A review for comparative text mining: From data acquisition to practical application

Na Wei
School of Management, Northwestern Polytechnical University, P.R. China

Songzheng Zhao
School of Management, Northwestern Polytechnical University, P.R. China

Jing Liu
School of Management Science and Engineering, Tianjin University of Finance and Economics, P.R. China

Shenghui Wang
Human Media Interaction (HMI), Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), University of Twente, The Netherlands

Abstract
Social media provides customers with great opportunities to share their opinions regarding certain products and services. Comparative text, as an important expression form, deserves further exploration, since it contains considerable comparative information between different products and services. In this study, we review existing research on Comparative Text Mining (CTM) in the past 16 years. Basic concepts related to CTM are first described, and a general research framework is subsequently proposed. We then dive into each component of the research framework, ranging from data acquisition, comparative text identification (CTI), comparative relation extraction (CRE), to potential applications. In addition, we conduct extensive experimental analysis on existing methods for CTI and CRE, and clarify their limitations. Accordingly, we provide corresponding insights, and point out future research directions.

Keywords
Comparative network; comparative relation extraction; comparative text identification; comparative text mining

1. Introduction
Following rapid growth of Web 2.0 technologies, there are various platforms for the public to share their ideas, experiences, and attitudes towards products or services. The flourishing of social media also provides corporations with opportunities for understanding advantages and disadvantages of their products from the perspectives of customers and users. Actually, exploring and mining the knowledge behind User Generated Content (UGC) can benefit government, corporations, and customers [1]. They can use UGC to analyse the real development status and trend of an industry, understand pros and cons of products, and make better purchasing decisions based on other users’ experiences, separately.

Opinion mining, as one of dominant research communities for utilising UGC, has attracted considerable attention from researchers and practitioners [2]. In opinion mining, researchers identify entities (e.g. products, names) in the text and analyse corresponding attitudes towards them [1]. However, it is insufficient to only analyse individual entities,
since comparison and competition between different entities play an important role in the real competitive world [3]. Therefore, comparison analysis is a promising research field for government, corporations, and customers [4].

Existing research on comparison analysis can be grouped into three types, that is, single opinion mining-based, novel data source-based, and comparative text mining (CTM)-based. Single opinion mining-based methods obtain reviews of different entities, who are compared based on sentimental scores derived based on each entity’s own reviews [5–7]. Intuitively, this kind of comparison is unreliable and inequitable [8], since reviews from people who have used more than one product are more credible than those who discuss one product [2]. For novel data source-based methods, comparison analysis is conducted on the basis of emerging data sources, such as question-answering contents [4,9], web search logs [10], and customer favourite data [11]. These models may be more accurate than those based on reviews from the public, but are infeasible in some scenarios due to the data’s inaccessibility [4]. Comparative text mining-based methods compare two or more entities co-occurred in comparative text [12]. It is shown that about 10% of massive online text contains rich and detailed comparison information [1], which presents the potential to conduct comparative analysis based on UGC.

So far, considerable effort has been devoted to conducting CTM. A review of CTM summarised previous studies on CTM from comparative text identification (CTI), comparative entity detection, comparative attribute detection, and comparative relation extraction (CRE) [1]. However, the prior review suffers from obvious limitations. First, it only covered the research before 2015 (2006–2015). New techniques and methods with the continuous effort on CTM (as shown in Figure 1) are not included in the prior review. Second, although potential application is the ultimate goal of CTM, the prior review ignored studies that focus on practical applications based on extracted comparative relations.

Therefore, this study systematically summarises existing, especially latest studies on CTM. Specifically, we search related papers from Web of Science with keywords ‘comparative sentence’, ‘comparative text’, ‘comparative opinion’, and ‘comparative relation’. Finally, 42 papers are included in this study. Furthermore, we conduct extensive empirical analysis with two real-world cases and clarify limitations of existing research. Accordingly, several promising directions for future research are pointed out. Different from the existing review [1], this study covers latest papers and newly developed models, especially deep learning-based methods. We also discuss the research status of potential applications that can benefit from mined comparison knowledge using CTM. Moreover, we provide a novel framework in which the original pipeline framework in [1] is transformed to the joint-learning manner in the CRE component.

Overall, contributions of this study are twofold:

1. We summarise studies regarding CTM in the past 16 years (2006–2021), analyse limitations of existing research, and point out some novel insights on future research directions.
2. We propose a novel CTM framework from data acquisition to practical applications. Especially, in the CRE component, several interactive subtasks are integrated into one, thereby alleviating error propagation, exploring interactions among subtasks, and thus enhancing generalisation capability of corresponding models.

The remainder of this study is organised as follows. Basic concepts and definitions, as well as the novel framework for CTM is described in Section 2. Each component of the framework is elaborated in detail in Sections 3–6. Section 7
conducts empirical analysis, summarises limitations of existing methods, and points out future research directions. Section 8 concludes this study.

2. Basic concepts and research framework

2.1. Comparative text

Comparative text refers to the text in which two or more similar entities are compared with regard to themselves or a specific attribute [12]. The entity could be a person, a place, a product, or a service. Comparative text can be classified into subjective comparison and objective comparison [13], which are, respectively, referred to as comparative opinion and comparative fact [1]. In existing research, the boundary between them is not very clear [1].

Generally, comparative text can be divided into finer granularities. In this study, we follow the commonly adopted standard, which divides comparative text into four types: non-equal gradable, equative, superlative, and non-gradable [13]. The former three types, which are widely studied in existing research, explicitly express the meaning of greater/less than, equal to, and the best/worst, respectively, while non-gradable implies implicit comparison. Table 1 provides several examples of different comparison expression types.

Table 1. Examples of different comparison expression types.

<table>
<thead>
<tr>
<th>Types</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparative opinion</td>
<td>Samsung is trendier than iPhone.</td>
</tr>
<tr>
<td>Comparative fact</td>
<td>The size of iPhone 8 is bigger than that of Samsung Note 8.</td>
</tr>
<tr>
<td>Non-equal gradable</td>
<td>The appearance of BMW is more beautiful than that of Mercedes Benz.</td>
</tr>
<tr>
<td>Equative</td>
<td>The cost performance of DeLong X3000 is the same as that of JieFang JH6.</td>
</tr>
<tr>
<td>Superlative</td>
<td>The cab of Yuejin C500 is the most successful.</td>
</tr>
<tr>
<td>Non-gradable</td>
<td>Toyota Camry has GPS, but Nissan SYLPHY made in 2016 does not have.</td>
</tr>
</tbody>
</table>

2.2. Framework for comparative text mining

Inspired by the definition of opinion mining1 in the prior study [1], we define CTM as a framework composed of a series of techniques and phases; the framework aims to discriminate comparative text from non-comparative text, extract comparative relations, and derive valuable information to support decision making.

We propose a comprehensive framework (as shown in Figure 2), which consists of four main components: data acquisition, CTI, CRE, and application. As compared with the prior work [1] that incorporated five sub-tasks (i.e. data acquisition and preprocessing, CTI, comparative entity recognition, comparative attribute recognition, and CRE), we integrate...
three interacted sub-tasks (i.e. comparative entity recognition, comparative attribute recognition, and CRE) into one united component (i.e. CRE in Figure 2) with a joint learning manner. For example, when the original pipeline framework in Varathan et al. [1] is applied on the text ‘Which is more cost-effective, x3000 or h6? They are similar’, falsely predict any entity (x3000 or h6) can lead to the miss of their relation due to error propagation. Our study can alleviate this issue via joint optimization of different sub-tasks. Moreover, we incorporate an application component into the framework due to its important role in exploring practical values of CTM.

3. Data acquisition

Explored data in research on CTM can be divided into two types: traditional online text data and specific data. The former one is widely used and can be obtained from discussions on forums [12,13], customer reviews [2,3,8,9,12–19], web, and blogs [2,12], in which customer reviews dominated. Investigated specific data includes question-answering contents [4,9], web search logs [10], and customer favourite data [11]. Text in question-answering contents and web search logs presents higher professionalism and less colloquialism [9], while customer favourite data are a set of entities; entities in the same list are simply considered as competitors.

Explored data in existing studies covers various industries and languages. With respect to industries, researchers focused on manufacturing and service areas, such as car [3,20–22], mobile phone [9,14,16,17,23], camera [8,18,20], MP3 player [16], running shoe [22], and restaurant [15,19,24]. In terms of languages, English and Chinese are dominant, while Japanese [2,25], Korean [15,26–28], Vietnamese [29], and Arabic [30] are also involved.

4. Comparative text identification (CTI)

The aim of CTI is to identify comparative text from non-comparative text (coarse-grained classification), and to further divide them into fine-grained types.

4.1. Coarse-grained classification

Coarse-grained classification is a binary classification task, that is, distinguishing comparative text (the positive class) from non-comparative text (the negative class). Existing coarse-grained classification methods can be categorised into two groups: rule-based and machine learning-based. Rule-based methods can be further classified into four types: co-occurrence rules, manually constructed rules, automatically constructed rules, and their combination. Co-occurrence rules treat a piece of text or sentence as positive when two different entities appear in the same context [17]. Manually constructed rules summarise syntax and semantic features of comparative text, based on which new comparative text is identified [15,23,27,31–33]. For example, simple patterns (such as ‘than’, ‘comparative + than’, ‘exceed’, ‘the same as’) often appear in comparative text; therefore, these words, phrases, combinations, or special syntactic structures can be selected manually as rules. In order to reduce manual workload, some methods for automatically constructed rules (e.g. Class Sequence Rule, CSR [9,34–36]) are proposed. In CSR, sequences are first constructed based on manually selected keywords, along with their part-of-speech (POS), and the POS of other words in text. These sequences are fed into mining algorithms (e.g. Prefix-Projected Pattern Growth, PrefixSpan [37]) to obtain candidate rules. Rules can be determined as long as the candidate rules meet certain minimum confidence and support criteria. Rules, which are generated using different methods, can be combined to construct a mixed rule pool and make the final identification [38].

With regard to machine learning-based methods, investigated classifier algorithms include Naive Bayes [39], support vector machine (SVM) [40,41], decision tree [21], random forest [42], logistic regression [3], and bagging [3]. Input features generally include CSR-generated rules [39] and manually designed language features [3,41,43], such as keywords, sentiment, and dependency structures. In recent years, some researchers have built CTI models based on deep learning methods [44], including the Bidirectional Encoder Representations from the Transformers (BERT) [45] and Long Short-Term Memory (LSTM), and so on.

4.2. Fine-grained classification

It is important to further classify comparative text into fine-grained types. For example, Ronen Feldman et al. [22] classified comparative text into similar and different. Xu et al. [14] grouped comparative text into three types: better, worse, and same. Dividing comparative text into non-equal gradable, equative, superlative, and non-gradable, dominated this line of research. For fine-grained classification, investigated approaches include machine learning-based methods [13,22].
In general, fine-grained classification is conducted following coarse-grained classification, rather than being performed directly, mainly due to the severe class imbalance issue [13].

4.3. Evaluation

The performance of coarse-grained classification is generally evaluated with the confusion matrix and corresponding classification metrics, including accuracy, precision, recall, and F-measure (F1). In addition, considering the imbalance issue in the coarse-grained classification [1], another metric AUC (Area Under Curve) [46], which is a metric invariant to class distribution, is also used.

\[ \text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \]  
\[ \text{Precision} = \frac{TP}{TP + FP} \]  
\[ \text{Recall} = \frac{TP}{TP + FN} \]  
\[ F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]

where TP, FP, TN, and FN represent the follows: TP – the number of true comparative text that are classified as comparative text; FP – the number of true non-comparative text that are classified as comparative text; TN – the number of true non-comparative text that are classified as non-comparative text; and FN – the number of true comparative text that are classified as non-comparative text.

The performance of fine-grained classification is generally evaluated with multi-class classification metrics, including micro, macro, and weighted average metrics.

\[ \text{Precision}_{\text{micro}} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FP_i)} \]  
\[ \text{Recall}_{\text{micro}} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FN_i)} \]  
\[ F1_{\text{micro}} = 2 \times \frac{\text{Precision}_{\text{micro}} \times \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}} \]  
\[ \text{Precision}_{\text{macro}} = \frac{1}{n} \sum_{i=1}^{n} \text{Precision}_i \]  
\[ \text{Recall}_{\text{macro}} = \frac{1}{n} \sum_{i=1}^{n} \text{Recall}_i \]  
\[ F1_{\text{macro}} = \frac{1}{n} \sum_{i=1}^{n} F1_i \]  
\[ \text{Precision}_{\text{weighted}} = \frac{1}{n} \sum_{i=1}^{n} w_i \text{Precision}_i \]  
\[ \text{Recall}_{\text{weighted}} = \frac{1}{n} \sum_{i=1}^{n} w_i \text{Recall}_i \]  
\[ F1_{\text{weighted}} = \frac{1}{n} \sum_{i=1}^{n} w_i F1_i \]

where \( n \) represents the number of classes, \( TP_i, FP_i, TN_i, FN_i, \text{Precision}_i, \text{Recall}_i, F1_i \) are values and metrics of the \( i \) th class, and \( w_i \) denotes the weight assigned to the \( i \) th class.

5. Comparative relation extraction

5.1. Comparative relation definition

Designing a comprehensive comparative relation tuple is the crucial foundation of mining comparative knowledge. However, an extremely complex relation tuple schema could impose many challenges in the CRE process. Therefore, designing a proper comparative relation is crucial. According to existing studies, the commonly used comparative relation schema is a four-tuple, that is, < comparative subject, comparative object, comparative aspect, comparative result...
For example, four elements are labelled in the following sentence: **The size (aspect) of iPhone 8 (subject) is bigger than (result) that of Samsung Note 8 (object).** Comparative subject and object, as the basis of comparative relations, are explored in almost all papers \cite{3,9,19,21,23}, while comparative aspect (or comparative feature), is an important component of comparative relation. However, it is recommended to incorporate comparative aspect as a part of comparative relation in comparative analysis of different products \cite{3}. Comparative result, which expresses customer preference towards different competitive products, could be grouped into two types: qualitative and quantitative. Qualitative result generally classifies result into four types \cite{14,17,23}: better, worse, similar, and no (i.e. non-existence of a comparative relation). While quantitative result represents a comparison relationship with a score \cite{3,19}. For example, Gao et al. \cite{19} used some discrete value to quantify the sentimental strength and polarity.

5.2. Comparative subject and object detection

Comparative subject and object are basic elements constituting a comparative relation. Existing comparative entity detection (without distinguishing between comparative subject and object) methods can be grouped into two types: lexicon-based \cite{3,8,9,16–18,23} and machine learning-based \cite{8,21,47}. Lexicon-based methods identify entities via matching words in text with the words in a pre-defined lexicon or calculating their similarity. In terms of machine learning-based methods, investigated algorithms include decision tree \cite{8} and commonly used conditional random field (CRF) \cite{21,47,48}. Explored features include word, position, POS, and semantic role labelling (SRL) corresponding to each word in text.

To distinguish between comparative subject and object, some research uses the relative position of entities to comparative keywords \cite{13}. In some situations, platforms provide information of the comparative subject \cite{23}. For example, reviews in a page on e-commerce platforms are often related to a specific product, which can be considered as the comparative subject.

5.3. Comparative aspect recognition

Existing comparative aspect recognition methods can be classified into three groups: lexicon-based \cite{9,17}, machine learning-based \cite{21,22,49}, and external knowledge-based \cite{3,8,16}. The lexicon-based and machine learning-based methods are similar with those in comparative subject and object detection. External knowledge-based method mainly uses the knowledge provided by the resource of reviews to identify aspects involved in a specific review. For example, a website can organise the comment area with different sections according to product aspects, and customers share their ideas of a specific aspect in the designated section.

5.4. Comparative result identification

Comparative result is used to express emotional tendency, which can be detected using sentiment lexicon-based \cite{2,3,9,12,15,23,50}, rule-based \cite{21}, and machine learning-based \cite{14,16,17,51,52} methods. The sentiment lexicon-based method \cite{19} analyses comparative results based on manually built sentiment lexicon with sentiment tools (e.g. SentiStrength \cite{18}). However, some researchers argued that beyond sentiment keywords, it is necessary to incorporate other elements such as predicate verbs (e.g. overwhelm and overcome), negative words (e.g. not and no), and adverb of degree (e.g. too and much). Therefore, some rules are proposed to detect some collocations and sentence patterns in rule-based methods \cite{22,23}. Machine learning–based methods generally use SVM and CRF to classify sentiment polarities or ranks \cite{16,17}.

5.5. Evaluation

Evaluation metrics for CRE generally include precision, recall, and F1 \cite{2,13,21,23,47}. Prior work evaluates the performance of CRE from three perspectives: separate evaluation on each element, micro-average and macro-average evaluation \cite{14,20}, and evaluation on the whole relation tuple. Intuitively, the former two evaluation strategies fail to evaluate the performance of CRE reasonably. For example, given three relations \( R_1 < A_1, B_1, C_1 >, R_2 < A_2, B_2, C_2 >, R_3 < A_3, B_3, C_3 > \); extracted relations are \( \bar{R}_1 < A_1, B_1, \bar{C}_1 >, \bar{R}_2 < A_2, \bar{B}_2, C_2 >, \bar{R}_3 < A_3, B_3, \bar{C}_3 > \) (the symbol with an overline represents that the identified element is different from the ground-truth). Although the accuracy of each element identification is 2/3, none of the relations are extracted correctly. Therefore, some work \cite{13} considers using the whole tuple as the unit to evaluate the performance of CRE; a correctly extracted relation means that all of elements in a relation are correctly identified.
In this study, we adopt the following evaluation strategies: uni-tuple, bi-tuple, triple, and whole-tuple. Specifically, uni-tuple-subject and uni-tuple-object respectively measures performance of identifying comparative subject and object; bi-tuple focuses on the simultaneously correct identification of both comparative subject and object; triple needs correct identification of three elements (i.e. comparative subject, object, and aspect); in whole-tuple strategy, all of elements are required to be correctly identified. Moreover, we adopt two different matching criteria: partial matching and exact matching. Partial matching means that the first token of an element is identified correctly; exact matching means an element is fully identified.

6. Application

6.1. Comparative network construction

Comparative tuples can be formulated as a graph. Nodes and edges in the graph represent entities and comparative relations between nodes (entities). The direction of an edge points to the preferred node (entity). For example, on the basis of a comparative relation \(< \text{iPhone 8, Samsung Note 8, shape, positive (negative)} >\) extracted from the sentence ‘The shape of iPhone 8 is better (worse) than that of Samsung Note 8’, the edge connecting the two entities points to iPhone 8 (Samsung Note 8), as shown in Figure 3.

According to the number and direction of edges between entities, some researchers classified the comparative network into three groups [19]: one-side graph, two-side graph, and multiple-side graph. One-side graph means there is only one edge between two nodes; the weight of the edge is obtained by calculating the number of relations or the total score of comparative results between two entities. In a two-side graph, there are two types of edges that represent opposite directions; the weight of each edge is calculated as the number of relations or the total score of comparative results in each edge type. In a multiple-side graph, each comparative relation is formulated as an edge. A comparative aspect, if exists in a comparative relation, can be regarded as the edge attribute. Unavailable comparative aspect represents that the two entities are compared as a whole. For example, given two sentences: Text-1: \(\text{the entity1 is better than entity2 in aspect1}\), and Text-2: \(\text{the entity1 is better than entity2}\), edge connecting the two entities in the graph derived from Text-1 has an attribute ‘aspect1’, while that in the graph derived from Text-2 has the attribute ‘whole’.

6.2. Evaluation and application of comparative network

Comparative network construction and application have obtained increasing attention. However, to the best of our knowledge, the commonly used evaluation method for comparative network is external knowledge-based [3,8,9,16,18,19]. Specifically, researchers generally evaluate the comparative network by analysing the consistency between sorting results of entities based on the constructed comparative network and those based on external knowledge [18], such as sales ranking and star ranking.

Applications of comparative network are the ultimate purpose of CTM, that is, discovering interesting and useful knowledge from comparative text. Sorting entities based on their centrality [2] or their degree [16,22] in a comparative network is a widely investigated application [3,8,9,18,19]. Other applications include market segmentation [19], identifying competing products [3,19], discovering competitive advantages and disadvantages of a specific product [3,17,19], analysing the upgrading of a product during different periods, and forecasting sales volume [18].

7. Experimental analysis and novel insights on future research directions

To provide an in-depth understanding of existing methods, we perform experimental analysis regarding both CTI and CRE with two real-world cases. Empirical discussions are subsequently conducted to analyse limitations of existing methods. Accordingly, we provide novel insights on future research directions about CTM.
7.1. Experimental datasets

We develop two real-world corpora ‘truck data’ and ‘mobile phone data’ for experimental analysis. The ‘truck data’ is constructed based on 118,275 posts collected from a famous forum the truck home (http://www.360che.com/) during 2010/02-2020/06. Drivers and truck enthusiasts can solicit advice on buying a truck and share their use experiences on the truck home. The ‘mobile phone data’ is built based on 132,696 reviews collected from a leading E-commerce platform JD (https://www.jd.com/) during 2017/02-2019/10. When conducting data cleaning, we delete personally identifiable information (e.g. user name and phone number), remove useless characters and duplicate text, and split long posts and reviews into sentences. Table 2 shows statistical information of the two used corpora in this study.

7.2. Experimental result and discussion

We compare several widely used CTI methods, including CSR and SVM. Moreover, we investigate several deep learning methods, including BERT, LSTM, and Convolution Neural Network (CNN). The AUC, precision, recall and F1 values are reported in Table 3. As shown in Table 3, deep learning–based methods obtain the best performance in most situations. On one hand, CSR and SVM cannot recognise implicit comparative text, since they depend on manually selected keywords. For example, although example #1 in Table 4 lacks a keyword, it implies comparative information. On the other hand, CSR and SVM generally predict non-comparative text including selected keywords or rules as comparative text [53], such as the example #2 in Table 4. While deep learning–based methods have the ability to correctly identify them. The main advantage of deep-learning methods lies in their capability of automatically learning features, especially prone-overlooked implicit features, presenting their great potential for comparative text identification.

With respect to CRE methods, we compare lexicon-based, machine learning-based (SVM), and deep learning-based (BERT + CRF) methods in terms of F1 on uni-tuple (subject and object), bi-tuple, and triple. As shown in Table 5, BERT + CRF obtains the best performance in all situations, significantly outperforming lexicon-based method and SVM and presenting its great potential for CRE. On one hand, lexicon-based and machine learning-based methods generally cannot identify comparative relations that are not included in prepared thesaurus; the irregularity of UGC exacerbates this limitation. For example, for the example #3 in Table 4, only deep learning-based method can correctly detect ‘h6’,

Table 2. Statistical information of the two corpora.

<table>
<thead>
<tr>
<th></th>
<th>Truck data</th>
<th>Mobile phone data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment for CTI</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of labelled instances</td>
<td>30,006</td>
<td>20,000</td>
</tr>
<tr>
<td>Number of comparative instances</td>
<td>2,175</td>
<td>3,284</td>
</tr>
<tr>
<td>No. of comparative: No. of non-comparative</td>
<td>1:12.8</td>
<td>1:5.09</td>
</tr>
<tr>
<td><strong>Experiment for CRE</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of entities</td>
<td>2,409</td>
<td>1,936</td>
</tr>
<tr>
<td>Number of relation tuples</td>
<td>9,962</td>
<td>10,327</td>
</tr>
<tr>
<td>Number of comparative aspect types</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

CTI: comparative text identification; CRE: comparative relation extraction.

Table 3. Performance comparison of different CTI methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method type</th>
<th>Method</th>
<th>AUC</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truck data</td>
<td>Rule based</td>
<td>CSR</td>
<td>0.7103</td>
<td>0.3286</td>
<td>0.5278</td>
<td>0.4049</td>
</tr>
<tr>
<td></td>
<td>Machine learning based</td>
<td>SVM</td>
<td>0.5528</td>
<td><strong>0.7500</strong></td>
<td>0.1111</td>
<td>0.1909</td>
</tr>
<tr>
<td></td>
<td>Deep learning based</td>
<td>BERT</td>
<td>0.7611</td>
<td>0.5843</td>
<td>0.5621</td>
<td>0.5730</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSTM</td>
<td>0.7596</td>
<td>0.5836</td>
<td>0.5590</td>
<td>0.5710</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CNN</td>
<td><strong>0.7773</strong></td>
<td>0.5488</td>
<td><strong>0.6042</strong></td>
<td><strong>0.5752</strong></td>
</tr>
<tr>
<td>Mobile phone data</td>
<td>Rule based</td>
<td>CSR</td>
<td>0.7583</td>
<td>0.5113</td>
<td>0.6363</td>
<td>0.5669</td>
</tr>
<tr>
<td></td>
<td>Machine learning based</td>
<td>SVM</td>
<td>0.6880</td>
<td><strong>0.8254</strong></td>
<td>0.3940</td>
<td>0.5315</td>
</tr>
<tr>
<td></td>
<td>Deep learning based</td>
<td>BERT</td>
<td>0.8816</td>
<td>0.7756</td>
<td>0.8102</td>
<td>0.7924</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LSTM</td>
<td>0.8826</td>
<td>0.7925</td>
<td>0.8076</td>
<td><strong>0.7998</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CNN</td>
<td><strong>0.8902</strong></td>
<td>0.7713</td>
<td><strong>0.8295</strong></td>
<td>0.7994</td>
</tr>
</tbody>
</table>

SVM: support vector machine; BERT: bidirectional encoder representations from the transformers; LSTM: long short-term memory; CNN: convolution neural network; CSR: class sequence rule; AUC: area under curve.
Best performance is highlighted in bold.

Wei et al. 8
Journal of Information Science, 2023, pp. 1–13 © The Author(s), DOI: 10.1177/01655515231165228
Table 4. Some examples explaining limitations of existing research.

<table>
<thead>
<tr>
<th>Index</th>
<th>Sentences</th>
<th>True labels</th>
<th>Predicted labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>看看上柴发动机！实际上大部分翻斗车都装的是科索发动机。Look at the Shangchai engine! Actually many dumpers are equipped with Corso engines.</td>
<td>Comparative</td>
<td>×</td>
</tr>
<tr>
<td>#2</td>
<td>奥铃和东风没有可比性。Aoling and Dongfeng are not comparative.</td>
<td>Non-comparative</td>
<td>×</td>
</tr>
</tbody>
</table>

Index Sentences Lexicon Predicted entities

<table>
<thead>
<tr>
<th>Index</th>
<th>Sentences</th>
<th>Lexicon</th>
<th>Predicted entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>#3</td>
<td>x3000和h6哪个性价比更高？都差不多</td>
<td>x3000, h6</td>
<td>x3000, Jh6, J6</td>
</tr>
<tr>
<td>#4</td>
<td>潍柴发动机与云内动力差不多，但我没用过锡柴。Weichai engine is similar to Yunnei, I have never used Xichai.</td>
<td>Weichai, Yunnei, Xichai</td>
<td>Weichai, Yunnei, Xichai</td>
</tr>
</tbody>
</table>

Index sentences Real relation Predicted relations

<table>
<thead>
<tr>
<th>Index</th>
<th>sentences</th>
<th>Real relation</th>
<th>Predicted relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>#5</td>
<td>潍柴发动机与云内动力差不多，但我没用过锡柴。Weichai engine is similar to Yunnei in terms of power, I have never used Xichai.解放是我用过的最好的品牌jiefang is the best brand I have ever used.</td>
<td>(Weichai, Yunnei, power)</td>
<td>(Weichai, Yunnei, power) √ (Weichai, Xichai, power) × (Xichai, Yunnei, power) ×</td>
</tr>
<tr>
<td>#6</td>
<td>Jiefang is the best brand I have ever used.</td>
<td>(jiefang, super, brand)</td>
<td>–</td>
</tr>
</tbody>
</table>

M-1, M-2, and M-3 represent lexicon-based, machine learning-based, and BERT + CRF, one of prominent deep learning methods.

SVM: support vector machine; CNN: convolution neural network; BERT: bidirectional encoder representations from the transformers; CSR: class sequence rule.

Table 5. F1 comparison of different CRE methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Uni-tuple (subject)</th>
<th>Uni-tuple (object)</th>
<th>Bi-tuple</th>
<th>Triple</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partial matching Truck data</td>
<td>Lexicon based</td>
<td>0.4410</td>
<td>0.3676</td>
<td>0.1519</td>
<td>0.0610</td>
</tr>
<tr>
<td></td>
<td>Machine learning based</td>
<td>SVM</td>
<td>0.4353</td>
<td>0.3792</td>
<td>0.1594</td>
</tr>
<tr>
<td></td>
<td>Deep learning based</td>
<td>BERT + CRF</td>
<td>0.9161</td>
<td>0.9103</td>
<td>0.8329</td>
</tr>
<tr>
<td>Mobile phone data</td>
<td>Lexicon based</td>
<td>0.8631</td>
<td>0.2232</td>
<td>0.2232</td>
<td>0.0805</td>
</tr>
<tr>
<td></td>
<td>Machine learning based</td>
<td>SVM</td>
<td>0.8751</td>
<td>0.2330</td>
<td>0.2323</td>
</tr>
<tr>
<td></td>
<td>Deep learning based</td>
<td>BERT + CRF</td>
<td>0.8934</td>
<td>0.8346</td>
<td>0.8303</td>
</tr>
<tr>
<td>Exact matching Truck data</td>
<td>Lexicon based</td>
<td>0.2674</td>
<td>0.2848</td>
<td>0.1000</td>
<td>0.0393</td>
</tr>
<tr>
<td></td>
<td>Machine learning based</td>
<td>SVM</td>
<td>0.2617</td>
<td>0.2735</td>
<td>0.1074</td>
</tr>
<tr>
<td></td>
<td>Deep learning based</td>
<td>BERT + CRF</td>
<td>0.8720</td>
<td>0.8912</td>
<td>0.7875</td>
</tr>
<tr>
<td>Mobile phone data</td>
<td>Lexicon based</td>
<td>0.7692</td>
<td>0.1388</td>
<td>0.1196</td>
<td>0.0432</td>
</tr>
<tr>
<td></td>
<td>Machine learning based</td>
<td>SVM</td>
<td>0.7702</td>
<td>0.1374</td>
<td>0.1237</td>
</tr>
<tr>
<td></td>
<td>Deep learning based</td>
<td>BERT + CRF</td>
<td>0.8934</td>
<td>0.8178</td>
<td>0.8136</td>
</tr>
</tbody>
</table>

SVM: support vector machine; BERT: bidirectional encoder representations from the transformers; CRF: conditional random field. Best performance is highlighted in bold.
which is not contained in the pre-defined lexicon. On the other hand, the lexicon-based method cannot distinguish words in the lexicon that do not convey comparative information in a specific context, such as ‘Xichai’ in example #4 in Table 4. Machine learning-based methods have the capability of excluding these words (e.g. example #4 in Table 4) via resorting to language features (e.g. the position in the sentence, number of occurrences in the training corpus). Unfortunately, machine learning-based methods sometimes inevitably exclude correct entities. By automatically learning the context information, deep learning methods can not only identify entities not included in the lexicon (‘h6’ in the example #3 in Table 4), but also effectively exclude terms that occur in the thesaurus but non-competing entities (‘Xichai’ in the example #4 in Table 4).

In addition, existing methods work in a pipeline manner, which means that incorrectly identified entities in the early stage could cause errors in subsequent stages. For example, the incorrectly predicted ‘Xichai’ for the example #4 in Table 4 causes incorrectly predicted relations (Weichai, Xichai, power) and (Xichai, Yunnei, power) in the example #5 in Table 4. Therefore, applying the joint learning manner in CRE is a meaningful direction in the future.

Table 6 elaborates limitations of existing research and related novel insights worthy of further exploration.

### Table 6. The limitations of existing research and related novel insights.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Limitations</th>
<th>Promising future directions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data acquisition</td>
<td>Significant cost of labelling for building a dataset.</td>
<td>Develop transfer learning-based models on different datasets derived from various data sources or domains.</td>
</tr>
<tr>
<td>Comparative Text</td>
<td>Existing widely used rule-based methods (e.g., CSR) and traditional machine learning methods are incompetent to identify implicit comparative text and present weak ability across datasets due to dataset-dependent rules (such as differences between comment data and forum data).</td>
<td>Build deep learning-based models by adopting BERT, CNN, LSTM, and et al, in order to conduct automatic text representation learning, in which semantic text understanding is encoded [54–60].</td>
</tr>
<tr>
<td>Identification (CTI)</td>
<td>The skew distribution of comparative text and non-comparative text in the corpus is a serious while unexplored issue. Prior widely used lexicon-based and machine learning-based methods cannot identify implicit relations. Traditional machine learning-based methods require handcrafted feature engineering, which is expert-dependent and time-consuming. The pipeline framework (i.e. extracting each element in comparative relations in a sequential manner) is prone to error propagation (examples #4 and #5 in Table 4), and fail to explore interactions between tasks for extracting different elements.</td>
<td>Develop novel data augmentation methods [61] based on characteristics of comparative expressions to alleviate the skew distribution issue. Leverage advanced deep learning methods, such as BERT + CRF [62], to avoid handcrafted feature engineering, enhance the capability of capturing implicit comparative relations, and improve the performance of comparative relation extraction [63,64].</td>
</tr>
<tr>
<td>Comparative Relation Extraction (CRE)</td>
<td>Prior methods cannot identify superlative relations (example #6 in Table 4). Existing comparative networks lack some important information.</td>
<td>Develop novel deep learning architecture to represent and predict superlative relations. 1. Develop multi-relational [65] comparative network via incorporating multiple aspects. 2. Build hierarchical structure-enhanced comparative network to mine comparative information at different granularities (e.g., brand, model, and product). 3. Construct a trustworthy comparative network via tackling conflicts in the original comparative network. 4. Construct dynamic comparative networks via incorporating the time factor. Explore various applications (e.g., predicting changes of competitive relationships) by leveraging constructed comparative network.</td>
</tr>
<tr>
<td>Applications</td>
<td>Limited applications.</td>
<td>Develop joint-learning-based models.</td>
</tr>
</tbody>
</table>

CTI: comparative text identification; BERT: bidirectional encoder representations from the transformers; CNN: convolution neural network; LSTM: long short-term memory; CRE: comparative relation extraction; CRF: conditional random field; CSR: class sequence rule.
8. Conclusion

This study reviewed existing research related to CTM. First, we described important concepts (including comparative text and CTM) and proposed a novel and comprehensive framework with four components for CTM, while prior research only focused on part of the framework and ignored the other components. Subsequently, we summarised prior work, including data acquisition, CTI, CRE, comparative network construction, and related applications. In addition, we conducted a series of empirical analysis to compare and analyse different methods in CTI and CRE. Finally, we analysed limitations of existing research and provided novel insights on future research directions.

This review has important theoretical and practical implications. With respect to theoretical research, this review can be used for acquiring the research profile and status, clarifying research framework, and finding interesting research directions. In terms of practice, corporations can benefit from CTM in diverse applications and fields, such as identifying competitors of its products and analysing advantages and disadvantages of its products compared with competitors.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Natural Science Foundation of China (grant number 71971170); and the Social Science Foundation of Shaanxi Province, China (grant number 2019P004); and the China Scholarship Council (grant number 202106290094).

ORCID iDs

Jing Liu https://orcid.org/0000-0003-3049-9423
Shenghui Wang https://orcid.org/0000-0003-0583-6969

Note

1. Opinion mining refers to a field focusing on developing multiple methods that could automatically extract opinionated information and identify the polarity of the opinion in terms of a specific entity, and a valuable tool for understanding the opinion from the public.

References


