE, and a model that persistently generates “food is great” might manifest inflated performance on CLIP-SCORE at the corpus level). Furthermore, the original CLIP-SCORE [Hessel et al., 2021] demonstrated limited correlation with captions that embody personal sentiments, hence its suboptimal suitability for this task.

6.5 Experiments

In this section, we undertake a comprehensive set of experiments to ascertain the efficacy of our personalized showcases framework. Through ablation studies, we discern the impact of various modalities on the personalization aspect of showcases. We further corroborate the diversity and precision of the explanations generated by our model through case studies and human evaluations, establishing its superiority over baseline models.

6.5.1 Experimental Setting

Baselines To validate the efficacy of our proposed model, we benchmark it against a suite of established baselines pertinent to various related domains, such as image captioning, medical report generation, and explanation generation in recommendations. These baselines include:

1. *ST* [Xu et al., 2015], an image captioning model that integrates Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks.

2. *R2Gen* [Chen et al., 2020c], which is a memory-augmented transformer architecture tailored for generating extensive textual content from visual inputs.

3. *Ref2Seq* [Ni et al., 2019b], a reference-driven sequence-to-sequence model that is prevalently employed for generating explanations within recommendation systems.

4. *Peter* [Li et al., 2021b], a contemporary transformer-based approach for explanation generation that incorporates user and item identifiers to inform the prediction of words within the target explanation.

5. *img* and *text* denote the use of visual and textual modalities, respectively, within these contexts.

Evaluation Metrics For image selection, we report Precision@K, Recall@K and F1@K to measure the ranking quality. Due to the nature of our task, we set a small K ($K = 3$). To evaluate

Table 6.2. Results on personalized showcases with different models and different input modalities. Results are reported in percentage (%). *GT* is the ground truth.

Model	Input	N-Gram Metrics				Diversity Metrics		Embedding Metrics		
		BLEU-1	BLEU-4	METEOR	NIST	DISTINCT-1	DISTINCT-2	CLIP-ALIGN	CLIP-SCORE	BERT-SCORE
GT	-	-	-	-	-	6.06	43.23	90.47	28.41	-
ST	*img*	8.24	0.28	3.41	28.08	2.74	17.41	80.84	24.31	85.20
R2Gen	*img*	6.47	0.22	3.10	36.55	3.23	22.45	82.07	24.28	85.89
Ref2Seq	*text*	7.09	**0.67**	3.80	30.78	0.92	5.89	73.51	23.83	84.71
Peter	*text*	8.89	0.44	3.28	34.45	0.38	1.27	72.70	23.27	86.94
Ours	*img*	9.92	0.32	3.64	37.35	3.37	26.37	84.78	**24.68**	88.03
	img+text	**10.40**	0.36	**3.83**	**50.64**	**3.58**	**28.58**	**85.31**	24.50	**88.23**

Table 6.3. Ablation study for personalized image selection. Results are reported in percentage (%).

Method	Accuracy			Diversity
	Prec@3	Recall@3	F1@3	Div@3
random	4.87	6.14	5.43	30.24
img	**25.21**	34.05	28.97	17.12
text	15.28	20.58	17.54	**18.68**
img+text	**25.21**	**34.37**	**29.09**	17.07

diversity, we introduce the truncated `div@K` ($K = 3$) for the average dissimilarities for all image pairs in recommended images. Formally, given K images $\{i_1, \ldots, i_K\}$, `div@K` is defined as:

$$div@K = \sum_{1 \leq m < n \leq K} \frac{dis(i_m, i_n)}{K(K-1)/2}. \tag{6.13}$$

For textual explanations, we first evaluate the relevance of generated text and ground truth by n-gram based text evaluation metrics: BLEU (n=1,4) [Papineni et al., 2002b], METEOR [Denkowski and Lavie, 2011] and NIST (n=4) [Doddington, 2002]. To evaluate diversity, we report DINSTINCT-1 and DISTINCT-2 which is proposed in [Li et al., 2015a] for text generation models. We then use CLIP and BERT to compute embedding-based metrics. CLIP-ALIGN is our proposed metrics in Section 6.4.4. CLIP-SCORE [Hessel et al., 2021] BERT-SCORE [Zhang et al., 2019] are two recent embedding-based metrics.

6.5.2 Framework Performance

We commence by presenting the performance of our framework as documented in Table 6.2, focusing primarily on text evaluation metrics where more intricate challenges and notable insights were observed. In this context, the input visual data are curated by our algorithm,[3] while the textual input comprises the users' historical critiques.

The disparity in performance between models that leverage textual inputs and those that utilize visual inputs on measures of diversity and CLIP-based metrics underscores the criticality of integrating visual information. Models equipped for visually-conscious generation demonstrate the capacity to fabricate precise explanations marked by a rich variety of linguistic expressions. Our PC^2L model evidences a marked enhancement across the majority of metrics relative to models grounded in LSTM and transformer architectures, underscoring the proficiency of a pretrained language model enhanced through contrastive learning in producing high-caliber explanations. Although the text-centric models *Ref2Seq* and *Peter* register competitive metrics in some n-gram evaluations like BLEU and METEOR, they manifest significantly lower performance in terms of diversity and embedding-based metrics. The textual output of these models is also characterized by redundancy and a lack of informativeness, as frequently evidenced by the generation of repetitive and non-descriptive sentences—a finding we corroborate through human evaluations and a detailed case study.

6.5.3 Component Analysis

Ablation studies are executed to ascertain the individual contribution and efficacy of each constituent within the system.

Model for image set selection. Table 6.3 provides an overview of the performance of our personalized image set selection process. In terms of general ranking performance, we conducted comparisons against the following approaches:

[3]To ensure a fair assessment of visual-textual alignment, the candidate pool for image selection incorporates the ground truth images associated with a particular user.

Table 6.4. Ablation study on contrastive learning. Baseline is to train a multi-modal decoder without contrastive learning.

Method	BLEU-1	DISTINCT-2	CLIP-ALIGN
Baseline	7.96	25.90	82.50
img CL + text CL	9.72	27.58	84.03
CCL+ text CL	10.19	28.10	85.12
img CL + PCL	9.96	28.32	84.15
PC^2*L*	**10.40**	**28.58**	**85.31**

1. **Random Selection**: This method involves random selection of images from the candidate pool. It is noteworthy that this approach yields significantly worse ranking performance compared to our model. However, it does exhibit the highest truncated diversity among the methods considered.

2. **Multi-Modal Approach with User Historical Information**: We introduced a multi-modal model that incorporates both user historical images and text. This approach outperforms a single-modal model. Notably, the text-only model achieves the highest diversity, primarily due to its relatively lower ranking accuracy, which is comparable to random selection.

From our observations, we can draw the following conclusions:

1. Given the notably low accuracy of random selection, it becomes evident that personalized ranking is imperative for effective image set selection.

2. Our multi-modal model, which leverages both image and text inputs, achieves the best ranking performance. However, there is still room for improvement in terms of diversity, as it does not exhibit the highest diversity among the methods evaluated.

These findings show the significance of personalized ranking and the potential for further enhancing the diversity of our image set selection process.

Effectiveness of Contrastive Learning We implement ablation studies on varying configurations of our contrastive loss to validate the efficacy of our proposed method. As illustrated in Table 6.4,

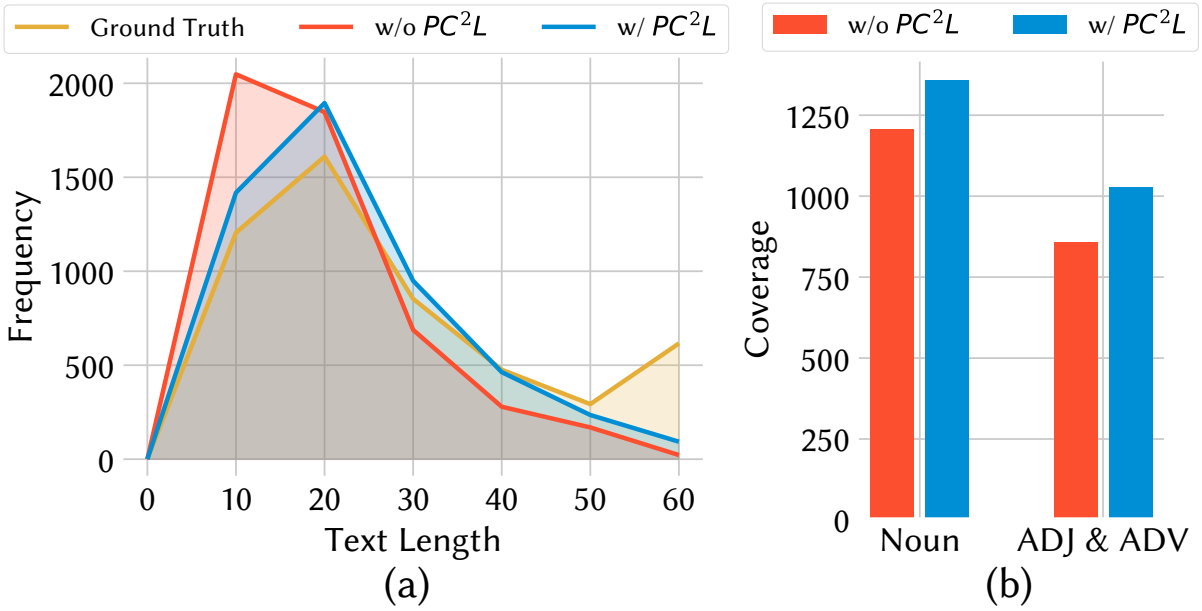


Figure 6.6. (a) The length distributions of generated texts on the test set. (b) The generated explanation coverage of nouns (Noun), adjectives (ADJ) and adverbs (ADV) in ground truth.

our PC^2L outperforms all other considered baselines across multiple metrics. Notably, *CCL* significantly enhances visual grounding capabilities by prompting the model to differentiate between random entities and the correct ones, thus advancing the CLIP-ALIGN metric over the standard contrastive framework as presented in [Chen et al., 2020a]. Conversely, *PCL* primarily augments diversity by incentivizing the model to prioritize users with divergent interests.

In our investigation of the qualitative improvements attributable to contrastive learning, we dissect the generated explanations along two dimensions: distribution of generation lengths and the coverage of key lexemes. Figure 6.6 (a) juxtaposes the length distribution of generated texts against the authentic data, with categorization into six brackets (in intervals of 10, ranging from 0 to 60). Models devoid of PC^2L manifest a more peaked distribution, whereas the inclusion of PC^2L yields a distribution that more closely approximates the actual data, substantiating its utility and capacity for generalization across novel images. It is noteworthy that the authentic data encompasses a higher incidence of longer texts compared to model outputs, which can be attributed to an imposed maximum length of 64 during both training and inference phases.

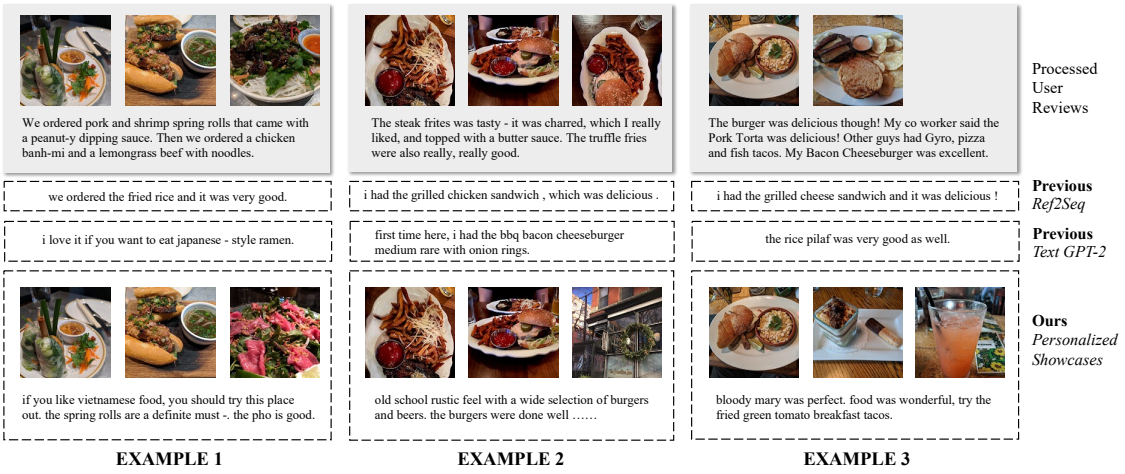


Figure 6.7. Comparison between text-only explanations (i.e., *Ref2Seq* and *Text GPT-2*) and our showcases.

Table 6.5. Ablation Study on different initializations of the decoder. *Random* randomly initializes model weights. *Text GPT-2* and *Img GPT-2* are initialized with weights from [Radford et al., 2019]. *Img GPT-2 + FT* finetunes the model on a corpus similar to our training text data. Results are in percentage (%).

Method	BLEU-1	DISTINCT-1	DISTINCT-2
Img Random	5.21	0.23	5.08
Text GPT-2	4.81	3.43	19.27
Img GPT-2	**7.59**	4.05	29.41
Img GPT-2 + FT	7.10	**4.32**	**30.82**

Figure 6.6 (b) assesses the keyword coverage—specifically nouns, adjectives, and adverbs—in the output text. An output is deemed to encompass a keyword if said keyword is present in the respective authentic data. Upon comparing models trained with and without the inclusion of PC^2L, it is evident that PC^2L elevates the presence of all categories of key terms, which suggests that our contrastive learning strategy effectively enhances the diversity and personalization of the generated text.

To summarize, the integration of contrastive learning within the multi-modal explanation generation process culminates in superior quality outputs, characterized by heightened diversity and enhanced visual-textual congruence.

Can GPT-2 provide linguistic knowledge? Subsequently, we examine if the incorporation of

GPT-2 as the decoding component can infuse linguistic expertise pertinent to our generation endeavor. We subject the model to training with disparate weight initializations, with the input comprising either veridical images (Img) or antecedent reviews (Text). As depicted in Table 6.5, a comparative analysis between the outputs of models with random initialization and those leveraging GPT-2's pretrained weights reveals a pronounced efficacy of the pretrained initialization in enhancing the generation quality, pertinent to both image and textual inputs. Furthermore, the act of finetuning with domain-specific data (constituting 260k samples drawn from users who have penned a single review and are thus omitted from our personalization dataset) serves to bolster the domain acumen of the decoder. This finetuning exercise yields tangible benefits, particularly in amplifying the diversity indices of the generated content.

6.5.4 Case Study

We conduct a comparative analysis of three instances (refer to Figure 6.7) to assess the superiority of our personalized showcases over the conventional monomodal explanations offered by *Ref2Seq* and *Text GPT-2*. Predominantly, our multimodal explanations are comprehensive, encapsulating the majority of the imagery depicted in user critiques. This observation corroborates the efficacy of our image set curation algorithm, verifying that the curated images are suitably representative of the visuals that users find engaging.

Furthermore, the inclusion of images furnishes ancillary grounding for textual generation, thereby endowing our textual explanations with heightened specificity (e.g., particular culinary dishes). Illustrated in Figure 6.7, it is evident that sole reliance on historical textual reviews fails to yield accurate explanations (noted in Case 1), as the explanations derived from *Ref2Seq* and *Text GPT-2* lack relevance to the user's actual feedback. Additionally, these explanations exhibit a lack of variety (as observed in Case 2). In stark contrast, our approach yields textual elucidations that are both pertinent and varied, informed by the visual context. However, in Case 3, our generated narrative does not encapsulate the user's review adequately as it pertains to only a single image from the selected set. This highlights the ongoing challenge of synthesizing texts

Table 6.6. Human evaluation results on two models. We present the workers with reference text and images, and ask them to give scores from different aspects. Results are statistically significant with p<0.01.

Method	Expressiveness	Visual Alignment
Ref2Seq	3.72	3.65
PC^{2}L	**4.25**	**4.10**

that aggregate information across multiple images within this domain.

The analysis of these examples further reinforces the finding that *Ref2Seq* tends to generate patterned explanations with limited distinctiveness, aligning with the observations noted in Table 6.2 regarding its diminished DISTINCT-1 and DISTINCT-2 scores.

6.5.5 Human Evaluation

To thoroughly assess our model, we engage in human evaluation through Amazon Mechanical Turk. [4] For each experimental model, we randomly select 500 instances from the testing corpus. Three human judges rate each instance employing a 5-point Likert scale to mitigate variance. We direct the evaluators to consider dual aspects: expressiveness (which encompasses semantic accuracy, diversity, and absence of redundancy) and visual-textual congruence (ensuring the textual content is representative of the imagery context). As indicated in Table 6.6, our *PC^{2}L* model demonstrates a significant advantage over *Ref2Seq* in performance, corroborating the outcomes from automated evaluation metrics.

6.6 Conclusion

In this chapter, we present a novel task designated as *personalized showcases*, devised to enrich recommendations with detailed explanations, for which we create an extensive dataset GEST from *Google Local*. Our proposed framework leverages a multi-modal explanation mechanism, augmented by contrastive learning, to derive visual and textual insights from user-

[4] https://www.mturk.com/

generated reviews. Empirical evidence suggests that our *showcases* yield explanations that surpass traditional text-only methods in terms of informativeness and variety. Moreover, we recognize that the task of visual grounding across multiple images presents a significant challenge within our *showcases* framework. Therefore, advancing the utilization of multi-modal data and enhancing the visual-textual congruence remains a critical focus for future endeavors in this domain. We anticipate that our dataset and framework will serve as valuable assets for the community's ongoing multi-modal and recommendation systems research.

Chapter 6, in part, is a reprint of the material as it appears in "Personalized Showcases: Generating Multi-Modal Explanations for Recommendations" by An Yan*, Zhankui He*, Jiacheng Li*, Tianyang Zhang, Julian McAuley, which was published in *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 7

Related Work

7.1 Related Work for TALLOR

Bootstrapping is a well-established method for model induction, originating from a limited set of initial seeds. It has been utilized for tasks such as word sense disambiguation [Yarowsky, 1995]. Additionally, it has been applied for lexicon expansion, with the objective of enlarging an existing lexicon with additional semantically related terms [Shen et al., 2017, Yan et al., 2019]. Huang and Riloff [2010] developed a bootstrapping approach for training semantic class taggers using a minimal set of seed examples. However, their methodology was primarily concentrated on the assignment of semantic tags to pre-identified phrases, omitting the aspect of entity boundary detection. Our research is oriented towards the autonomous derivation of logical rules from an initial seed set, with a dual focus on identifying entity boundaries and classifying entity labels concurrently.

Distant supervision has been introduced to minimize manual annotation efforts by employing pre-existing lexicons or knowledge bases to train models [Mintz et al., 2009]. In the realm of Named Entity Recognition (NER), various strategies have been adopted to implement systems under distant supervision [Ren et al., 2015, Fries et al., 2017, Giannakopoulos et al., 2017]. For instance, AutoNER [Shang et al., 2018b] trains a NER system by harnessing both typed lexicons and untapped phrases as sources of supervision. Peng et al. [2019] proposed the AdaPU algorithm, which utilizes incomplete dictionaries as a supervisory signal. However, the

availability of lexicons or knowledge bases is not a given, especially within specialized fields and resource-scarce environments, where their creation can be both time-intensive and laborious.

In the context of weak label acquisition, methods have been put forth that leverage manually authored labeling functions [Bach et al., 2017]. Building on this concept, several methodologies [Safranchik et al., 2020, Lison et al., 2020] have been implemented for NER by postulating the presence of a sufficient number of handcrafted labeling functions and lexicons. Nevertheless, the manual composition of labeling rules can be prohibitively costly and typically necessitates expertise within the domain. Our endeavor is to automate the learning of logical rules, thereby significantly diminishing the need for human input.

7.2 Related Work for UCTopic

Numerous strategies have been employed to distill topical phrases utilizing Latent Dirichlet Allocation (LDA) [Blei et al., 2003]. Wallach [2006] enhanced LDA by integrating a hierarchical Dirichlet generative probabilistic model, enabling topic sharing across words in a bigram structure. The Topical N-gram (TNG) model [Wang et al., 2007] introduced additional latent variables and word-specific multinomials to better represent bi-grams, subsequently aggregating them into n-gram phrases. PD-LDA [Lindsey et al., 2012] adopted a hierarchical Pitman-Yor process to apply a common topic to all words within an n-gram. Danilevsky et al. [2014] scored extracted phrases using four heuristic metrics. TOPMine [El-Kishky et al., 2014] imposed a constraint whereby all terms in a phrase must align with a single latent topic, and phrases were assigned topics based on their constituent words. Our approach surpasses prior topic mining methodologies by leveraging the capabilities of pre-trained language models and applying unsupervised contrastive learning on extensive datasets, delivering preeminent pre-trained phrase representations and refinements for topic mining tasks.

In terms of phrase representation, initial methods relied on a composition function that amalgamates word embeddings to form cohesive phrase embeddings. Yu and Dredze [2015]

devised a rule-based compositional function over word vectors. Zhou et al. [2017] engaged a pairwise GRU model along with datasets like PPDB [Pavlick et al., 2015] for phrase representation learning. Phrase-BERT [Wang et al., 2021] assembled token embeddings from BERT, pretraining on positive instances generated by a GPT-2-derived diverse paraphrasing model [Krishna et al., 2020]. Lee et al. [2021] derived phrase representations under the guidance of reading comprehension tasks, subsequently employing those in open-domain question answering. Additional research has tailored phrase embeddings to specific applications, including semantic parsing [Socher et al., 2011] and machine translation [Bing et al., 2015]. In this work, we introduce an unsupervised contrastive learning methodology for pre-training general-purpose phrase representations and for finetuning towards topic-specific phrase representations.

7.3 Related Work for Chapter 4

Significant research has been dedicated to enhancing the interpretability of recommendation algorithms. Catherine and Cohen [2017] have developed techniques to infer latent representations from review texts for rating predictions, which in turn assist in pinpointing the most pertinent reviews for a specific user-item combination. An alternate prevalent strategy involves crafting textual justifications for recommendations. An attribute-to-sequence model for generating descriptive product reviews was introduced by Dong et al. [2017], employing categorical attributes. Multi-task learning frameworks that simultaneously address collaborative filtering and review synthesis have been established by Ni et al. [2017]. Additionally, Li et al. [2019b] designed a system for producing tips reflective of 'persona' data, accommodating user language styles and item traits. Despite these advances, these methods typically train on entire reviews or tips, which may be suboptimal due to varying review quality. More recent efforts by Liu et al. [2019a] entailed constructing a model for generating detailed explanations in text classification, using a dataset composed of user-generated ratings and summaries. This level of detail, however, is often lacking in most platforms. Contrarily, our research excels by

extracting justifications from reviews to serve as training instances, and we demonstrate through comprehensive experimentation that this approach provides a superior foundation for explainable recommendation systems.

Diversity remains a crucial characteristic for NLG technologies. Cutting-edge research has aimed to harness existing knowledge to bolster generative diversity. A technique to integrate planned narratives into story creation was proposed by Yao et al. [2019]. An approach for aspect-oriented, coarse-to-fine review generation was advanced by Li et al. [2019a], predicting aspects for each sentence to delineate the review's content progression. Subsequently, a sequence of sentence outlines is created, which is then elaborated upon by a decoder. In the realm of conversational systems, methodologies to derive templates from past interactions have been investigated, which are later modified to produce novel responses [Weston et al., 2018, Wu et al., 2018]. This extract-and-edit approach has also been explored in NLG style transfer tasks [Li et al., 2018]. An attribute-centric masked language model for non-parallel sentiment alteration was presented by Wu et al. [2019], masking sentiment-driven tokens and then training the model to fill these gaps to reflect the intended sentiment. In our work, we introduce a conditional masked language model that accounts for more granular aspects.

7.4 Related Work for UCEPIC

The elucidation of user-specific recommendation reasoning, encompassing various forms such as item features, attributes, and user similarity, has been a research focus for an extended period [Zhang et al., 2020c, 2014b, Gao et al., 2019]. Of late, the generation of explanations in natural language has garnered increased interest [Li et al., 2021b, Ni and McAuley, 2018, Lu et al., 2018, Li et al., 2017b, 2020b, 2023, Ni et al., 2019a] with the intention to create post-hoc narratives or justifications that reflect individual user preferences. For instance, Li et al. [2017b] employed an RNN-based architecture to produce explanations aligned with predicted ratings. To enhance the precision of generated explanations, Ni et al. [2019a] leveraged item

aspects to guide the semantics, while Li et al. [2021b] introduced a transformer tailored to item characteristics for personalized explanation synthesis. Additionally, methods that derive explanations from user reviews are linked to the review generation domain, where various controlled generators [Tang et al., 2016, Dong et al., 2017, Ni and McAuley, 2018] have been adapted for initial explanatory generation models. While existing studies have advanced the controllability of generation processes based on auto-regressive models [Li et al., 2019a, Ni and McAuley, 2018, Li et al., 2020a, 2021a, Hua and Wang, 2019, Moryossef et al., 2019] focusing on aspect planning, our work enhances this control and the informativeness of explanations by integrating aspect planning with lexical constraints within an insertion-based generation paradigm.

Text generation with lexical constraints mandates the inclusion of specified terms within the output. Prior research predominantly employs specialized decoding strategies. Hokamp and Liu [2017] introduced a decoding mechanism that integrates lexical constraints within a grid beam search. Post and Vilar [2018] proposed an approach that minimizes complexity in decoding as constraints increase. Enhancements to decoding efficiency were further achieved by Hu et al. [2019] through vectorized dynamic beam allocation. Sampling-based decoding strategies were also explored by Miao et al. [2019], who utilized a Metropolis-Hastings sampling process, starting with constraint placement within a template followed by word decoding. Although these methods are effective, they generally require substantial computational complexity. A more recent advancement by Zhang et al. [2020b] achieved hard-constrained generation with an improved time complexity of $\mathcal{O}(\log n)$ by integrating pre-trained language models and insertion-based generation methodologies [Stern et al., 2019, Gu et al., 2019b, Chan et al., 2019, Gu et al., 2019a] initially utilized in machine translation. Concurrently, CBART [He, 2021] harnesses the pre-trained BART model [Lewis et al., 2020], wherein the encoder and decoder facilitate instruction and mask prediction, respectively.

7.5 Related Work for Chapter 6

Considerable research has focused on crafting explanations for recommendations. Several approaches have emerged, including the generation of product reviews from categorical attributes [Dong et al., 2017], images [Truong and Lauw, 2019], or aspects [Ni and McAuley, 2018]. Recognizing the variability in review quality, Li et al. [2019b] created succinct and informative 'tips' from the Yelp dataset as recommendation explanations. To elevate the generation quality, Ni et al. [2019b] suggested segmenting reviews and classifying these segments to isolate effective justifications. A transformer-based model that incorporates user and item embeddings alongside pertinent features was introduced by Li et al. [2021b] for generating explanation narratives. These methods typically draw upon historical user or item reviews. However, imagery provides a wealth of context for generating text. In our task, multi-modal data combining images and text often results in explanations that are more comprehensible to users.

The recent proliferation of deep learning in multi-modal learning and pretraining has been noteworthy [Tan and Bansal, 2019, Huang et al., 2020, Radford et al., 2021b, Chen et al., 2021]. These architectures commonly employ the Transformer structure [Vaswani et al., 2017] to encode visual and textual inputs, enhancing multimodal tasks. Notably, CLIP [Radford et al., 2021b], trained on vast image-caption datasets, has demonstrated robust zero-shot performance across various vision and language assignments [Shen et al., 2021]. Other methodologies [Hessel et al., 2021, Zhu et al., 2021] have utilized CLIP embeddings to assess modality congruence as benchmarks for tasks like image captioning and text generation.

Contrastive learning, aiming to distinguish representations by contrasting positive with negative instances, has seen widespread application across diverse machine learning disciplines, including computer vision [Chen et al., 2020a, Khosla et al., 2020, He et al., 2020], natural language processing [Huang et al., 2018, Fang et al., 2020, Gao et al., 2021b], and recommender systems [Xie et al., 2020b, Zhou et al., 2021, Wei et al., 2021]. Recent studies have yielded promising outcomes by applying contrastive learning to conditional text generation, either by

creating adversarial exemplars [Lee et al., 2020] or identifying challenging negatives using pretrained language models [Cai et al., 2020, Yan et al., 2021].

Chapter 8

Conclusion and Future Outlook

In this dissertation, I have presented the core trajectories of my research in developing systems that generate explanations for recommendations. My exploration has centered on enhancing the explicability of personalized systems through advanced extraction and comprehension methodologies.

For the extraction component, I have introduced a bootstrapping framework that operates under weakly supervised conditions for named entity recognition. This innovative approach necessitates only a minimal set of seed rules for entity identification, enabling users to steer the recognition process by selecting desired rules. This adaptability allows for easy customization of target entities within personalized systems, tailoring them to specific user needs.

To deepen the understanding of the phrases extracted, I have proposed a novel method of phrase representation learning anchored in contrastive learning. This technique requires no direct supervision and is adeptly transferable across various domains, facilitating the acquisition of domain-specific representations tailored to diverse recommendation contexts.

Finally, leveraging the information procured, my work investigates several modalities of controllable explanation generation, including topic-, phrase-, and image-based approaches. These methodologies significantly enhance the diversity, relevance, and richness of the explanations generated, thereby advancing the field of explainable recommendation systems.

We foresee several opportunities for future research to further the themes we have

explored in this dissertation.

1. **Integration with More Advanced Language Models**: As large language models (LLMs) continue to evolve, integrating the latest models into the explanation generation framework could yield more nuanced and contextually aware explanations. Leveraging models like Llama-2 or other state-of-the-art architectures could enhance the naturalness, coherence, and specificity of the generated text.

2. **Personalization at Scale**: Exploring ways to personalize explanations at scale using LLMs, possibly through user profiling or incorporating user feedback loops, could lead to more tailored and user-centric recommendation systems. This might involve developing LLMs that can dynamically adjust explanation styles and content based on individual user preferences.

3. **Multimodal Explanation Enhancement**: Given the rise of multimodal LLMs, future work could involve integrating text with other modalities like images, videos, or audio to create richer, more engaging explanations. This could involve harnessing models that can process and generate multimodal content coherently.

4. **Interactive Explanation Systems**: Developing interactive systems where users can critique and receive explanations in real time could be a significant advancement. This would require LLMs to not only generate explanations but also understand and respond to user critiques dynamically.

Bibliography

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In *EMNLP*, 2017.

Stephen H. Bach, B. He, A. Ratner, and C. Ré. Learning the structure of generative models without labeled data. *Proceedings of machine learning research*, 70:273–82, 2017.

Ashutosh Baheti, Alan Ritter, Jiwei Li, and William B. Dolan. Generating more interesting responses in neural conversation models with distributional constraints. In *EMNLP*, 2018.

Jinze Bai, Chang Zhou, Junshuai Song, Xiaoru Qu, Weiting An, Zhao Li, and Jun Gao. Personalized bundle list recommendation. *The World Wide Web Conference*, 2019.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca J. Passonneau. Abstractive multi-document summarization via phrase selection and merging. In *ACL*, 2015.

David M. Blei, A. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

Hengyi Cai, Hongshen Chen, Yonghao Song, Zhuoye Ding, Yongjun Bao, Weipeng Yan, and Xiaofang Zhao. Group-wise contrastive learning for neural dialogue generation. *arXiv preprint arXiv:2009.07543*, 2020.

Rose Catherine and William W. Cohen. Transnets: Learning to transform for recommendation. In *RecSys*, 2017.

William Chan, Nikita Kitaev, Kelvin Guu, Mitchell Stern, and Jakob Uszkoreit. Kermit: Generative insertion-based modeling for sequences. *ArXiv*, abs/1906.01604, 2019.

Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei.

Reading tea leaves: How humans interpret topic models. In *NIPS*, 2009.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. 2021.

Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. On sampling strategies for neural network-based collaborative filtering. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020b.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020c.

Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. Co-attentive multi-task learning for explainable recommendation. In *IJCAI*, 2019.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

Marina Danilevsky, Chi Wang, Nihit Desai, Xiang Ren, Jingyi Guo, and Jiawei Han. Automatic construction and ranking of topical keyphrases on collections of short documents. In *SDM*, 2014.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91, 2011.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4418. URL https://www.aclweb.org/anthology/W17-4418.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019a.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep

bidirectional transformers for language understanding. In *NAACL*, 2019b.

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, 2002.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. Learning to generate product reviews from attributes. In *EACL*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021.

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, and Jiawei Han. Scalable topical phrase mining from text corpora. *ArXiv*, abs/1406.6312, 2014.

Günes Erkan and Dragomir R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479, 2004.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*, 2020.

Jason Alan Fries, Sen Wu, A. Ratner, and Christopher Ré. Swellshark: A generative model for biomedical named entity recognition without labeled data. *ArXiv*, abs/1704.06360, 2017.

Ygor Gallina, Florian Boudin, and Béatrice Daille. Kptimes: A large-scale dataset for keyphrase generation on news documents. In *INLG*, 2019.

Jingyue Gao, Xiting Wang, Yasha Wang, and Xing Xie. Explainable recommendation through attentive multi-view learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3622–3629, 2019.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. Generating multiple diverse responses for short-text conversation. In *AAAI*, 2018.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv*, abs/2104.08821, 2021a.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021b.

Athanasios Giannakopoulos, C. Musat, Andreea Hossmann, and Michael Baeriswyl. Unsupervised aspect term extraction with b-lstm & crf using automatically labelled datasets. In

WASSA@EMNLP, 2017.

Jiatao Gu, Qi Liu, and Kyunghyun Cho. Insertion-based decoding with automatically inferred generation order. *Transactions of the Association for Computational Linguistics*, 7:661–676, 2019a.

Jiatao Gu, Changhan Wang, and Junbo Zhao. Levenshtein transformer. volume 32, 2019b.

S. Gupta and Christopher D. Manning. Improved pattern learning for bootstrapped entity extraction. In *CoNLL*, 2014.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 2:1735–1742, 2006.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. An unsupervised neural attention model for aspect extraction. In *ACL*, 2017.

Xingwei He. Parallel refinements for lexically constrained text generation with bart. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8666, 2021.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. *ArXiv*, abs/1705.00652, 2017.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv: Learning*, 2016.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Chris Hokamp and Qun Liu. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, 2017.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural

text degeneration. *arXiv preprint arXiv:1904.09751*, 2019a.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *CoRR*, abs/1904.09751, 2019b.

J Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. Improved lexically constrained decoding for translation and monolingual rewriting. 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 839–850, 2019.

X Hua and L Wang. Sentence-level content planning and style specification for neural text generation. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

Jiaji Huang, Yi Li, Wei Ping, and Liang Huang. Large margin neural language model. *arXiv preprint arXiv:1808.08987*, 2018.

Ruihong Huang and E. Riloff. Inducing domain-specific semantic class taggers from (almost) nothing. In *ACL*, 2010.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan L. Boyd-Graber, and Hal Daumé. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *NAACL*, 2016.

Zhengbao Jiang, W. Xu, J. Araki, and Graham Neubig. Generalizing natural language analysis through span-relation representations. In *ACL*, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

Martin Krallinger, O. Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Dong-Hong Ji, D. M. Lowe, R. Sayle, R. Batista-Navarro, R. Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, D. Campos, Buzhou Tang, H. Xu, Tsendsuren Munkhdalai, K. Ryu, S. V. Ramanan, P. S. Nathan, S. Zitnik, M. Bajec, L. Weber, Matthias Irmer, S. Akhondi, J. Kors, S. Xu, X. An, Utpal Kumar Sikdar, A. Ekbal, M. Yoshioka, Thaer M. Dieb, Miji Choi, Karin M. Verspoor, Madian Khabsa, C. Lee Giles, H. Liu, K. Ravikumar, Andre Lamurias, F. Couto, Hong-Jie Dai, R. Tsai, C. Ata, T. Can, Anabel Usie, Rui Alves, Isabel Segura-Bedmar, Paloma Martínez, J. Oyarzábal, and A. Valencia. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7:S2 – S2,

2015.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. Reformulating unsupervised style transfer as paraphrase generation. *ArXiv*, abs/2010.05700, 2020.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Found. Trends Mach. Learn.*, 5:123–286, 2012.

Jinhyuk Lee, Mujeen Sung, Jaewoo Kang, and Danqi Chen. Learning dense representations of phrases at scale. In *ACL/IJCNLP*, 2021.

Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. Contrastive learning with adversarial perturbations for conditional text generation. *arXiv preprint arXiv:2012.07280*, 2020.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

J. Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, A. P. Davis, C. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016, 2016a.

Jiacheng Li, Jingbo Shang, and Julian McAuley. UCTopic: Unsupervised contrastive learning for phrase representations and topic mining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6159–6169, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.426. URL https://aclanthology.org/2022.acl-long.426.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015a.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B. Dolan. A diversity-promoting objective function for neural conversation models. In *HLT-NAACL*, 2015b.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, 2016b.

Juncen Li, Robin Jia, He He, and Percy S. Liang. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *NAACL-HLT*, 2018.

Junyi Li, Wayne Xin Zhao, Ji-Rong Wen, and Yang Song. Generating long and informative reviews with aspect-aware coarse-to-fine decoding. In *ACL*, 2019a.

Junyi Li, Siqing Li, Wayne Xin Zhao, Gaole He, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. Knowledge-enhanced personalized review generation with capsule graph neural network. pages 735–744, 2020a.

Junyi Li, Wayne Xin Zhao, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. Knowledge-based review generation by coherence enhanced text planning. pages 183–192, 2021a.

Lei Li, Yongfeng Zhang, and Li Chen. Generate neural template explanations for recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 755–764, 2020b.

Lei Li, Yongfeng Zhang, and Li Chen. Personalized transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4947–4957, 2021b.

Lei Li, Yongfeng Zhang, and Li Chen. Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems*, 41(4):1–26, 2023.

Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. Neural rating regression with abstractive tips generation for recommendation. In *SIGIR*, 2017a.

Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. Neural rating regression with abstractive tips generation for recommendation. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 345–354, 2017b.

Piji Li, Zihao Wang, Lidong Bing, and Wai Lam. Persona-aware tips generation. In *WWW*, 2019b.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*, 2004.

Robert V. Lindsey, Will Headden, and Michael Stipicevic. A phrase-discovering topic model using hierarchical pitman-yor processes. In *EMNLP*, 2012.

P. Lison, A. Hubin, Jeremy Barnes, and Samia Touileb. Named entity recognition without labelled data: A weak supervision approach. *ArXiv*, abs/2004.14723, 2020.

Hui Liu, Qingyu Yin, and William Yang Wang. Towards explainable nlp: A generative explanation framework for text classification. In *ACL*, 2019a.

Jingjing Liu, Panupong Pasupat, Yining Wang, D. Scott Cyphers, and James R. Glass. Query understanding enhanced by hierarchical parsing structures. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 72–77, 2013.

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*, abs/1907.11692, 2019b.

Yichao Lu, Ruihai Dong, and Barry Smyth. Why i like it: multi-task learning for recommendation and explanation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 4–12, 2018.

William C. Mann and Sandra A. Thompson. Rhetorical structure theory: toward a functional theory of text. 1988.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.

Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. pages 897–908, 2013.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. Deep keyphrase generation. In *ACL*, 2017.

Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul N. Bennett, Jiawei Han, and Xia Song. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *ArXiv*, abs/2102.08473, 2021.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6834–6842, 2019.

David Mimno. Using phrases in mallet topic models. http://www.mimno.org/articles/phrases/, 2015.

M. Mintz, Steven Bills, R. Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL/IJCNLP*, 2009.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. Step-by-step: Separating planning from realization in neural data-to-text generation. pages 2267–2277, 2019.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

Jianmo Ni and Julian McAuley. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *ACL*, 2018.

Jianmo Ni, Zachary C. Lipton, Sharad Vikram, and Julian J. McAuley. Estimating reactions and recommending products with generative models of reviews. In *IJCNLP*, 2017.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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 188–197, 2019a.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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 188–197, 2019b.

Cheng Niu, Wei Li, Jihong Ding, and Rohini K Srihari. A bootstrapping approach to named entity classification using successive learners. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 335–342, 2003.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002a.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002b.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *ACL*, 2015.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and X. Huang. Distantly supervised named entity recognition using positive-unlabeled learning. *ArXiv*, abs/1906.01378, 2019.

Jeffrey Pennington, R. Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

Matt Post and David Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. 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 1314–1324, 2018.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021a.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021b.

Ehud Reiter. A structured review of the validity of bleu. *Computational Linguistics*, 44(3): 393–401, 2018.

Xiang Ren, Ahmed El-Kishky, C. Wang, Fangbo Tao, Clare R. Voss, and Jiawei Han. Clustype: Effective entity recognition and typing by relation phrase-based clustering. *KDD : proceedings. International Conference on Knowledge Discovery & Data Mining*, 2015:995–1004, 2015.

Esteban Safranchik, Shiying Luo, and Stephen H. Bach. Weakly supervised sequence tagging from noisy rules. In *AAAI*, 2020.

E. T. K. Sang and F. D. Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *ArXiv*, cs.CL/0306050, 2003.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.

Jingbo Shang, Jialu Liu, Meng Jiang, X. Ren, Clare R. Voss, and Jiawei Han. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30:1825–1837, 2018a.

Jingbo Shang, Liyuan Liu, X. Ren, X. Gu, Teng Ren, and Jiawei Han. Learning named entity tagger using domain-specific dictionary. *ArXiv*, abs/1809.03599, 2018b.

J. Shen, Zeqiu Wu, Dongming Lei, Jingbo Shang, Xiang Ren, and Jiawei Han. Setexpan: Corpus-based set expansion via context feature selection and rank ensemble. *ArXiv*, abs/1910.08192, 2017.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

Richard Socher, Cliff Chiung-Yu Lin, A. Ng, and Christopher D. Manning. Parsing natural scenes and natural language with recursive neural networks. In *ICML*, 2011.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016.

Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit. Insertion transformer: Flexible sequence generation via insertion operations. In *International Conference on Machine Learning*, pages 5976–5985. PMLR, 2019.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Jian Tang, Yifan Yang, Samuel Carton, Ming Zhang, and Qiaozhu Mei. Context-aware natural language generation with recurrent neural networks. *ArXiv*, abs/1611.09900, 2016.

M. Thelen and E. Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *EMNLP*, 2002.

Quoc-Tuan Truong and Hady Lauw. Multimodal review generation for recommender systems. In *The World Wide Web Conference*, pages 1864–1874, 2019.

Stéphan Tulkens and Andreas van Cranenburgh. Embarrassingly simple unsupervised aspect extraction. In *ACL*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Hanna M. Wallach. Topic modeling: beyond bag-of-words. *Proceedings of the 23rd international conference on Machine learning*, 2006.

Shufan Wang, Laure Thompson, and Mohit Iyyer. Phrase-bert: Improved phrase embeddings from bert with an application to corpus exploration. *ArXiv*, abs/2109.06304, 2021.

Xuerui Wang, Andrew McCallum, and Xing Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 697–702, 2007.

Yizhong Wang, Sujian Li, and Jingfeng Yang. Toward fast and accurate neural discourse segmentation. In *EMNLP*, 2018.

Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. Contrastive learning for cold-start recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5382–5390, 2021.

Sean Welleck, Kianté Brantley, Hal Daumé Iii, and Kyunghyun Cho. Non-monotonic sequential text generation. In *International Conference on Machine Learning*, pages 6716–6726. PMLR, 2019.

Jason Weston, Emily Dinan, and Alexander H. Miller. Retrieve and refine: Improved sequence generation models for dialogue. In *SCAI@EMNLP*, 2018.

Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. Practical diversified recommendations on youtube with determinantal point processes. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. Mask and infill: Applying masked language model to sentiment transfer. In *IJCAI*, 2019.

Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. Response generation by context-aware prototype editing. In *AAAI*, 2018.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. Clear: Contrastive learning for sentence representation. *ArXiv*, abs/2012.15466, 2020.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. *arXiv: Learning*, 2020a.

Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Bolin Ding, and Bin Cui. Contrastive learning for sequential recommendation. *arXiv preprint arXiv:2010.14395*, 2020b.

Ke Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. Luke: Deep contextualized entity representations with entity-aware self-attention. In *EMNLP*, 2020.

An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. Weakly supervised contrastive learning for chest x-ray report generation. *arXiv preprint arXiv:2109.12242*, 2021.

Lingyong Yan, Xianpei Han, L. Sun, and B. He. Learning to bootstrap for entity set expansion. In *EMNLP/IJCNLP*, 2019.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. Plan-and-write: Towards better automatic storytelling. In *AAAI*, 2019.

David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*, 1995.

Mo Yu and Mark Dredze. Learning composition models for phrase embeddings. *Transactions of the Association for Computational Linguistics*, 3:227–242, 2015.

Hongyu Zang and Xiaojun Wan. Towards automatic generation of product reviews from aspect-sentiment scores. In *INLG*, 2017.

Hongyi Zhang, Moustapha Cissé, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *ArXiv*, abs/1710.09412, 2018.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *ICLR*, 2020a.

Xiang Zhang, Junbo Jake Zhao, and Yann André LeCun. Character-level convolutional networks for text classification. *ArXiv*, abs/1509.01626, 2015.

Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, and William B Dolan. Pointer: Constrained progressive text generation via insertion-based generative pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8649–8670, 2020b.

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *SIGIR*, 2014a.

Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 83–92, 2014b.

Yongfeng Zhang, Xu Chen, et al. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1):1–101, 2020c.

Yunyi Zhang, Jiaming Shen, Jingbo Shang, and Jiawei Han. Empower entity set expansion via language model probing. In *ACL*, 2020d.

Chang Zhou, Jianxin Ma, Jianwei Zhang, Jingren Zhou, and Hongxia Yang. Contrastive learning for debiased candidate generation in large-scale recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3985–3995, 2021.

Zhihao Zhou, Lifu Huang, and Heng Ji. Learning phrase embeddings from paraphrases with grus. *ArXiv*, abs/1710.05094, 2017.

Wanrong Zhu, Xin Eric Wang, An Yan, Miguel Eckstein, and William Yang Wang. Imagine: An imagination-based automatic evaluation metric for natural language generation. *arXiv preprint arXiv:2106.05970*, 2021.