Comparative Opinion Quintuple Extraction from Product Reviews

Ziheng Liu* **Rui Xia*†** **Jianfei Yu**
School of Computer Science and Engineering,
Nanjing University of Science and Technology, China
{zhliu, rxia, jfyu}@njust.edu.cn

Abstract

As an important task in opinion mining, comparative opinion mining aims to identify comparative sentences from product reviews, extract the comparative elements, and obtain the corresponding comparative opinion tuples. However, most previous studies simply regarded comparative tuple extraction as comparative element extraction, which ignored the fact that many comparative sentences may contain multiple comparisons. The comparative opinion tuples defined in these studies also failed to explicitly provide comparative preferences. To address these issues, in this work we first introduce a new Comparative Opinion Quintuple Extraction (COQE) task, to identify comparative sentences from product reviews and extract all comparative opinion quintuples (Subject, Object, Comparative Aspect, Comparative Opinion, Comparative Preference). Secondly, based on the existing comparative opinion mining corpora, we make supplementary annotations and construct three datasets for the COQE task. Finally, we benchmark the COQE task by proposing a new multi-stage neural network approach which significantly outperforms the baseline systems extended from previous comparative opinion mining methods. The datasets and source code are publicly released at https://github.com/NUSTM/COQE.

1 Introduction

Fine-grained opinion mining from product reviews has received considerable attention in the last decade. As around 10% of product reviews contain at least one comparison (Kessler and Kuhn, 2013), it is therefore crucial to extract and analyze these comparative sentences to detect public opinions towards the compared entities and aspects.

As the pioneering work for this direction, Jindal and Liu (2006b) proposed the Comparative Sentence Mining (CSM) task which first identifies comparative sentences from reviews, and extracts pre-defined comparative quintuples, i.e., (Subject, Object, Comparative Aspect, Relation Word, Comparison Type), from the identified comparative sentences. For example, given a sentence "*G6 has a worse zoom than G7, but G6's battery was more reliable than G7*", one comparative quintuple is (*G6*, *G7*, *battery*, *more*, Non-Equal Gradable). In their work they assumed that a comparative sentence contains only one comparative relation, and treated the comparative quintuple extraction task as a comparative element extraction (CEE) problem.

However, in real scenarios, a large number of comparative sentences contain more than one comparative relation. For example, 17.6% of the comparative sentences in the camera domain (Kessler and Kuhn, 2014) have at least two comparative relations. In this situation, simply applying CEE cannot extract comparative quintuples effectively. Moreover, in their definition, the relation word sometimes fails to explicitly reflect the comparative preference. For example, *"more"* in the above sentence is ambiguous, since it may refer to *"more reliable"* or *"more expensive"*.

Some recent studies (Panchenko et al., 2019; Ma et al., 2020) proposed a new task named Comparative Preference Classification (CPC), to identify the explicit comparative preferences (e.g., Better, Worse, None) between the subject entity and the object entity. However, the CPC task requires that the subject and object entities have been annotated, which largely hinders its applications in real scenarios.

To address the limitations of CEE and CPC, we introduce a new Comparative Opinion Quintuple Extraction (COQE) task, with the emphasis on the identification of comparative sentences, and the

*Equal contribution.
†Corresponding author.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3955–3965
November 7–11, 2021. ©2021 Association for Computational Linguistics

	G6 has a worse zoom than G7, but G6's battery was more reliable than G7.
Elements	subject: *G6*
	object: *G7*
	comparative aspect: {*zoom*, *battery*}
	comparative opinion: {*worse*, *more reliable*}
COQE	{(*G6*, *G7*, *zoom*, *worse*, Worse), (*G6*, *G7*, *battery*, *more reliable*, Better)}

Table 1: An example of the Comparative Opinion Quintuple Extraction (COQE) task.

extraction of all the comparative opinion quintuples from the comparative sentence. We define the comparative opinion quintuple as (sub, obj, ca, co, cp), where sub, obj, ca, co and cp refer to Subject, Object, Comparative Aspect, Comparative Opinion and Comparative Preference, respectively. Based on Subject, Object and Comparative Aspect which were defined in previous work, we further define Comparative Opinion as an opinion expression, in terms of a continuous textual span. It is similar to the relation word defined in (Jindal and Liu, 2006b) but including more necessary context, e.g., adjectives/adverbs after the relation word "*more*" or "*less*" and the negations. We also include Comparative Preference as a part of the comparative quintuple, and jointly extract the comparative elements and classify the comparative preference. As shown in Table 1, the output of COQE contains a set of two comparative opinion quintuples: {(*G6*, *G7*, *battery*, *more reliable*, Better), (*G6*, *G7*, *zoom*, *worse*, Worse)}.

Secondly, we construct three datasets for this COQE task based on three existing comparative opinion mining corpora. On the basis of the camera-domain corpus proposed by Kessler and Kuhn (2014), we further annotate the comparative opinion and preference for each comparative sentence. We also add the comparative opinion annotation to the datasets from the car and electronic domains released by COAE 2012/2013 (Tan et al., 2013). In addition, we annotate all the valid comparative quintuples and provide the starting and end position of each element in the quintuples.

Finally, we benchmark the task by proposing a new multi-stage neural network approach, including the stages of 1) Joint Comparative Sentence Identification and Comparative Elements Extraction, 2) Comparative Element Combination and Filtering, and 3) Comparative Preference Classification. The new approach significantly outperforms the baseline systems extended from traditional comparative opinion mining methods on three datasets.

The contributions of this work can be summarized as follows:

- We propose a new Comparative Opinion Quintuple Extraction (COQE) task, aiming to extract all the comparative quintuples from each review sentence.
- We construct three new datasets for the task, on the basis of the existing comparative opinion mining corpora.
- We benchmark the task by proposing a multi-stage neural network approach which significantly outperforms baseline systems extended from traditional methods.

2 Related Work

As a branch of aspect-based sentiment analysis, Comparative Sentence Mining was first proposed by Jindal and Liu (2006b) to first identify comparative sentences (CSI) from reviews, and extracts pre-defined comparative quintuples, i.e., (Subject, Object, Comparative Aspect, Relation Word, Comparison Type) from the identified comparative sentences. They assumed that a comparative sentence contains only one comparative relation, and regarded comparative quintuple extraction as a comparative element extraction (CEE) problem. This ignored the fact that a large percentage of comparative sentences contain more than one comparison.

For the CSI task, (Ganapathibhotla and Liu, 2008; Huang et al., 2008; Park and Blake, 2012) designed keyword-based or syntactic-based rules to identify comparative sentences in product reviews and scientific articles. (Jindal and Liu, 2006a,b; Huang et al., 2008; Liu et al., 2013) employed a Class Sequential Rule (CSR) method to mine sequence rules and use them as features of statistical classifiers.

For the CEE task, Jindal and Liu (2006b) and He et al. (2012) employed a Label Sequential Rule (LSR) method to extract comparative elements. (Hou and Li, 2008; Song et al., 2009; Huang et al., 2010; Wang et al., 2015a) extracted comparative elements based on conditional random field (CRF). Wang et al. (2010) and Kessler and Kuhn (2013) further employed semantic role labeling (SRL) to extract comparative elements. Arora et al. (2017) proposed a LSTM-CRF neural network to extract comparative elements.

In recent years, Panchenko et al. (2019) proposed a Comparative Preference Classification (CPC) task, to predict the preference (Better, Worse, None) between two annotated entities. Ma et al. (2020) further proposed a Graph Attention Network for this task. However, CPC requires to annotate two compared entities in advance, which greatly limits its application in real scenes.

In comparison, the COQE task proposed in this work focuses on identification of comparative sentences, and the extraction of all the comparative opinion quintuples from the comparative sentence, rather than comparative element extraction only. We support comparative quintuple extraction when a sentence contains multiple comparisons. Secondly, we re-define the comparative quintuple by incorporating comparative preference, and jointly perform comparative tuple extraction and comparative preference classification. Finally, most of the previous models for comparative opinion mining were based on rule methods or traditional machine learning methods. We establish a multi-stage deep learning approach for our task and significantly improved the performance of both CEE and COQE.

It is also worth noting that some recent studies on opinion tuple extraction (Liao et al., 2016; Peng et al., 2020) and quadruple extraction (Cai et al., 2021) have been proposed in traditional aspect-based sentiment analysis. Our work can be viewed as their extension from absolute opinion mining to comparative opinion mining.

3 Task and Datasets

3.1 Task Definition

Given a product review sentence containing n words $X = [x_1, \cdots, x_n]$, the goal of COQE is to first identify whether it is a comparative sentences, and then extract the set of quintuples (sub, obj, ca, co, cp) if it is a comparative sentence as follows:

$$S_{COQE} = \{\cdots, (sub, obj, ca, co, cp)_i, \cdots\}, \quad (1)$$

where sub and obj refer to the subject and object entities being compared, ca denotes the comparative aspect (i.e., feature attribute) of the entities, co denotes the comparative opinion which is an opinion expression indicating the comparative preference between two entities, and $cp \in \{$Worse, Equal, Better, Different$\}$ is the comparative preference denoting whether sub is worse than, equal to, better than, or different from obj.

Note that the first four elements of the comparative opinion quintuple need to be extracted from the sentence, while the fifth element needs to be classified from the pre-defined categories. Therefore, COQE is a challenging task that involves extracting four elements, classifying one element, and combining all the five elements into valid quintuples.

3.2 Dataset Construction

In addition to the comparative sentence mining corpus proposed by Jindal and Liu (2006b), Kessler et al. (2010) proposed a JDPA corpus, which consists of blog posts about cameras and cars where the camera domain contains 506 comparative sentences, and the car domain contains 1100 comparative sentences. However, the corpus only reflects the comparative elements and can not capture the comparative relation.

Kessler and Kuhn (2014) proposed a camera domain corpus containing 1707 comparative sentences, which explicitly annotated the comparative quintuple and supported the case where a sentence contains multiple comparative relations. They defined the quintuple as (Subject, Object, Aspect, Scale, Predicate), where Predicate is the syntactic marker that introduces a comparison (e.g., "better", "more") and Scale, a modified adjective/adverb, is added when predicate do not by themselves fully describe a comparison (e.g, "reliable" after "more"). The joint annotation Scale and Predicate can solve the shortcoming of Relation Word in (Jindal and Liu, 2006b), but it did not contain some necessary context that describes a comparative relation, e,g., negation and contrast.

The Chinese Opinion Analysis Evaluation (COAE) 2012/2013 (Tan et al., 2013) provided two Chinese comparative sentence mining corpora, in the domains of Car and Electronics, respectively. They annotated the comparative relation as a pair of triples, i.e., (subject, aspect, absolute sentiment) and (object, aspect, absolute sentiment).

We construct three datasets for our COQE task, on the basis of the above corpora.

- **Camera-COQE**: On basis of the Camera domain corpus released by Kessler and Kuhn (2014), we completed the annotation of Comparative Opinion and Comparative Preference for 1705 comparative sentences, and introducing 1599 non-comparative sentences.

		Car-COQE	Ele-COQE	Camera-COQE
#Sentence	#Comparative	1747	1800	1705
	#Non-Comparative	1800	1800	1599
	#Multi-Comparisons	550	361	500
	#Comparison Per Sent	1.5	1.3	1.4
	Percentage	31.5%	20.1%	29.3%
#Element	Subject Entity	1520	950	1649
	Object Entity	2121	1980	1316
	Comparative Aspect	1917	1602	1368
	Comparative Opinion	2171	2089	2163
	Comparative Preference	2695	2289	2442

Table 2: Statistics of three comparative quintuple corpora.

- **Car-COQE**: Based on the COAE 2012/2013 (Tan et al., 2013) Car domain corpus, we supplemented with the annotation of Comparative Opinion and Comparative Preference.
- **Ele-COQE**: Similar to Car-COQE, we construct the Ele-COQE dataset based on the COAE 2012/2013 Electronics (Ele) domain corpus.

Table 2 displays basic statistics of three datasets, where #Comparative, #Non-Comparative and #Multi-Comparisons denote the number of comparative sentences, non-comparative sentences and comparative sentences with multiple comparative quintuples. #Comparison Per Sent denotes the average number of comparative quintuples per sentence and Percentage denotes the percentage of sentences with multiple comparative quintuples among all the comparative sentences. As we can see, at least 20% of the comparative sentences in each dataset contain multiple comparative opinion quintuples.

4 Approach

As stated in the task definition, COQE is a challenging task that includes four-element extraction, one-element classification, and five-element combinations. To tackle the task, we propose a multi-stage neural network framework, in which the first stage is to identify comparative sentences and extract comparative elements, the second stage is to combine and filter the extracted four comparative elements (sub, obj, ca, co) to obtain valid comparative quadruples, and the third stage is to further classify each comparative quadruple into a pre-defined preference category, and obtain all the comparative opinion quintuples.

For the sentence in Table 1, in the first stage, we identify it as a comparative sentence and get the set of four comparative elements: $S_{sub} = \{G6\}$, $S_{obj} = \{G7\}$, $S_{ca} = \{zoom, battery\}$ and $S_{co} = \{worse, more\ reliable\}$. In the second stage, we combine the four elements extracted in the first stage with Cartesian product to form a candidate set of comparative quadruples, i.e., (*G6*, *G7*, *zoom*, *worse*), (*G6*, *G7*, *zoom*, *more reliable*), (*G6*, *G7*, *battery*, *worse*), (*G6*, *G7*, *battery*, *more reliable*). Furthermore, we train a classifier to filter invalid combinations to get two valid comparative quadruples, i.e., (*G6*, *G7*, *zoom*, *worse*), (*G6*, *G7*, *battery*, *more reliable*). Finally, in the third stage, the two comparative quadruples are classified into the corresponding comparative preference category to obtain two valid comparative quintuples as shown in Table 1.

4.1 Stage 1: Joint Comparative Sentence Identification and Comparative Elements Extraction

In the first stage, we proposed a multi-task learning framework based on BERT to identify comparative sentences and extract comparative elements simultaneously. Specifically, given an input sentence $X = [x_1, \cdots, x_n]$, we first insert two special tokens (i.e., CLS and SEP) at the beginning and the end respectively, and then feed the transformed sentence to BERT to obtain the hidden representations in the last layer:

$$h = [h_{[CLS]}, h_1, \cdots, h_n, h_{[SEP]}]. \quad (2)$$

Comparative Sentence Identification. First, we feed $h_{[\mathrm{CLS}]}$ to a softmax layer to predict whether the input sentence X is a comparative sentence:

$$y^c = \mathrm{softmax}(W^c h_{[CLS]} + b^c), \quad (3)$$

where W^c and b^c are weight matrices to learn, and $y^c \in \{0, 1\}$.

Comparative Element Extraction. For the identified comparative sentences, we further adopt four separate linear transformation functions and CRF layers to extract the four elements sub, obj, ca, co, respectively:

$$y^e = \mathrm{CRF}^e(h_1^e, \cdots, h_n^e), \quad (4)$$

where $h^e = W^e h + b^e$ and the Begin-Middle-End-Single-Outside (BMESO) tagging schema is

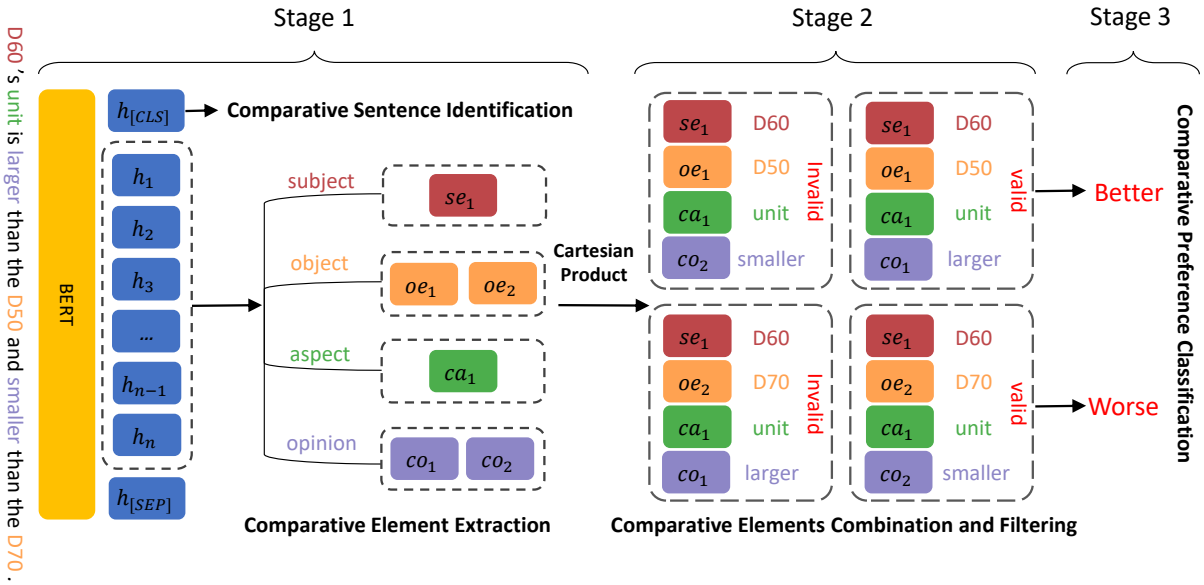

Figure 1: The architecture of the BERT-based Multi-Stage Neural Network Approach for COQE.

adopted for sequence labeling, and e refers to sub, obj, ca, co respectively. It should be noted that we adopt separate output layers for extracting each element in order to solve the problem of overlapping and nesting entities.

During the training stage, Comparative Sentence Identification and Comparative Elements Extraction are optimized simultaneously based on a multi-task learning framework. The final loss of the first stage is a weighted sum of L_{csi} and L_{cee_i}:

$$L = \lambda_c L_{csi} + \lambda_e \sum_i L_{cee_i}. \quad (5)$$

where L_{csi} is the cross-entropy loss for comparative sentence identification, and L_{cee_i} is the CRF loss for each individual element extraction. The two hyperparameters λ_e and λ_c are set to be 1 in our experiments.

4.2 Stage 2: Comparative Elements Combination and Filtering

In the first stage, we have obtained four sets of comparative elements for comparative sentences, denoted by $S_{sub} = \{sub_1, ..., sub_k\}$, $S_{obj} = \{obj_1, ..., obj_l\}$, $S_{ca} = \{ca_1, ..., ca_p\}$, $S_{co} = \{co_1, ..., co_q\}$.

With the four element sets, we perform Cartesian product over them to obtain a set of all possible comparative quadruple candidates:

$$S_{quad} = \{(sub_1, obj_1, ca_1, co_1), \cdots, (sub_k, obj_l, ca_p, co_q)\}. \quad (6)$$

For each quadruple, we obtain the representation of each element by concatenating its hidden representations for comparative sentence identification and comparative element extraction in Eqn. (2) and Eqn. (4) as follows:

$$r^e = [\text{avg}(h^e_{[start:end]}); \text{avg}(h_{[start:end]})], \quad (7)$$

where $start$ and end denote each element's start and end indices in the sentence, and avg denotes the average pooling operation. We then concatenate the representations of the four elements as the representation of each quadruple below:

$$\mathbf{r} = [r^{sub}; r^{obj}; r^{ca}; r^{co}]. \quad (8)$$

Finally, we stack a softmax layer on top as a quadruple filter to detect the validity of a quadruple:

$$y^{quad} = \text{softmax}(W^q \mathbf{r} + b^q), \quad (9)$$

where $y^{quad} \in \{0, 1\}$ indicates whether the input quadruple is valid or not.

During training, we employ a class-weighted cross entropy loss to address the data imbalance issue between valid and invalid quadruples as follows:

$$L_{quad} = \lambda L^{invalid}_{quad} + L^{valid}_{quad}, \quad (10)$$

where λ is the trade-off hyperparameter, and set to be 0.4 in our experiments.

4.3 Stage 3: Comparative Preference Classification

After obtaining all the valid comparative quadruples in the second stage, we then classify each

quadruple into a pre-defined comparative preference category in the third stage. Specifically, we add another softmax layer over the representation of each quadruple in Eqn. (8) for comparative preference classification below:

$$y^s = \text{softmax}(W^s\mathbf{r} + b^s), \tag{11}$$

where $y^s \in \{\text{Worse}, \text{Equal}, \text{Better}, \text{Different}\}$. During training, the standard cross-entropy loss is used for optimizing the parameters of the comparative preference classifier.

Finally, we combine the comparative preference prediction with the valid quadruples predicted in the second stage to get the final comparative opinion quintuples.

5 Experiments

5.1 Experimental Settings

We evaluate the performance of the multi-stage neural network approach on three COQE datasets. For comparison, we also develop two baseline systems extended from the representative methods in the previous comparative opinion mining task. We divide each dataset into a training set, a validation set and a testing set, with the proportion of 64%, 16% and 20%, respectively.

In Stage 1 of our BERT-based multi-stage approach Multi-$\text{Stage}_{\text{BERT}}$, we adopt $\text{BERT}_{\text{base}}$ for the English Camera dataset, and adopt a Chinese Version of BERT (BERT-Chinese) in the Chinese Car and Ele datasets. During training for all three stages, we use the Adam optimizer and set the batch size to 16 and the dropout to 0.1. The learning rates for Stages 1, 2 and 3 are set to be 2e-5, 5e-4 and 5e-4, respectively.

5.2 Evaluation Metrics

As Comparative Sentence Identification (CSI) and Comparative Element Extraction (CEE) are subsets of our approach, we evaluate the performance on CSI, CEE and COQE respectively. For CSI, we use the accuracy as the evaluation metric. For CEE, following (Marasović and Frank, 2018; Zhang et al., 2019, 2020), we calculate Precision, Recall and F_1 metrics for each element, and their Micro-average F_1. For COQE, we calculate Precision, Recall and F_1 for the whole quintuple.

The calculation of Precision, Recall, and F_1 score are as follows:

$$Precision = \frac{\#correct}{\#predict}, \tag{12}$$

$$Recall = \frac{\#correct}{\#gold}, \tag{13}$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}, \tag{14}$$

where $\#predict$ denotes the number of comparative element (or quintuple for COQE) predicted by the model, $\#gold$ denotes the number of comparative element (or quintuple) in the dataset, $\#correct$ denotes the number of correct comparative quintuple (or quintuple) in the predictions.

Meanwhile, we consider three matching strategies for measuring the correct predictions: Exact Match, Proportional Match, and Binary match respectively.

At first, ensure that the predicted quintuple's comparative preference is the same as the golden one, then define $\#correct_e$, $\#correct_p$ and $\#correct_b$ for Exact Match, Proportional Match, and Binary Match as follows:

$$\#correct_e = \begin{cases} 0 & \exists\,(g_k \neq p_k)\,; \\ 1 & \text{otherwise}, \end{cases} \tag{15}$$

where g_k denotes k-th element in the gold comparative quintuple, p_k denotes k-th element in the predicted comparative quintuple. It means that if all p_k and g_k match exactly ($k = 1, 2, 3, 4$), the count is 1, otherwise 0.

$$\#correct_p = \begin{cases} 0 & \exists(g_k \cap p_k = \varnothing); \\ \frac{\sum_k len(g_k \cap p_k)}{\sum_k len(g_k)} & \text{otherwise}, \end{cases} \tag{16}$$

where $len(\cdot)$ denotes the length of the comparative element. If all p_k and g_k have overlaps, the count is $\frac{\sum_k len(g_k \cap p_k)}{\sum_k len(g_k)}$, otherwise 0.

$$\#correct_b = \begin{cases} 0 & \exists(g_k \cap p_k = \varnothing); \\ 1 & \text{otherwise}. \end{cases} \tag{17}$$

where the count is 1 if all p_k and g_k have overlaps, otherwise 0.

5.3 Baseline Systems

In addition to Multi-$\text{Stage}_{\text{BERT}}$, we also established the following baseline systems:

- **CSI_{CSR}-CEE_{CRF}**: In Stage 1, we use a SVM with CSR (Jindal and Liu, 2006a) features to

Dataset	Method	CSI	CEE					COQE
			Subject	Object	Aspect	Opinion	Micro-Ave	Quintuple
Camera-COQE	CSI_{CSR}-CEE_{CRF}	65.38	30.66	41.48	24.16	53.45	40.04	3.46
	$(\text{CSI-CEE})_{\text{CRF}}$	82.14	34.65	40.52	32.96	60.10	45.12	4.88
	$\text{Multi-Stage}_{\text{LSTM}}$	87.14	48.72	48.29	44.27	54.10	49.58	9.05
	$\text{Multi-Stage}_{\text{BERT}}$	**93.04**	**58.15**	**60.00**	**59.11**	**65.61**	**61.21**	**13.36**
Car-COQE	CSI_{CSR}-CEE_{CRF}	86.90	25.83	48.46	43.10	54.27	45.05	5.19
	$(\text{CSI-CEE})_{\text{CRF}}$	89.85	38.44	56.65	48.16	56.50	51.33	8.65
	$\text{Multi-Stage}_{\text{LSTM}}$	92.68	52.93	69.04	54.71	63.94	60.99	10.28
	$\text{Multi-Stage}_{\text{BERT}}$	**97.39**	**73.51**	**84.16**	**76.99**	**81.03**	**79.50**	**29.75**
Ele-COQE	CSI_{CSR}-CEE_{CRF}	88.30	23.40	46.72	39.76	51.46	42.48	4.07
	$(\text{CSI-CEE})_{\text{CRF}}$	85.97	14.81	41.73	38.84	53.96	42.27	4.71
	$\text{Multi-Stage}_{\text{LSTM}}$	96.25	38.12	62.37	61.96	68.22	60.90	14.90
	$\text{Multi-Stage}_{\text{BERT}}$	**98.31**	**65.62**	**75.16**	**78.86**	**84.67**	**77.78**	**30.73**

Table 3: Results of different approaches for CSI, CEE and COQE under the Exact Match metric.

Dataset	Metric	CSI	CEE	COQE
Camera-COQE	Exact	93.04	61.21	13.36
	Prop	93.04	71.64	23.26
	Binary	93.04	76.23	25.25
Car-COQE	Exact	97.39	79.50	29.75
	Prop	97.39	85.19	38.46
	Binary	97.39	87.32	39.62
Ele-COQE	Exact	98.31	77.78	30.73
	Prop	98.31	84.59	40.83
	Binary	98.31	86.43	41.87

Table 4: Results of our $\text{Multi-Stage}_{\text{BERT}}$ approach under three kinds of matching metrics. For CEE, we report the micro average F_1 score.

identify comparative sentences and a CRF with standard lexical features to extract comparative elements. Stages 2 and 3 are similar as $\text{Multi-Stage}_{\text{BERT}}$.

- $\textbf{(CSI-CEE)}_{\textbf{CRF}}$: In this approach, we employ a feature-enhanced CRF (Wang et al., 2015b) for joint comparative sentence identification and comparative element extraction, where an all-"O" labeling sequence indicates the identification of non-comparative sentence.
- $\textbf{Multi-Stage}_{\textbf{LSTM}}$: This is a variant of $\text{Multi-Stage}_{\text{BERT}}$, where we replace the text encoder from BERT to LSTM.

5.4 Main Result

In Table 3, we report the performance of all four approaches on three tasks across three datasets. For CSI, we report accuracy. For CEE, we report the F_1 score for each element (Subject, Object, Comparative Aspect, Comparative Opinion) and their Micro-average (Micro). For COQE, we report the F_1 score for the quintuple. All results are reported under exact match.

It can be seen that across different tasks and datsets, CSI_{CSR}-CEE_{CRF} yields generally the lowest performance. CSI_{CSR}-CEE_{CRF} is slightly better. But their overall performances are relatively low, especially when dealing with complex tasks such as COQE (lower than 10%). Two deep learning approaches achieve much better performance in all three tasks. The BERT-based Multi-stage approach shows significant priority over LSTM-based one, due to its strong representation and generalization ability.

Among three tasks, the CSI task is the easiest, where almost all methods obtain satisfactory accuracy. The performances of different approaches for CEE are also okay, but the gap between different approaches increases. The COQE task is the most difficult. The traditional machine learning methods generate very poor performance. Even $\text{Multi-Stage}_{\text{LSTM}}$ fails to achieve satisfactory results. It is reasonable as the exact match of all five elements in a quintuple is very challenging.

By contrast, $\text{Multi-Stage}_{\text{BERT}}$ shows strong ability and greatly improves the performance of COQE, especially on Car-COQE and Ele-COQE, even though the task is very difficult.

In Table 4, we also report the performance of $\text{Multi-Stage}_{\text{BERT}}$ under three kinds of matching metrics, and that of different approaches in Table A1 and Table A2. It can be observed under Proportional Match and Binary Match, the performances of all models will have significant improvements.

Method	Camera-COQE			Car-COQE			Ele-COQE		
	CSI	CEE	COQE	CSI	CEE	COQE	CSI	CEE	COQE
Only CSI Loss	91.53	\	\	**97.75**	\	\	97.64	\	\
Only CEE Loss	\	60.23	12.65	\	78.91	28.32	\	76.80	29.17
Multi-Task Learning	**93.04**	**61.21**	**13.36**	97.39	**79.50**	**29.75**	**98.31**	**77.78**	**30.73**

Table 5: Ablation study to analyze the importance of different loss strategies in Stage 1.

	Method	None	Filter	keep rate
Camera -COQE	CSI_{CSR}-CEE_{CRF}	2.98	3.46	88.79
	$(\text{CSI-CEE})_{\text{CRF}}$	4.60	4.88	85.93
	$\text{Multi-Stage}_{\text{LSTM}}$	6.77	9.05	84.08
	$\text{Multi-Stage}_{\text{BERT}}$	11.17	13.36	60.66
Car -COQE	CSI_{CSR}-CEE_{CRF}	4.78	5.19	70.03
	$(\text{CSI-CEE})_{\text{CRF}}$	8.17	8.65	79.07
	$\text{Multi-Stage}_{\text{LSTM}}$	6.89	10.28	42.92
	$\text{Multi-Stage}_{\text{BERT}}$	23.39	29.75	61.47
Ele -COQE	CSI_{CSR}-CEE_{CRF}	4.07	4.05	80.55
	$(\text{CSI-CEE})_{\text{CRF}}$	4.47	4.71	80.88
	$\text{Multi-Stage}_{\text{LSTM}}$	14.07	14.90	84.08
	$\text{Multi-Stage}_{\text{BERT}}$	27.92	30.73	87.77

Table 6: Results of different approaches on the COQE task with or without the filter in Stage 2.

5.5 In-depth Analysis

Effects of Different Loss Strategies in Stage 1. In Table 5, we conduct ablation study of the multi-learning framework in Stage 1, by comparing only CSI loss, only CEE loss and multi-task learning. It can be seen that, CSI performance can be increased by adding the CEE loss, and the CEE performance can be also increased by adding the CSI loss. It suggests that the CSI and CEE tasks are mutually indicative. It is therefore reasonable for us to employ a multi-task learning of SCI and CEE in Stage 1.

Effects of Comparative Quadruple Filtering in Stage 2. In Table 6, we investigate the effects of comparative quadruple filtering in Stage 2, by comparing the COQE F_1 score of different approaches with or without Filtering, denoted by Filter and None respectively. The keep rate indicates the percentages of valid quadruple in all possible ones. It can be observed in Table 6 that after filtering, the COQE extraction performance increases significantly across all approaches and datasets.

Effects of Different Comparative Elements Representation in Stage 3. To investigate the impact of using different comparative element representations, we compare the results of using following comparative element representations:

	Camera -COQE	Car -COQE	Ele -COQE
BERT Embedding	12.30	29.40	**32.05**
High-layer Embedding	9.96	17.41	21.86
Concatenation	**13.36**	**29.75**	30.73

Table 7: The effect of different comparative element representation in Stage 3.

- **BERT Embedding**: Only the current element's BERT embedding is used in Eqn. (7): $r^e = \text{avg}(h_{[start:end]})$.
- **High-layer Embedding**: Only the current element's high-layer representation is used in Eqn. (7): $r^e = \text{avg}(h^e_{[start:end]})$.
- **Concatenation**: The concatenation of the two representations is used, as defined in Eqn. (7).

Based on the results in Table 7, we can clearly observe that the performance of only using Element Feature as the comparative element representation is rather limited, and concatenating the Element Feature and BERT Embedding achieves significant higher performance. This demonstrates that the two kinds of features can generally complement each other. Therefore, we use their concatenation as the comparative element representation in our experiments.

5.6 Case Study

To validate the effectiveness of our task, we compare our task with the CSM task proposed by Jindal and Liu (2006b) and the CPC task proposed by Panchenko et al. (2019). The output of three tasks on two examples are shown in Table 8.

Comparing the outputs on the first example, we can clearly see that the comparative quintuple defined in our COQE task exactly paraphrases the input sentence. In contrast, the quintuple defined in the CSM task is not a paraphrase of the input sentence, since it is hard to judge whether "G6" or "G7" is preferred by the user. Moreover, unlike the CPC task that requires providing the entity pairs *G6* and *G7*, our task aims to jointly perform entity pair extraction and preference classification.

Example 1		*The G6 's battery was more powerful than the G7 's battery.*	**CSI**	**CEE**	**CPC**	**COQE**
	CSM (Jindal and Liu, 2006b)	{(*G6*, *G7*, *battery*, *more*, Non-Equal Gradable)}	✓	✓	✗	✓
	CPC (Panchenko et al., 2019)	(*G6*, *G7*) ⇒ Better	✗	✗	✓	✗
	COQE	{(*G6*, *G7*, *battery*, *more powerful*, Better)}	✓	✓	✓	✓
Example 2		*The D200 autofocus performs similarly to the D80 but a stronger autofocus motor on the D200.*	**CSI**	**CEE**	**CPC**	**COQE**
	CSM (Jindal and Liu, 2006b)	{*D200*, *D80*, *autofocus performs*, *similarly*, Equative}	✓	✓	✗	✗
	CPC (Panchenko et al., 2019)	(*D80*, *D200*) ⇒ Worse	✗	✗	✓	✗
	COQE	{(*D200*, *D80*, *autofocus performs*, *similarly*, Equal), (*D200*, *D80*, *autofocus motor*, *stronger*, Better)}	✓	✓	✓	✓

Table 8: Case study of three comparative opinion mining tasks.

Source → Target	**Metric**	**CSI**	**CEE**					**COQE**
			Subject	Object	Aspect	Opinion	Micro-Ave	Quintuple
Ele → Car	**Exact**	96.62	64.75	79.60	67.96	78.90	74.08	23.44
	Prop	96.62	70.29	87.24	74.10	85.92	80.82	30.49
	Binary	96.62	72.80	89.46	76.01	88.44	83.11	31.64
Car → Ele	**Exact**	97.22	48.42	72.12	73.20	83.23	73.35	24.42
	Prop	97.22	53.61	84.34	76.46	87.48	79.26	31.61
	Binary	97.22	56.14	87.10	76.98	90.82	82.25	32.85

Table 9: Results of our proposed $\text{Multi-Stage}_{\text{BERT}}$ approach in the cross-domain setting.

Compared with the CSM task on the second example, our task is more suitable for comparative sentences with multiple comparative quintuples. Furthermore, compared with the CPC task, our task incorporates two additional preference categories, i.e., Equal and Different, which can cover a wider range of comparative entities.

5.7 Cross-Domain Experiments

In addition to the previous in-domain experiments, we further conducted a cross-domain experiment on two Chinese datasets, where the training set and validation set are chosen from the source domain, and testing set is in the target domain. The results are reported in Table 9. We use Source→Target to denote different cross-domain tasks, e.g., in Ele→Car Ele is the source domain and Car is the target domain.

It can be observed that there is a significant performance drop in the extraction of the subject, object, and aspect in comparison with the in-domain results in Table 3. This is reasonable since most entities and aspects in the source and target domains are quite different. In contrast, an interesting observation is that the comparative opinion extraction performance drops slightly in comparison with the in-domain setting, probably due to that the gap of comparative opinions in different domains is relatively small. As a whole, the quintuple extraction performance has a significant drop.

It can also be found that the drop of comparative sentence identification is very limited. This suggest that the patterns of distinguishing comparison in different domains are similar.

6 Conclusions and Future Work

In this work, we introduce a new Comparative Opinion Quintuple Extraction (COQE) task, to identify comparative sentences from reviews, and extract all comparative opinion quintuples each of which includes Subject, Object, Comparative Aspect, Comparative Opinion and Comparative Preference. We construct three datasets for the task, and benchmark the task by proposing a new multi-stage neural network approach which shows significant advantages in comparison with baseline systems extended from previous methods. In the future work, we would like to consider more sophisticated approaches, for example, end-to-end deep learning models, for COQE.

Acknowledgments

We would like to thank Prof. Suge Wang for providing the COAE Car and Ele datasets. This work was supported by the Natural Science Foundation of China (62076133 and 62006117), and the Natural Science Foundation of Jiangsu Province for Young Scholars (BK20200463) and Distinguished Young Scholars (BK20200018).

References

Jatin Arora, Sumit Agrawal, Pawan Goyal, and Sayan Pathak. 2017. Extracting entities of interest from comparative product reviews. In *Proceedings of ACM on Conference on Information and Knowledge Management(CIKM)*, pages 1975–1978.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 340–350.

Murthy Ganapathibhotla and Bing Liu. 2008. Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 241–248.

Shuo He, Fang Yuan, and Yu Wang. 2012. Extracting the comparative relations for mobile reviews. In *Proceedings of the 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet)*, pages 3247–3250.

Feng Hou and Guohui Li. 2008. Mining chinese comparative sentences by semantic role labeling. In *International Conference on Machine Learning and Cybernetics*, pages 2563–2568.

Gaohui Huang, Tianfang Yao, and Quansheng Liu. 2010. Mining chinese comparative sentences and relations based on crf algorithm. *Application Research of Computers*, pages 2061–2064.

Xiaojiang Huang, Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2008. Learning to identify comparative sentences in chinese text. In *Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence*, pages 187–198.

Nitin Jindal and Bing Liu. 2006a. Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 244–251.

Nitin Jindal and Bing Liu. 2006b. Mining comparative sentences and relations. In *Proceedings of the twenty AAAI Conference on Artificial Intelligence (AAAI)*, pages 1331–1336.

Jason S Kessler, Miriam Eckert, Lyndsie Clark, and Nicolas Nicolov. 2010. The icwsm 2010 jdpa sentiment corpus for the automotive domain. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge*.

Wiltrud Kessler and Jonas Kuhn. 2013. Detection of product comparisons-how far does an out-of-the-box semantic role labeling system take you? In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1892–1897.

Wiltrud Kessler and Jonas Kuhn. 2014. A corpus of comparisons in product reviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 2242–2248.

Jian Liao, Yang Li, and Suge Wang. 2016. The constitution of a fine-grained opinion annotated corpus on weibo. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 227–240.

Quanchao Liu, Heyan Huang, Chen Zhang, Zhenzhao Chen, and Jiajun Chen. 2013. Chinese comparative sentence identification based on the combination of rules and statistics. In *Part II of the Proceedings of the 9th International Conference on Advanced Data Mining and Applications*, pages 300–310.

Nianzu Ma, Sahisnu Mazumder, Hao Wang, and Bing Liu. 2020. Entity-aware dependency-based deep graph attention network for comparative preference classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5782–5788.

Ana Marasović and Anette Frank. 2018. Srl4orl: Improving opinion role labeling using multi-task learning with semantic role labeling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 583–594.

Alexander Panchenko, Alexander Bondarenko, Mirco Franzek, Matthias Hagen, and Chris Biemann. 2019. Categorizing comparative sentences. In *Proceedings of the 6th Workshop on Argument Mining*, pages 136–145.

Dae Hoon Park and Catherine Blake. 2012. Identifying comparative claim sentences in full-text scientific articles. In *Proceedings of the workshop on detecting structure in scholarly discourse*, pages 1–9.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 8600–8607.

Rui Song, Hongfei Lin, and Fuyang Chang. 2009. Chinese comparative sentences identification and comparative relations extraction. *Journal of Chinese Information Processing*, pages 102–107.

Songbo Tan, Kang Liu, Suge Wang, and Xiangwen Liao. 2013. Overview of chinese opinion analysis evaluation 2013.

Suge Wang, Hongxia Li, and Xiaolei Song. 2010. Automatic semantic role labeling for chinese comparative sentences based on hybrid patterns. In *International Conference on Artificial Intelligence and Computational Intelligence (IEEE)*, pages 378–382.

Wei Wang, TieJun Zhao, GuoDong Xin, and YongDong Xu. 2015a. Exploiting machine learning for comparative sentences extraction. *International Journal of Hybrid Information Technology*, pages 347–354.

Wei Wang, TieJun Zhao, GuoDong Xin, and YongDong Xu. 2015b. Extraction of comparative elements using conditional random fields. *Acta Automatica Sinica*, pages 1385–1393.

Bo Zhang, Yue Zhang, Rui Wang, Zhenghua Li, and Min Zhang. 2020. Syntax-aware opinion role labeling with dependency graph convolutional networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3249–3258.

Meishan Zhang, Peili Liang, and Guohong Fu. 2019. Enhancing opinion role labeling with semantic-aware word representations from semantic role labeling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 641–646.

A Experiment results under Proportional Match and Binary Match

In Table A1 and Table A2, we report the performance of all four approaches on three tasks across three datasets under the metrics of Proportional Match and Binary Match.

Dataset	Method	CSI	CEE					COQE
			Subject	Object	Aspect	Opinion	Micro	Quintuple
Camera-COQE	CSI_{CSR}-CEE_{CRF}	65.38	34.05	46.19	25.67	58.31	43.90	4.67
	$(\text{CSI-CEE})_{\text{CRF}}$	82.14	38.49	45.42	35.61	66.19	49.80	6.59
	$\text{Multi-Stage}_{\text{LSTM}}$	87.14	55.45	58.16	51.55	68.60	59.72	13.23
	$\text{Multi-Stage}_{\text{BERT}}$	**93.04**	**68.68**	**71.00**	**66.24**	**77.62**	**71.64**	**23.26**
Car-COQE	CSI_{CSR}-CEE_{CRF}	86.90	30.23	54.68	45.45	58.11	49.33	7.15
	$(\text{CSI-CEE})_{\text{CRF}}$	89.85	43.78	63.02	51.25	61.86	56.45	11.97
	$\text{Multi-Stage}_{\text{LSTM}}$	92.68	62.91	76.88	63.84	76.71	70.97	17.54
	$\text{Multi-Stage}_{\text{BERT}}$	**97.39**	**80.46**	**89.01**	**81.73**	**87.48**	**85.19**	**38.46**
Ele-COQE	CSI_{CSR}-CEE_{CRF}	88.30	27.13	52.42	42.06	53.79	46.04	8.16
	$(\text{CSI-CEE})_{\text{CRF}}$	85.97	15.74	55.61	39.08	60.90	49.15	9.01
	$\text{Multi-Stage}_{\text{LSTM}}$	96.25	48.01	80.01	68.71	79.70	72.85	25.33
	$\text{Multi-Stage}_{\text{BERT}}$	**98.31**	**70.09**	**87.90**	**81.30**	**90.30**	**84.59**	**40.83**

Table A1: The performance of different approaches under the Proportional Match metric.

Dataset	Method	CSI	CEE					COQE
			Subject	Object	Aspect	Opinion	Micro	Quintuple
Camera-COQE	CSI_{CSR}-CEE_{CRF}	65.38	36.00	48.40	26.58	63.36	46.84	5.08
	$(\text{CSI-CEE})_{\text{CRF}}$	82.14	40.50	47.89	37.43	72.77	53.60	7.32
	$\text{Multi-Stage}_{\text{LSTM}}$	87.14	59.29	62.39	55.34	75.61	64.73	14.67
	$\text{Multi-Stage}_{\text{BERT}}$	**93.04**	**71.07**	**74.91**	**69.14**	**85.36**	**76.23**	**25.25**
Car-COQE	CSI_{CSR}-CEE_{CRF}	86.90	32.04	57.36	47.25	61.06	51.73	7.61
	$(\text{CSI-CEE})_{\text{CRF}}$	89.85	45.76	64.87	52.35	65.45	58.65	12.63
	$\text{Multi-Stage}_{\text{LSTM}}$	96.25	48.01	80.01	68.71	79.70	72.85	18.76
	$\text{Multi-Stage}_{\text{BERT}}$	**97.89**	**82.45**	**90.50**	**83.18**	**90.91**	**87.32**	**39.62**
Ele-COQE	CSI_{CSR}-CEE_{CRF}	88.30	28.72	55.14	43.70	56.39	48.27	8.57
	$(\text{CSI-CEE})_{\text{CRF}}$	85.97	15.74	59.47	39.32	62.49	50.96	9.42
	$\text{Multi-Stage}_{\text{LSTM}}$	96.25	49.27	83.42	71.91	83.86	76.16	26.61
	$\text{Multi-Stage}_{\text{BERT}}$	**98.61**	**70.87**	**91.30**	**82.33**	**91.76**	**86.43**	**41.87**

Table A2: The performance of different approaches under the Binary Match metric.