



# RubE: Rule-based methods for extracting product features from online consumer reviews



Yin Kang\*, Lina Zhou

Department of Information Systems, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA

## ARTICLE INFO

### Article history:

Received 12 May 2015

Received in revised form 12 January 2016

Accepted 26 May 2016

Available online 27 May 2016

### Keywords:

Product feature extraction

Rule-based method

Objective feature

Indirect dependency relation

## ABSTRACT

Motivated by the role of product features in enabling personalized recommendations and marketing, this research aims to extract product features from online consumer reviews. Previous studies are dominated by statistical-based techniques or focused on subjective features that are associated with opinions. In this research, we propose RubE-unsupervised rule-based methods that extract both subjective and objective features from online consumer reviews. We identify objective features by incorporating part-whole relation and review-specific patterns. We extract subjective feature by extending double propagation with indirect dependency and comparative construction. The experiment results demonstrate that RubE significantly outperforms the state-of-the-art techniques for product feature extraction and is generalizable from search goods to experience goods.

© 2016 Elsevier B.V. All rights reserved.

## 1. Introduction

Online consumer reviews are central to the emergence of social commerce [45]. As a form of online word-of-mouth [6], these reviews are used to supplement expert reviews and product descriptions with actual usage experience of consumers. Compared with manufacturer descriptions, consumer reviews are considered unbiased, comprehensive, and credible [15]. Such reviews can help potential buyers make better purchase decisions by reducing uncertainty [7,37], search cost, and switching cost [20]. However, harnessing the power of online consumer reviews is limited by our ability to process the large volume of review data. Automatic product feature extraction has emerged as a promising direction to tap into the potential of online reviews.

The extraction of product features (e.g., attributes and parts) is important for consumers because they care about and have their preference for specific product features [19]. Understanding the role of product features and how they affect consumers' shopping behavior is important in marketing [8]. In addition, feature extraction is fundamental to other complex analyses of online reviews that have significant implications for marketing application, including feature-based opinion mining [23,29–31,40],

personalized product recommendation [34], review summarization [39], and review helpfulness assessment [1,17,22,39].

There has been an extensive amount of work on product feature extraction. Various methods have been applied to this problem, including statistical methods such as hidden Markov model (HMM) [35] and conditional random field (CRF) [14,41] and rule-based methods (RbMs) such as double propagation (DP) [31]. Despite their gradual improvement to the performance in product feature extraction, the state-of-the-art methods expose several limitations: (1) they were dominated by statistical techniques but gave little attention to RbMs. Some recent studies [28,31] provided evidence that RbMs can be effective for product feature extraction. In contrast to statistical methods, RbMs do not rely on training corpora, which remain hard to obtain and difficult to prepare. (2) The small number of RbMs (e.g., [31]) were focused on subjective features on which reviewers had explicitly expressed opinions while ignoring nonsubjective features that were not associated with opinions (i.e., objective features); in addition, they did not address the complexity of indirect dependency in developing the rules, which could hurt the recall. (3) Previous pruning methods mainly relied on product-specific heuristics and information to improve the precision of feature extraction (e.g., [31]), which are difficult to generalize across different types of products. On a related note, most studies have focused on online reviews of search goods (e.g., electronic products), but omitted those of experience goods. For instance, none of the studies has addressed online

\* Corresponding author.

E-mail addresses: [ky6@umbc.edu](mailto:ky6@umbc.edu) (Y. Kang), [zhoul@umbc.edu](mailto:zhoul@umbc.edu) (L. Zhou).

movie reviews, which pose a number of unique challenges to feature extraction [44].

This research aims to improve product feature extraction by addressing the above-mentioned limitations. We propose unsupervised Rule-based Extraction (RubE) methods to extract both objective and subjective product features from online consumer reviews. We extract objective features by discovering part-whole relations and review-specific patterns. We extract subjective features by extending DP [31] with two new types of linguistic structures: indirect dependency and comparative constructions. This research makes multifold contribution to the literature. First, we introduce a new classification scheme of product features in online consumer reviews (i.e., subjective vs. objective categories) that can guide the design of RBMs for feature extraction. Second, we improve the recall of product feature extraction by addressing objective feature extraction for the first time and by improving the extraction of subjective features. Third, we improve the precision of feature extraction by introducing a two-step domain-independent pruning method based on semantic similarity and document frequency. In addition, we provide evidence for the superior generality of RubE to previous RBMs.

The remainder of this paper is organized as follows: we first review previous literature as related to our research setting; we then propose our feature extraction methods – RubE – followed by the introduction of our data collection, experiment design, and results; finally, we discuss our findings and conclude the paper with future work.

## 2. Definitions and related work

### 2.1. New classification scheme

We use “feature” as a generic term to refer to attributes, components, and related concepts of both the object as a whole and its parts. Product features have been classified as explicit and implicit categories based on whether the related feature expressions actually appear in a review or not, or classified as frequent and infrequent features based on the level of their occurrence frequency [21]. Motivated by a recent feature extraction study (e.g., [31]), we propose a new classification scheme based on whether there are opinions expressed about the features in a review, which classified features into subjective and objective categories.

*Definition 1 (subjective feature):* If a reference to feature  $f$  is associated with opinion expressions in a review,  $f$  is considered a subjective feature. For example, “I like this bundle, the rechargeable battery is a life saver . . .” and “This camera is awesome . . .”

*Definition 2 (objective feature):* If a reference to feature  $f$  is not associated with any opinions in a review,  $f$  is considered an objective feature. For example, “This phone comes with a rechargeable battery . . .” and “This phone has three colors . . .”

State-of-the-art rule-based feature extraction studies [9,31] only addressed subjective features but ignored objective features. It is important to consider objective features for several reasons. First, incorporating objective features is expected to boost the recall of extraction methods. The distinctive characteristics of objective features point to the need for developing new extraction methods. Second, the extraction of subjective and objective features can mutually reinforce each other. This is because the differentiation of subjective and objective features depends on their specific review context. In other words, an objective feature extracted from one review context may be expressed as a subjective feature in another context (see *rechargeable battery* in Definitions 1 and 2). Third, objective features have the potential to support promotional review detection and assessment of

trustworthiness of online information [25] in that objective features are presumably more objective and credible.

### 2.2. Related work

Based on our extensive literature review, extant methods for product feature extraction can be grouped into two broad categories: statistical-based and RBMs (see Table 1). Each of them can be further classified into supervised and unsupervised methods based on whether they require training data labeled with product features.

Statistical methods have been widely used in product feature extraction. Hu and Liu [13] proposed class sequential rule mining to generate meaningful patterns for extracting features from online reviews; Wong and Lam [35] used HMM for identifying (product feature-based) hot items from auction websites, and they treated product feature extraction as a graph labeling problem using CRF in a follow-up study [36]. Unlike supervised methods, unsupervised methods do not require training data. Hu and Liu [12] applied association rule mining (ARM) to feature extraction by assuming that people tend to use the same words when commenting on the same product features. Popescu and Etzioni [27] adapted PMI to the problem of feature extraction by computing the PMI between a noun/noun phrase and class-specific discriminators. Chen et al. [4] exploited how to incorporate domain knowledge into topic modeling to improve the performance of product feature extraction, and subsequently proposed automated knowledge LDA to learn prior knowledge and generate product features [3].

The other paradigm of feature extraction methods is rule based, which uses rules derived from uncovered patterns. DP is a state-of-the-art rule-based semisupervised method for extracting noun-phrase-based features [31]. DP is based on the dependency relation between product features and opinion words. Zhang et al. [42] extended DP with *part-whole* pattern and *no* pattern. Gindl et al. [9] leveraged syntactic patterns (e.g., dependency relations) that take sentence relations into consideration. Poria et al. [28] manually encoded three sets of rules in RBM, including subjective-noun relations, nonsubjective-noun relation, and other rules, to address both implicit and explicit features.

Statistical methods generally require large corpora to mitigate the data scarcity problem in computing probabilities. Preparing the corpora is challenging particularly for supervised methods because creating a large-scaled annotated corpus can be very costly. Despite the availability of corpora in support of traditional information extraction research, those corpora were collected from editorials and news, which are distinctively different from online consumer reviews. In addition, the target of traditional information extraction was focused on named entities rather than product features. By contrast, RBMs do not rely on large corpora. Nevertheless, DP and other variants mainly dealt with subjective features and overlooked objective features. In addition, although DP was built on dependency structures, it only considered direct dependency while ignoring indirect dependency relations. The RBM proposed by Poria et al. [28] used handcrafted rules, which were inefficient to develop and difficult to generalize across domains. Furthermore, some named entities (e.g., brand and model names) that are unique to the product review domain have not been exploited by extant feature extraction methods.

## 3. RubE methods

We propose RubE, unsupervised Rule-based Extraction methods for extracting both subjective features and objective features.

**Table 1**

A summary of previous studies on product feature extraction.

Method	Study	Techniques	Performance	Product type/Source
Statistical-based Unsupervised	Hu and Liu [12]	Association rule mining	Recall: 80%	Electronic products from Amazon
	Popescu and Etzioni [27]	Compactness pruning and redundancy pruning	Precision: 72%	Electronic products from Amazon
	Scaffidi et al. [34]	Frequency-based extraction and PMI	Precision: 94%	
	Chen et al. [4]	A language model approach based on statistical distribution difference between different corpora	Recall: 77%	7 products (e.g., electronics, books, and games) from Amazon
	Chen et al. [3]	MC-LDA (LDA with m-set and c-set) based on an extended generalized E-GPU model	Precision: 85–90%	4 products (e.g., camera, food, computer, and care) from Amazon
Statistical-based Supervised/ Semisupervised	Chen et al. [3]	Use prior knowledge to extract aspects, propose AKL to learn knowledge and deal with error knowledge	N/A	Electronic products from Amazon
	Wong and Lam [35]	HMM-based learning method	Precision: 78.5%	Electronic products from 3 bidding websites
	Hu and Liu [13]	Class Sequential Rules Mining	Recall: 72%	Electronic products from Amazon
		Pattern matching	Precision: 83.8%	
	Probst et al. [30]	Naive Bayes combined with a multiview semisupervised algorithm (co-EM)	Recall: 84.9%	2 sport products from DICK's sporting goods
	Kobayashi et al. [18]	Combine contextual and statistical clues	Precision: 38%	
	Wong and Lam [36]	Template slot filling	Recall: 75%	Restaurants from Japanese weblog posts
		CRF-based learning method	Precision: 72%	
	Jin and Ho [16]	Lexicalized HMM	Recall: 62%	Electronic products from auction websites
			Precision: 78.5%	
Rule-based Unsupervised	Mukherjee and Liu [26]	Two novel statistical models: SAS and ME-SAS	Recall: 74.5%	16 cameras from Amazon
	Zhang et al. [42]	Part-whole and “no” patterns	Precision: 73–88%	
		Pruning-based HIT	Recall: 65–97%	Hotel from TripAdvisor
	Gindl et al. [9]	Syntactic Patterns (Dependency relations)	N/A	
	Poria et al. [28,29]	Anaphora resolution	Precision: 60–70%	4 products (e.g., car, mattress, phone, and LCD) from Amazon
Rule-based Supervised/ semisupervised		Manually encoded rules for sentences with subjective noun relation or not, and other specific conditions	Recall: 40–70%	
	Qiu et al. [31]	Double propagation	N/A	Electronic products from Amazon
			Precision: 82.15–93.25%	
			Recall: 86.15–93.32%	
			Precision: 88%	
			Recall: 83%	

Note: N/A indicates the study adopted evaluation metrics that are distinctively different from precision and recall.

### 3.1. Subjective feature-oriented extraction methods

RubE aims to extract subjective features by applying DP [31] to deal with direct dependency, and, more importantly, by introducing two new types of linguistic structures: indirect dependency and comparative constructions.

Dependency refers to a binary asymmetrical connection between two words in a sentence. There are two types of dependency: direct and indirect [31].

- Direct dependency indicates that one word directly depends on another, or both directly depend on an intermediate word. Based on the observation that opinion words are often used to modify features, opinion words and features can be directly linked to each other via certain dependency relations. For example, “The phone has a great color screen” (*great* → *mod* → *color screen*), and “iPod is the best mp3 player” (*best* → *amod* → *player* ← *nsubj* ← *iPod*, where ‘→’ or ‘←’ denotes dependency relation, “*amod*” and “*nsubj*” are specific types of dependency relations.)
- Indirect dependency indicates that one word indirectly depends on another through a third word or both depend on a third word through additional words (see Table 2). For example, “the camera is very easy to use” (*camera* → *xsubj* → *use* → *xcomp* → *easy*), where feature “camera” depends on “easy” via another word “use.”

We followed DP [31] in handling direct dependency. DP extracted opinion words and product features iteratively based on propagation using the dependency relations between them. It required a set of seed opinion words to bootstrap the propagation. Eight rules were generated to perform the following four tasks: (1) extracting product features using opinion words; (2) extracting product features using the extracted product features; (3) extracting opinion words using the extracted product features; and (4) extracting opinion words using both the given and the extracted opinion words. We use the same strategy but exclude opinion words in the final step.

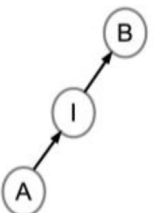
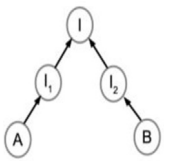
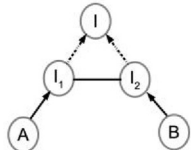
However, DP did not handle indirect dependency due to its complexity. Thus, the following introduction is focused on our method for addressing indirect dependency.

#### 3.1.1. Indirect dependency

We fill the gap in the literature by incorporating indirect dependency in extracting subjective features. We propose the following three types of indirect dependency structures (see Table 2).

- Simple indirect dependency, where word *A* indirectly depends on word *B* via an intermediate word *I*, and word *A* or *B* is a feature candidate. This kind of structure usually describes performing certain actions on features.

**Table 2**  
Rules for extracting indirect dependency.

Types	Structures	Rules	Examples
Simple		$A \rightarrow \text{Dep}_1 \rightarrow I \rightarrow \text{Dep}_2 \rightarrow B$ $A \in \{\text{NN}\}, I \in \{\text{VB}\} B \in \{\text{features}\}$ $\text{Dep}_1 \in \{\text{dobj}, \text{nsubj}, \text{xsubj}\}$ $\text{Dep}_2 \in \{\text{prepc}, \text{xcomp}\}$	"I'm a happier person after discovering the i/p button" (person (NN) $\leftarrow$ Dep (prepc-after) $\leftarrow$ discovering (VBG) $\leftarrow$ Dep (dobj) $\leftarrow$ button (NN))
Explicitly pivoted		$A \rightarrow \text{Dep}_1 \rightarrow I_1 \rightarrow \text{Dep}_2 \rightarrow I \leftarrow \text{Dep}_3 \leftarrow I_2 \leftarrow \text{Dep}_4 \leftarrow B$ $A \text{ or } B \in \{\text{features}\}$ $\text{Dep}_{1,2,3,4} \in \{\text{nn}, \text{prep}\}$ $I \in \{\text{VB}\}$	"You can move the focus range to almost anywhere in the scene with the push of a button" (Focus (N) $\rightarrow$ Dep (nn) $\rightarrow$ range (N) $\rightarrow$ Dep (dobj) $\rightarrow$ move (verb) $\leftarrow$ Dep (prep-in) $\leftarrow$ scene (N) $\leftarrow$ Dep (prep-with) $\leftarrow$ push (verb))
Implicitly pivoted		$A \rightarrow \text{Dep}_1 \rightarrow I_1 \rightarrow I_2 \leftarrow \text{Dep}_2 \leftarrow B$ $A \in \{\text{NN}\}, I_{1,2} \in \{\text{VB}\}$ $I$ : missing, $B \in \{\text{features}\}$ $\text{Dep}_{1,2} \in \{\text{nsubj}, \text{dobj}\}$	"Another nice thing is that the unit has both optical and coax digital audio outputs" (Thing (NN) $\rightarrow$ Dep (nsubj) $\rightarrow$ is (verb) $\leftarrow$ Dep (ccomp) $\leftarrow$ has (verb) $\leftarrow$ Dep (dobj) $\leftarrow$ audio output (NN))

Note: A, B, and I are words in a sentence,  $A, B \in \{\text{features}\}$ ; I denotes intermediate word.

Dep (dependency relations)  $\in \{\text{'ccomp'}, \text{'amod'}, \text{'xcomp'}, \text{'partmod'}, \text{'csbj'}, \text{'csbjpass'}, \text{'nsubj'}, \text{'nsubjpass'}, \text{'dobj'}, \text{'iobj'}, \text{'acompl'}\}$ . JJ (adjective)  $\in \{\text{JJ}, \text{JJR}, \text{JJS}\}$ , VB (verb)  $\in \{\text{VB}, \text{VBZ}, \text{VBP}\}$  and RB (adverb)  $\in \{\text{RB}, \text{RBR}, \text{RBS}\}$ .

- Explicitly pivoted indirect dependency, where both words A and B indirectly depend on word I via word  $I_1$  and word  $I_2$ , respectively, and both A and B can be feature candidates. This type of indirect dependency is typically expressed as a complex sentence.
- Implicitly pivoted indirect dependency, where both A and B indirectly depend on an intermediate word I via  $I_1$  and  $I_2$ , respectively, and either A or B is a feature candidate.

Similar to DP, we extract product features iteratively via propagation using the indirect dependency relations as defined in Table 2. Specifically, node A or B from the dependency relations that satisfied the conditions of the rules (see Table 2) would be extracted as potential features. Consider the following sentence as an example. "I'm a happier person after discovering the i/p button" (person (NN)  $\leftarrow$  Dep (prepc-after)  $\leftarrow$  discovering (VBG)  $\leftarrow$  Dep (dobj)  $\leftarrow$  button (NN)). Given that the dependency structure of the sentence satisfied the conditions of two of the specified indirect dependency relations from Table 2 (i.e., POS of Word I is VB, and POSes of Word A and B are NNs), the word associated with "dobj" would be extracted as a candidate product feature.

### 3.1.2. Comparative constructions

Comparative constructions are used to present explicit orderings among different objects in terms of the degree or amount to which they possess some gradable property. Identifying comparative sentences is useful for feature extraction because consumers may express their opinions in a relative sense by comparing a target product against other alternatives. The syntactic structure of comparative sentences tends to be complex, in that they generally involve long-distance dependency and lack opinion words.

Before extracting comparative constructions using RubE, we first classify comparative constructions into four subtypes: comparative (e.g., longer), superlative (e.g., best), equality (e.g.,

the same as), and unique words (e.g., beat); and then we extract rules based on POS tags and comparative words from the sentence structures of each type (see Table 3).

The representation of the rules followed Sarawagi [33]. All of the rules are defined using indicative words and patterns of dependency relations and sentence structures.

To apply the rules, we first match a review against patterns on the left of each rule; if there is a match, we then extract the components tagged with ":" as candidate product features.

The extraction of comparative constructions also needs to deal with noncomparative sentences that contain comparative words such as "I cannot agree with you more," "more than often." To this end, we analyze the dependency relations between two nodes and check whether there existed certain dependency relations and POS tags (e.g., "prep\_than" and "JJR") expected of a comparative structure. In addition, we compile a list of patterns to filter other noncomparative sentences.

### 3.2. Objective feature-oriented extraction methods

Objective features are not uncommon in online consumer reviews for at least the following two reasons: (1) reviews filled with subjective features are more likely to mislead consumer's decision and (2) descriptive reviews tend to state facts about a product or product features such as "no audio, no video," and "There are three colors to choose from . . ." However, these features were not addressed by extant extraction methods. For instance, DP is not applicable due to the lack of use of opinion words in expressing objective features, and frequent-item based methods would be ineffective because objective features tend to have low occurrence frequency in online reviews.

We developed methods for extracting objective features based on two observations. First, objective features are typically expressed by a variety of lexicosyntactic structures such as part-

whole relation. Second, objective features are often expressed in concrete terms such as specific quantities, weights, and measures.

### 3.2.1. Part-whole relation

A *part-whole* relation indicates that one or more objects are part of another object. Both supervised approaches (e.g., [10]) and unsupervised approaches (e.g., [38]) have been developed to detect part-whole relations. Zhang et al. [42] incorporated part-whole relations to address the low recall problem of DP. However, they assumed that the most frequent words in a corpus belong to class concepts and their *part* words product features. The assumption is problematic because the same word can be used as both a class and a feature word (e.g., “*lens* of the camera,” “*cap* of *lens*”). We extend and improve the syntactic patterns of *part-whole* relations [10] for the extraction of objective features. Compared with the unambiguous structure of part-whole relations [10], the extraction of its ambiguous counterpart is more challenging, which was the focus of RubE.

In addition to existing lexicosyntactic patterns of ambiguous part-whole relations such as “ $NP_x PP_y$ ,” “ $NP_x$ ’s  $NP_y$ ,” “ $NP_x$  have  $NP_y$ ,” and “ $NP_x$  of  $NP_y$ ,” we introduce two new sentence-level patterns: “ $NP_x$  verb  $NP_y$  ( $PP_z$ )” and “ $PRP/Ex$  Verb  $NP$ ,” where  $NP$ ,  $PP$ ,  $PRP$ , and  $Ex$  denote noun phrase, prepositional phrase, personal pronoun, and existential there, respectively, and  $NP_x$  and  $NP_y$  contain the part or a

candidate feature, separately. These patterns are introduced in detail below.

- **Genitive phrase ( $NP_x PP_y$ ,  $NP_x$ ’s  $NP_y$ ,  $NP_x$  off/have  $NP_y$ ):** Genitive phrases typically follow one of the patterns listed in the above parentheses, such as “*battery life*” and “*camera’s battery*.” However, these patterns can be ambiguous. For example, the following two phrases share the same syntactic structure, “*the engine of my car*” and “*the car of my friend*,” but only the first phrase denotes a part-whole relation. To resolve the ambiguity, we developed a pruning strategy (see the pruning section).
- **Verb phrase ( $NP_x$  Verb  $NP_y$  [ $PP_z$ ], where  $PP_z$  is optional):** Based on the sample relations [10], we extract cue verbs such as “have,” “include,” “contain,” and “consist” to extract features from ambiguous part-whole patterns; for example, “*the camera bundle contains a memory card*.”
- **Verb phrase with prefix ( $PRP/Ex$  Verb  $NP$ ),** where  $PRP$  (pronoun) is commonly used to refer a product. For example, “*It has a tripod and a memory card*.” In addition, some *part-whole* relations can also be expressed with verbal phrases such as “there be” and “comes with.”

Based on the above patterns, we generate a set of rules for extracting part-whole relations, and sample rules are listed in

**Table 3**  
Sample Rules of Comparative Construction.

Category	Pattern(s)	Sample Rules	Examples
Comparative	$G(F, P_1) + CW$ + $G(F, P_2)$	$\{[Dependency = Dep_2]? \{Node = NP POS = NN\}:f$ $\{String = “of”\} \{Node = NP POS = NN\} \{String = CW\} \{Dependency = Dep_1\}$ $\{[Dependency = Dep_2]? \{Node = NP POS = NN\} \{String = “of”\} \{Node = NP POS = NN\}$ $POS = DT\} \} \rightarrow Feature = f$	“The battery life of Camera X is longer than that of Camera Y” “The battery life” (NP) $\rightarrow$ nmod of $\in$ $Dep_2$ $\rightarrow$ “Camera X” (NP) “longer than” (CW) $\rightarrow$ advmod $\in$ $Dep_1$ $\rightarrow$ “that” (DT) $\rightarrow$ nmod of $\in$ $Dep_2$ $\rightarrow$ “Camera X” (NP) therefore, “battery life”:f $\rightarrow$ feature
	$P_1 + CW + F + P_2$	$\{Node = NP POS = NN\} \{String = CW\} \{Dependency = Dep_1\} \{Node = NP POS = NN\}:f$ $POS = NN\} \rightarrow Feature = f$	In “Camera X has a longer battery life than Camera Y” “Camera X” (NP) $\rightarrow$ “longer” (CW) $\rightarrow$ amod $\in$ $Dep_1$ “Camera Y” (NP) $\rightarrow$ “battery life” (NP) $\rightarrow$ “Camera Y” (NP) therefore, “battery life”:f $\rightarrow$ feature
Equality	$G(F, P_1) + G(F, P_2) + EW$	$\{[Dependency = Dep_2]? \{Node = NP POS = NN\}:f$ $\{String = “of”\} \{Node = NP POS = NN\} \{String = EW\} \{Dependency = Dep_1\} \}$ $\{String = EW\} \{Dependency = Dep_1\} \} \rightarrow Feature = f$	In “The price of camera X and camera Y are the same.” “The price” (NP) $\rightarrow$ nmod of $\in$ $Dep_2$ $\rightarrow$ “Camera X” (NP)/“Camera Y” (NP) “The price” (NP) $\rightarrow$ nmod of $\in$ $Dep_1$ $\rightarrow$ “same” (EW) therefore, “price”:f $\rightarrow$ feature
	$P_1 + P_2 + EW + F$	$\{Node = NP POS = NN\} \{Node = NP POS = NN\} \{String = EW\} \{Dependency = Dep_1\} \{Node = NP POS = NN\}:f \rightarrow Feature = f$	In “Camera X and Camera Y are about the same size.” “Camera X” (NP)/“Camera Y” (NP) $\rightarrow$ nsubj of $\in$ $Dep_1$ $\rightarrow$ “size” “same” (EW) $\rightarrow$ amod of $\in$ $Dep_1$ $\rightarrow$ “size” therefore, “size”:f $\rightarrow$ feature
Superlative	$G(F, P) + SW$	$\{[Dependency = Dep_2]? \{Node = NP POS = NN\} \{String = “”\} \{Node = NP POS = NN\}:f$ $\{String = SW\} \} \rightarrow Feature = f$	In “Camera X’s lens is the best.” “Camera X” (NP) $\rightarrow$ “best” (CW) $\rightarrow$ nsubj $\in$ $Dep_1$ $\rightarrow$ “lens” (NN) therefore, “lens”:f $\rightarrow$ feature
	$F + SW$	$\{Node = NP POS = NN\}:f \{String = SW\} \{Dependency = Dep_1\} \} \rightarrow Feature = f$	In “The picture quality is the best.” “best” (CW) $\rightarrow$ nsubj $\in$ $Dep_1$ $\rightarrow$ “picture quality” (NP) therefore, “picture quality”:f $\rightarrow$ feature
Unique words	$P_1 + P_2 + UW + F$ Or $P_1 + F + UW + P_2$	$\{Node = NP POS = NN\} \{Node = NP POS = NN\} \{String = UW\} \{Dependency = Dep_1\} \{Node = NP POS = NN\}:f \rightarrow Feature = f$	In “Camera X and Camera Y have different OS.” “Camera X” (NP) $\rightarrow$ “Camera Y” (NP) $\rightarrow$ “different” (UW) $\rightarrow$ amod $\in$ $Dep_1$ $\rightarrow$ “OS” (NN) therefore, “OS”:f $\rightarrow$ feature

Note:  $G$ : a genitive relation between two words;  $F$ : features, “ $P$ ”: product words;  $CW$ : comparative words;  $EW$ : equality words;  $SW$ : superlative words;  $UW$ : unique words; “ $Node$ ”: syntactic structure; “ $NP$ ”: noun phrases; “ $NN$ ”: noun; “ $f$ ”: product feature;  $Dep_1 \supset \{‘nsubj’, ‘prep\_than’, ‘nmod’, ‘amod’, ‘advmod’\}$ ;  $Dep_2 \supset \{‘prep\_of’, ‘nmod\_of’, ‘advmod’, ‘amod’\}$ .



**Table 4.** For example, the following sentence “the camera comes with a rechargeable battery,” contains a syntactic pattern that matches the second rule in Table 4. Specifically, “the camera” matches NP, which is followed by “come,” and “a rechargeable battery” matches NP; thus, we extracted the “battery” as a potential product feature.

### 3.2.2. Review-specific patterns

Online consumer reviews have unique genre characteristics, which point to some review-specific patterns such as specific named entities, negative and with-expressions, and cue phrases.

- **Named entities.** An expression of product brand and/or model names typically exhibits some structural patterns such as brand + numbers. We adapt named entity recognition techniques and regular expressions to the extraction of the domain-specific entities.
- **Negative and with-expressions.** Short negation (i.e., “no/not”) and with-expression patterns (i.e., “with/without”) may indicate features in an online consumer review such as “no audio” and “without auto mode.” We extend the negation expressions [42] by introducing not-expressions and “with/without” patterns to capture additional features.
- **Cue phrases.** Some phrases such as “pros and cons” commonly signal following expressions of features.

### 3.3. Pruning methods

To improve the precision of feature extraction, we developed a two-stage pruning method, which includes (1) nonfeature candidate identification and (2) semantic similarity and document frequency-based filtering. In view of the informal writing style of online consumer reviews such as conjoined words (e.g., wifi and wi-fi) and misspellings, we first apply a text similarity function based on the editing distance [32] to group variant expressions of the same features.

- **Stage 1: Identification of nonfeature candidates.** The identification stage used two strategies: self-filtering and mutual exclusion. We propose a self-filtering strategy – if a word feature candidate is not part of any phrase candidate, the word feature is likely to be noise, and vice versa. In mutual exclusion [31], if more than one feature candidate appeared in the same sentence without any conjunctions in-between, all such candidates except for one would be treated as noise. In addition, the survival feature candidate is the one with the highest similarity to product word and then highest frequency (when similarity scores are the same).
- **Stage 2: Filtering nonfeature candidates.** We design two complementary filtering methods. The intersection between

the two sets of results would be treated as the final set of nonfeatures.

- **Document frequency-based filtering.** The design of the method is inspired by a common assumption in feature extraction research that terms with rare occurrences are unlikely to be features. Here, documents refer to consumer reviews. We choose the most conservative threshold for document frequency in this study – 2.
- **Semantic similarity-based filtering.** Terms that are not semantically related to the product under review are unlikely to be features of the product. Based on the assumption, we propose a new measure – differential similarity to assess the semantic relatedness between a candidate feature and a product. Drawing on Wordnet-based similarity metrics such as hso [11] and lesk [2], the differential similarity of candidate feature  $f$  ( $f \in F$ ),  $DS(f_j)$ , was derived as the difference in average similarity between  $F$  and a product word before and after removing  $f$ -inferior features from  $F$ , which included  $f$  itself and other candidates whose similarity scores were below that of  $f$  (see equation (1)). For example, given  $F = \{f_i | i = 1 \dots n\}$  of product  $P$ , where  $n$  is the total number of features of  $P$ , we first measured the semantic similarities between  $f_i$  and  $P$ :  $\{sim(f_i, P) | i = 1 \dots n\}$ , and then sorted all  $f_i \in F$  in an ascending order of  $sim(f_i, P)$  into  $F' = \{f_j | j_1 < j_2 \text{ when } sim(f_{j_1}, P) \leq sim(f_{j_2}, P), j_1, j_2 = 1 \dots n\}$ .  $DS(f_j)$ , the differential similarity of  $f_j$  was normalized by the average similarities of all  $f_j \in F$ . Those features with  $DS(f_j)$  below 0.5 were filtered. The threshold was determined empirically.

$$DS(f_j) = \left( \frac{1}{n-j} \sum_{k=j+1}^n sim(f_k, P) - \frac{1}{n} \sum_{k=1}^n sim(f_k, P) \right) / \left( \frac{1}{n} \sum_{k=1}^n sim(f_k, P) \right) \quad (1)$$

Assume that  $F = \{\text{“lens,” “battery,” “coffee,” “car”}\}$ ,  $P = \text{“camera,”}$   $sim(\text{“camera,” “lens”}) = 0.8$ ,  $sim(\text{“camera,” “battery”}) = 0.7$ ,  $sim(\text{“camera,” “coffee”}) = 0.3$ , and  $sim(\text{“camera,” “car”}) = 0.2$ . Then,  $F' = \{\text{“car,” “coffee,” “battery,” “lens”}\}$ .  $DS(\text{“car”}) = [(0.8 + 0.7 + 0.3)/3 - (0.8 + 0.7 + 0.3 + 0.2)/4] / [(0.8 + 0.7 + 0.3 + 0.2)/4] = 0.2 < 0.5$ ,  $DS(\text{“coffee”}) = 0.5$ ,  $DS(\text{“battery”}) = 0.6$ , and  $DS(\text{“lens”}) = 1$ . Thus, car and coffee were filtered from  $F$ .

## 4. Experiments

### 4.1. DataSets

We test RubE using two datasets. One dataset consists of reviews of five electronic products [12], and the other contains movie reviews [43]. The distinctive types of products represented by the two datasets allowed us to test the generality of RubE. Some descriptive statistics of the datasets is reported in Table 5.

**Table 4**  
Sample Rules of Part–Whole Relation.

Category	Sample Rules	Usage Examples
Genitive phrase	$\{Node = NP   POS = NN\} \{String = "s"? \} \{Node = NP   POS = NN\} : y \rightarrow Feature = y$	“This phone’s battery life” “This phone” (NP), “battery life” (NP), “s” match $\{String = "s"? \}$ therefore, “battery life”: $f \rightarrow feature$
Verb phrase	$\{Node = NP   POS = NN\} \{POS = VB\} \{Node = NP   POS = NN\} \{Node = NP   POS = NN\} : y \rightarrow Feature = y$	In “This camera comes with a rechargeable battery” “This phone” (NP), “a rechargeable battery” (NP), “come” (VB) therefore, “battery”: $f \rightarrow feature$
Verb phrase with prefix	$\{POS = PRP   POS = Ex\} \{POS = VB\} \{Node = NP   POS = NN\} : y \rightarrow Feature = y$	In “There is a rechargeable battery” “There is” (Ex), “a rechargeable battery” (NP), therefore, “battery”: $f \rightarrow feature$

Note: “Node”: syntactic structure; “NP”: noun phrases; “NN”: noun; “VB”: verb; “PRP”: personal pronoun; “EX”: existential there.

**Table 5**  
Descriptive Statistics of the Datasets.

Product Name	No. Reviews	No. Features
Camera (Canon)	45	79
Camera (Nikon)	34	96
Cell Phone (Nokia)	41	67
MP3 Player (Creative)	95	57
DVD Player (Apex)	99	49
Movie	356	30

#### 4.2. Evaluation metrics and baselines

We select precision, recall, and F-measure as the evaluation metrics. Precision is defined as the ratio of the number of correctly identified features to the total number of identified features. Recall is defined as the ratio of the number of correctly identified features to the total number of actual features in the gold standard. F-measure is defined as a harmonic mean of the precision and recall.

We select four methods as the baselines, including ARM [12], DP [31], RbM [28] proposed by Poria et al., and CRF [14,41]. They represent either the most popular or the state-of-the-art methods for product feature extraction.

#### 4.3. Procedure

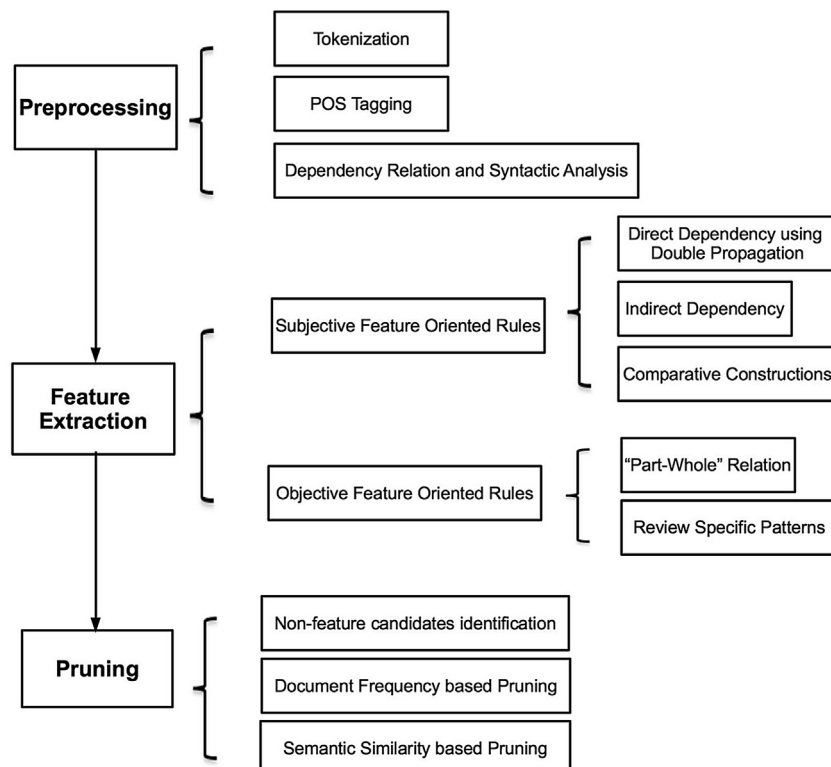
The procedure for feature extraction consists of three key processes: preprocessing, feature extraction, and pruning, as shown in Fig. 1.

As both feature extraction and pruning have been described in detail in the previous section, so we focus on the preprocessing step here. The preprocessing used the following analyses:

tokenization, POS tagging, named entity recognition, and dependency grammar analysis, using tools including Natural Language Toolkit (NLTK; [24]) and Stanford Natural Language Processing (NLP; [5]). Among them, the named entity recognition is focused on some review-specific entities such as brand names.

#### 4.4. Results

The average performances of RubE and the baseline methods over all the products are reported in Table 6. It is shown from the table that RubE outperforms all the baselines in terms of recall, precision, and F-measure. Specifically, the average recall improvement of RubE is about 9% over the ARM, RbM [28], and CRF, and about 5% over the DP. The average precision improvement is about 17% over ARM, about 2% over DP, 11% over RbM [28], and about 10% over CRF. The average improvement of RubE in the F-measure is 13% over ARM, 3% over DP, 9% over RbM [28], and about 9% over CRF. Paired-sample *t*-tests showed that the recall of RubE is greater than that of DP ( $p < 0.05$ ), ARM ( $p < 0.01$ ), RbM [28] ( $p < 0.05$ ), and CRF ( $p < 0.01$ ), respectively; the precision of RubE is higher than that of DP ( $p < 0.1$ ), ARM ( $p < 0.01$ ), RbM [28] ( $p < 0.01$ ), and CRF



**Fig. 1.** Procedure for Feature Extraction.

**Table 6**

Average Recall, Precision, and F-measure of RubE and the baseline methods.

Metrics	ARM	DP	RbM [28]	CRF	RubE
Recall	0.78	0.82	0.78	0.78	0.87
Precision	0.71	0.86	0.77	0.78	0.88
F-measure	0.74	0.84	0.78	0.78	0.87

( $p < 0.01$ ), and the F-measure of RubE is higher than that of DP ( $p < 0.1$ ), ARM ( $p < 0.01$ ), RbM [28] ( $p < 0.01$ ), and CRF ( $p < 0.01$ ), respectively.

To assess the impact of our pruning strategies, we report the performance of RubE before and after pruning on each of the products in Fig. 2. The figure shows that our proposed pruning methods significantly improved precision with only a slight drop on recall.

To gain an understanding of the generality of RubE and other methods, we reported their performance for electronic products and movies separately in Fig. 3. The side-by-side comparisons showed that RubE outperformed all the baseline methods on both electronic products and movie reviews, and the improvements of RubE over the baselines are even more pronounced on the movie data. The results demonstrated the generalizability of RubE across different types of products.

To gain insights into efficacy of individual components of RubE in recall, we report their marginal improvements over DP for electronic products and for movies in Table 7 separately. The results show that every component of RubE contributed to its overall performance with indirect dependency leading to the largest improvements.

## 5. Discussion

### 5.1. Findings and alternative explanations

Our experimental results demonstrate the superior performance of RubE to the state-of-the-art methods for extracting features from online consumer reviews. Specifically, the proposed

feature extraction methods improve the recall, and the pruning methods improve the precision of feature extraction. Moreover, the performance improvements are consistently demonstrated across the online reviews of two distinctive product types.

Given the importance of our proposed classification scheme of product features in guiding our development of RubE, we provide detailed information about the two types of features. It should be noted, however, that the distinction of subjective features from objective features is inconsequential for the extracted features due to the context dependency of their definitions. For example, the same feature (e.g., lens) may be extracted as a subjective feature from one context of camera review and as an objective feature from another context. In other words, the same product features may be extracted by more than one rule due to the overlaps among the features extracted by different components of RubE. As a result, we do not report descriptive statistics of either type of the features in the datasets. Nevertheless, to give a rough idea of the occurrences of objective features, which were overlooked in previous methods, we computed the percentage distribution of features extracted by individual components of RubE (i.e., recall) by replicating their overlaps in the set of extracted features. For example, if a feature is extracted by both indirect dependency and part-whole components, typically from different reviews, the feature would be counted twice toward the total number. The recall percentage distribution of RubE components is reported in Fig. 4.

It is shown from Fig. 4 that objective features, including those extracted by part-whole and review-specific patterns accounted for about 15% of the total recall. This number was not negligible, which provided preliminary evidence for incorporating objective features. Compared with part-whole relations, review-specific

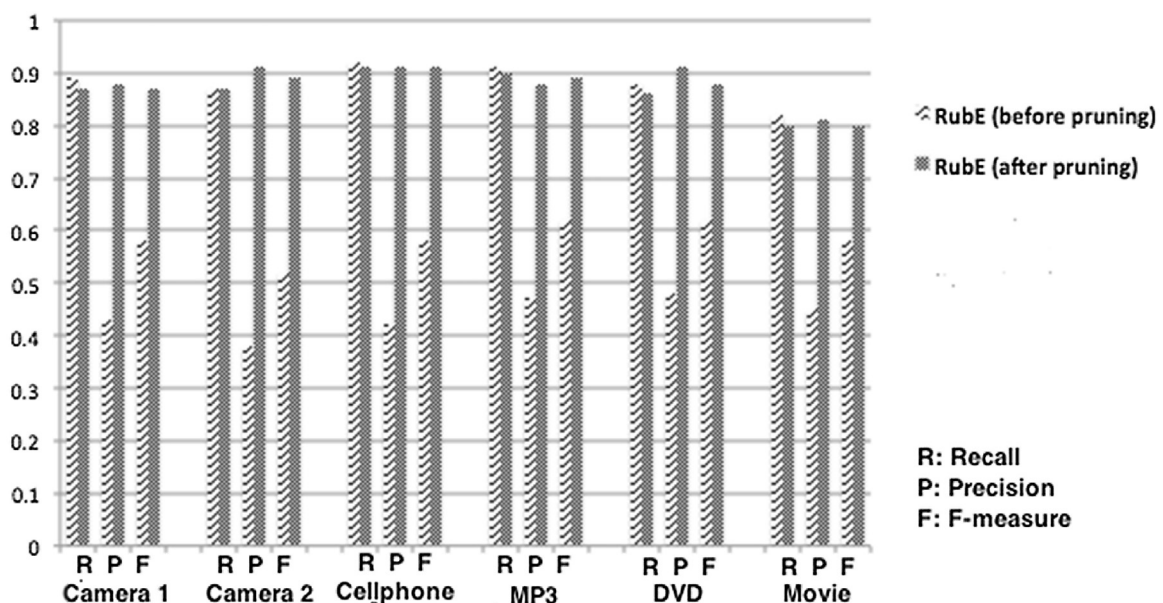


Fig. 2. Performance (recall, precision, and f-measure) of RubE before and after pruning on all products.



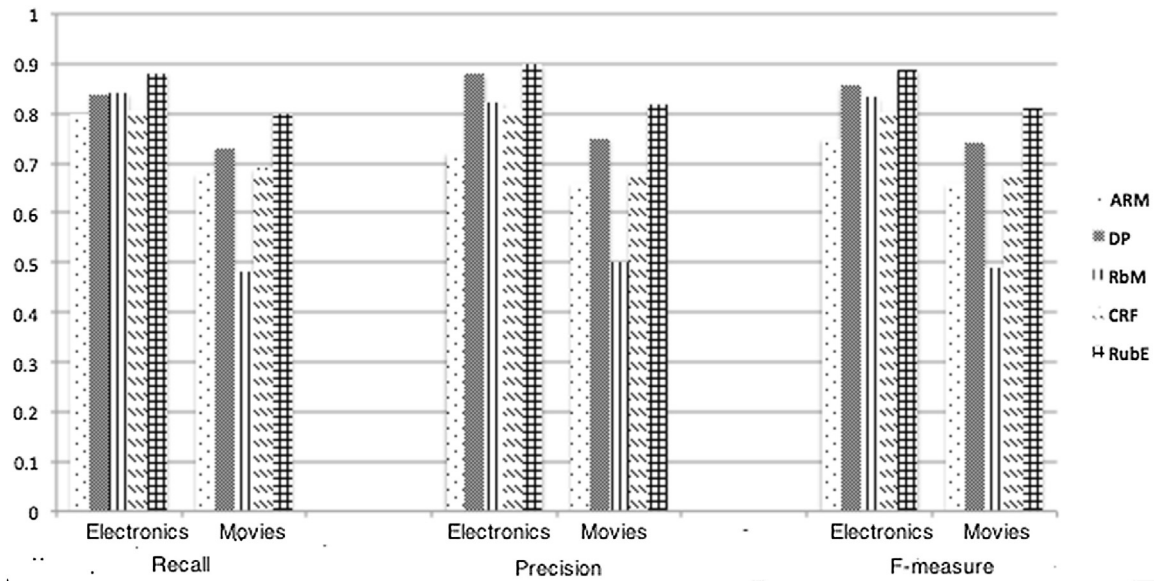


Fig. 3. Performance of all methods on electronic products and movies.

patterns resulted in higher recall. It underscores the value of incorporating unique linguistic characteristics of online review genre for feature extraction. In addition, among the subjective patterns, indirect dependency (31%) was ranked the second highest in recall, which was next to direct dependency using DP (45%). The results highlight the significance of indirect dependency in extracting product features from online reviews.

### 5.2. Theoretical implications

This study makes multifold contributions to the literature. First, RubE improves the state of performance in product feature extraction by extracting both objective and subjective features from online reviews. Our proposed classification scheme of product features can be used to guide further improvement of feature extraction methods. Second, this is the first research that exploited indirect dependency in extracting features from online consumer reviews. The three types of indirect dependency structures explored in this study helped to detect long-distance dependency and complex relationships between words. In addition, this study introduces additional new patterns for extracting subjective features such as comparative constructions. Third, it proposes pruning strategies that consider both semantics and context information, which are not only complementary but also domain-independent. In addition, RubE is developed on the basis of the lexicosyntactic structure rather than pure heuristics, which overcomes the overfitting and low generality problem of RbMs. Finally, in view of significant challenges of extracting features from the movie domain [44], RubE addresses the

challenges by demonstrating superior performance (i.e., 81% in F-measure) to alternative methods.

RubE has significant theoretical implications for text mining and NLP research. First, despite the ongoing trend of applying statistical-based methods in text processing, this study demonstrates that RbMs remain effective in analyzing text in a confined domain such as online consumer reviews. In addition, RbMs can be a promising alternative for those domains that lack annotated datasets. Second, the prevalent use of indirect dependency in online consumer reviews calls for adding corresponding functions to existing dependency analysis tools. Third, the findings of this study suggest that the design of complementary and domain-independent pruning strategies is one way to improve the generality of feature extraction methods.

This research also has implications for the analysis of social media content. RubE can be used to detect sentence subjectivity. In addition, the feature extraction methods proposed here pave the way for feature-based sentiment analysis of online reviews. Further, the feature extraction methods enable fine-grained summarization of online consumer reviews.

### 5.3. Practical implications

The findings of this study have practical implications for consumers, online retailers, and manufacturers. First, RubE improves the usefulness of online reviews for customers in their purchase decision making by enabling feature-based retrieval and summarization. Second, RubE can be used to improve online recommender systems with feature-based personalization, which

Table 7  
Marginal improvement (recall) of each component of RubE.

	Electronics	Movie
DP	0.81	0.73
DP + indirect dependency	+0.03	+0.04
DP + indirect dependency + comparatives	+0.01	+0.01
DP + indirect dependency + comparatives + part-whole	+0.01	+0.01
DP + indirect dependency + comparatives + part-whole + specific patterns	+0.02	+0.02

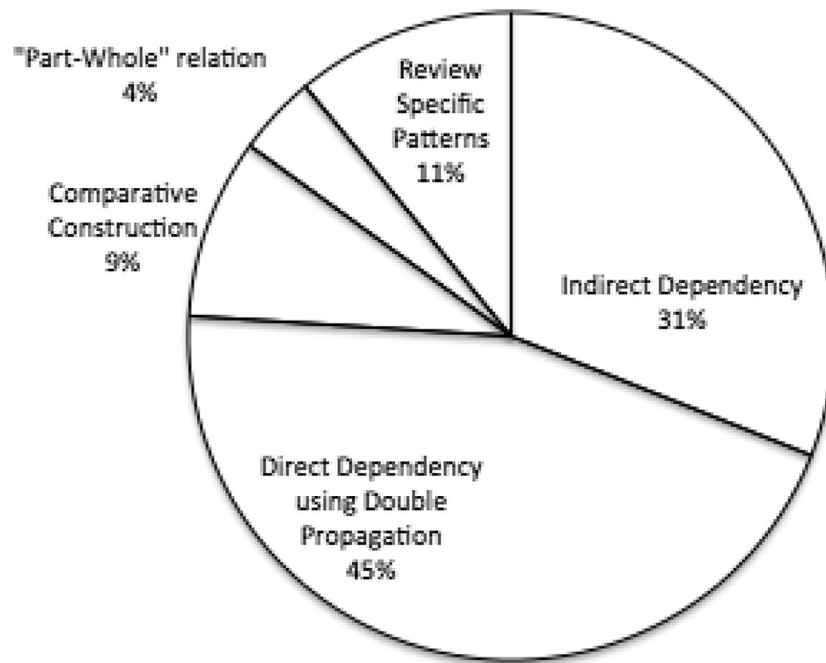


Fig. 4. Percentage distribution in recall for individual components of RubE.

in turn improves customers' adoption and stickiness to retailer's websites. Third, manufacturers may leverage consumer feedback on specific product features to improve product quality. Fourth, RubE can be extended to support topic detection by extracting opinion targets using subjective feature-oriented methods and by extracting the subjects of sentences using the objective feature-oriented methods as candidate topics, and to support ontology learning in discovering concepts and their relationships.

## 6. Conclusion

This research lends strong support to the use of RBMs for analyzing online consumer reviews. Our study demonstrates that the performance of feature extraction can be improved by recognizing different types of features and through design of extraction methods for each type of features separately. The generality of RBMs for feature extraction may be approached by developing domain-independent rules and pruning strategies.

This study exposes several limitations that suggest future research. First, the part-whole patterns may be refined to improve its relatively low recall in extracting objective features. Second, although our indirect dependency relations demonstrate a strong capability of handling long distance relations, some issues such as reference resolution and long distance negation detection need to be addressed in the future. Third, the generality of RubE can be evaluated more fully by using datasets of online reviews from other domains such as hotels and physicians. Finally, the efficacy of RBMs demonstrated in this study may be combined with the strengths of statistical methods to further improve the performance of feature extraction in future.

## Acknowledgments

This research was supported in part by the National Science Foundation (SES-152768). Any opinions, findings or recommendations expressed here are those of the authors and are not necessarily those of the sponsors of this research. The authors also thank Zhenxue Zhang for sharing of the annotated datasets.

## References

- [1] Nikolay Archak, Anindya Ghose, Panagiotis G. Ipeirotis, Deriving the pricing power of product features by mining consumer reviews, *Manage. Sci.* 57 (8) (2011) 1485–1509.
- [2] Satanjeev Banerjee, Ted Pedersen, Extended gloss overlaps as a measure of semantic relatedness, Paper presented at the Ijcai (2003).
- [3] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Aspect extraction with automated prior knowledge learning, Paper presented at the ACL (2014).
- [4] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, Riddhiman Ghosh, Exploiting domain knowledge in aspect extraction, Paper presented at the Emnlp (2013).
- [5] Marie-Catherine De Marneffe, Bill MacCartney, C.D. Manning, Generating typed dependency parses from phrase structure parses, Paper Presented at the Proceedings of LREC (2006).
- [6] Chrysanthos Dellarocas, The digitization of word of mouth: promise and challenges of online feedback mechanisms, *Manage. Sci.* 49 (10) (2003) 1407–1424.
- [7] Angelika Dimoka, Yili Hong, Paul A. Pavlou, On product uncertainty in online markets: theory and evidence, *MIS Q.* 36 (2) (2012) 395–426.
- [8] Rex Yuxing Du, Ye Hu, Sina Damangir, Leveraging trends in online searches for product features in market response modeling, *J. Market.* 79 (1) (2015) 29–43.
- [9] Stefan Gindl, Albert Weichselbraun, Arno Scharl, Rule-based opinion target and aspect extraction to acquire affective knowledge, Paper Presented at the Proceedings of the 22nd International Conference on World Wide Web Companion (2013).
- [10] Roxana Girju, Adriana Badulescu, Dan Moldovan, Automatic discovery of part-whole relations, *Comput. Ling.* 32 (1) (2006) 83–135.
- [11] Graeme Hirst, David St-Onge, Lexical chains as representations of context for the detection and correction of malapropisms, *WordNet: Electr. Lexical Database* 305 (1998) 305–332.
- [12] Mingqing Hu, Bing. Liu, Mining opinion features in customer reviews, Paper presented at the Aaii (2004).
- [13] Mingqing Hu, Bing. Liu, Opinion feature extraction using class sequential rules, Paper presented at the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs (2006).
- [14] Niklas Jakob, Iryna Gurevych, Extracting opinion targets in a single- and cross-domain setting with conditional random fields, Paper Presented at the Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (2010).
- [15] M.L. Jensen, J.M. Averbek, Z. Zhang, K.B. Wright, Credibility of anonymous online product reviews: a language expectancy perspective, *J. Manage. Inf. Syst.* 30 (1) (2013) 293–323, doi:http://dx.doi.org/10.2753/Mis0742-1222300109.
- [16] Wei Jin, Hung Hay Ho, A novel lexicalized HMM-based learning framework for web opinion mining, Paper presented at the Proceedings of the 26th Annual International Conference on Machine Learning (2009).
- [17] Soo-Min Kim, Patrick Pantel, Tim Chklovski, Marco Pennacchiotti, Automatically assessing review helpfulness, Paper Presented at the

- Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (2006).
- [18] Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Extracting aspect-Evaluation and aspect-of relations in opinion mining, Paper Presented at the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (2007).
  - [19] Shi Li, Lina Zhou, Yijun Li, Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures, *Inf. Process. Manage.* 51 (1) (2015) 58–67.
  - [20] X.X. Li, L.M. Hitt, Z.J. Zhang, Product reviews and competition in markets for repeat purchase products, *J. Manage. Inf. Syst.* 27 (4) (2011) 9–41, doi:http://dx.doi.org/10.2753/Mis0742-1222270401.
  - [21] B. Liu, *Web Data Mining*, Springer, 2007.
  - [22] Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, Ming Zhou, Low-quality product review detection in opinion summarization, Paper presented at the EMNLP, Prague (2007).
  - [23] Kang Liu, Liheng Xu, Jun Zhao, Extracting opinion targets and opinion words from online reviews with graph coranking, Paper Presented at the ACL (2014).
  - [24] Edward Loper, Steven Bird, NLTK: The natural language toolkit, Paper Presented at the Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics-Volume 1 (2002).
  - [25] D. Mayzlin, Y. Dover, Chevalier, JA, Promotional reviews: an empirical investigation of online review manipulation, *National Bureau of Economic Research* (2012).
  - [26] Arjun Mukherjee, Bing Liu, Aspect extraction through semisupervised modeling, Paper Presented at the Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 (2012).
  - [27] Ana-Maria Popescu, Oren Etzioni, Extracting Product Features and Opinions from Reviews *Natural Language Processing and Text Mining*, Springer, 2007, pp. 9–28.
  - [28] Soujanya Poria, Erik Cambria, Lun-Wei Ku, Chen Gui, Alexander Gelbukh, A rule-based approach to aspect extraction from product reviews, *SocialNLP 2014* (2014) 28.
  - [29] Soujanya Poria, Erik Cambria, Gregoire Winterstein, Guang-Bin Huang, Sentic patterns: dependency-based rules for concept-level sentiment analysis, *Knowledge-Based Syst.* 69 (2014) 45–63.
  - [30] Katharina Probst, Rayid Ghani, Marko Fano Krema, E. Andrew, Yan Liu, Semi-supervised learning of attribute-value pairs from product descriptions, Paper presented at the Ijcai (2007).
  - [31] Guang Qiu, Bing Liu, Jiajun Bu, Chun Chen, Opinion word expansion and target extraction through double propagation, *Comput. Ling.* 37 (1) (2011) 9–27.
  - [32] Anne Rimrott, Trude Heift, Evaluating automatic detection of misspellings in German, *Lang. Learn. Technol.* 12 (3) (2008) 73–92.
  - [33] Sunita Sarawagi, Information extraction, *Found. Trends Databases* 1 (3) (2008) 261–377.
  - [34] Christopher Scaffidi, Kevin Bierhoff, Eric Chang, Mikhael Felker, Herman Ng, Chun Jin, Red Opal: product-feature scoring from reviews, Paper presented at the Proceedings of the 8th ACM conference on Electronic commerce (2007).
  - [35] Tak-Lam Wong, Wai Lam, Hot item mining and summarization from multiple auction Web sites, Paper presented at the Data Mining, Fifth IEEE International Conference on (2005).
  - [36] Tak-Lam Wong, Wai Lam, Learning to extract and summarize hot item features from multiple auction web sites, *Knowl. Inf. Syst.* 14 (2) (2008) 143–160.
  - [37] Jianan Wu, Yinglu Wu, Jie Sun, Zhilin Yang, User reviews and uncertainty assessment: a two stage model of consumers' willingness-to-pay in online markets, *Decis. Supp. Syst.* 55 (2013) 175–185, doi:http://dx.doi.org/10.1016/j.dss.2013.01.017.
  - [38] F. Xia, C. Cao, Extracting Part-Whole Relations from Online Encyclopedia (2014).
  - [39] Wenting Xiong, Helpfulness-Guided review summarization, Paper Presented at the Hlt-naacl (2013).
  - [40] Liheng Xu, Kang Liu, Siwei Lai, Yubo Chen, Jun Zhao, Mining opinion words and opinion targets in a two-Stage framework, Paper Presented at the Proceedings of the 51 st Annual Meeting of the Association for Computational Linguistics (2013).
  - [41] Bishan Yang, Claire Cardie, Joint inference for fine-grained opinion extraction, Paper presented at the Acl (2013) (1).
  - [42] Lei Zhang, Bing Liu, Suk Hwan Lim, Eamonn O'Brien-Strain, Extracting and ranking product features in opinion documents, Paper Presented at the Proceedings of the 23rd International Conference on Computational Linguistics: Posters (2010).
  - [43] Zhang Zhenxue, Urcf: An Approach to Integrating User Reviews into Memory-based Collaborative Filtering (Doctoral), University of Maryland at Baltimore County, 2013.
  - [44] Lina Zhou, Pimwadee Chaovalit, Ontology-supported polarity mining, *J. Am. Soc. Inf. Sci. Technol.* 59 (1) (2008) 98–110.
  - [45] Lina Zhou, Ping Zhang, Hans-Dieter Zimmermann, Social commerce research: an integrated view, *Electr. Comm. Res. Appl.* 12 (2) (2013) 61–68.

**Yin Kang** is a PhD student in the Department of Information Systems at the University of Maryland, Baltimore County, USA. His research interests focus on natural language processing, web data mining, and machine learning.

**Lina Zhou** is an associate professor of information systems at the University of Maryland, Baltimore County, USA. Her research aims to improve human decision making and knowledge management through the design of intelligent technologies and understanding of human behavior. Her current research interests include deception detection, natural language processing, mobile web adaptation, ontology learning, and online social networks. Dr. Zhou has authored and/or coauthored over 50 referred articles in journals such as *Journal of Management Information Systems*, *MIS Quarterly*, *Communications of the ACM*, *Information & Management*, *IEEE Transactions on Knowledge and Data Engineering*, and *Decision Support Systems*.