Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance

Zhen Hai, Kuiyu Chang, Jung-Jae Kim, and Christopher C. Yang

Abstract—The vast majority of existing approaches to opinion feature extraction rely on mining patterns only from a single review corpus, ignoring the nontrivial disparities in word distributional characteristics of opinion features across different corpora. In this paper, we propose a novel method to identify opinion features from online reviews by exploiting the difference in opinion feature statistics across two corpora, one domain-specific corpus (i.e., the given review corpus) and one domain-independent corpus (i.e., the contrasting corpus). We capture this disparity via a measure called *domain relevance* (DR), which characterizes the relevance of a term to a text collection. We first extract a list of candidate opinion features from the domain review corpus by defining a set of syntactic dependence rules. For each extracted candidate feature, we then estimate its *intrinsic-domain relevance* (IDR) and *extrinsic-domain relevance* (EDR) scores on the domain-dependent and domain-independent corpora, respectively. Candidate features that are less generic (EDR score less than a threshold) and more domain-specific (IDR score greater than another threshold) are then confirmed as opinion features. We call this interval thresholding approach the *intrinsic and extrinsic domain relevance* (IEDR) criterion. Experimental results on two real-world review domains show the proposed IEDR approach to outperform several other well-established methods in identifying opinion features.

Index Terms—Information search and retrieval, natural language processing, opinion mining, opinion feature, Chinese

1 Introduction

Opinion mining (also known as sentiment analysis) aims to analyze people's opinions, sentiments, and attitudes toward entities such as products, services, and their attributes [1]. Sentiments or opinions expressed in textual reviews are typically analyzed at various resolutions. For example, document-level opinion mining identifies the overall subjectivity or sentiment expressed on an entity (e.g., cellphone or hotel) in a review document, but it does not associate opinions with specific aspects (e.g., display, battery) of the entity. This problem also happens, though to a lesser extent, in sentence-level opinion mining, as shown in Example 1.1.

Example 1.1. "The exterior is very beautiful, also not expensive, though the battery is not very durable, I still firmly recommend this cellphone."

Although Example 1.1 expresses an overall positive opinion on the cellphone, it also contains conflicting opinions associated with different attributes or aspects of the cellphone. The opinion orientations for the "cellphone" itself and its "exterior" are positive, but the opinion polarity for the aspect of "battery" is negative. Such fine-grained opinions may very well tip the balance in purchase decisions. Savvy consumers nowadays are no longer satisfied with just the overall opinion rating of a product. They want to understand why it receives the rating, that is, which positive or negative attributes or aspects contribute to the final rating of the product. It is, thus, important to extract the specific opinionated features from text reviews and associate them to opinions.

In opinion mining, an *opinion feature*, or *feature* in short, indicates an entity or an attribute of an entity on which users express their opinions. In this paper, we propose a novel approach to the identification of such features from unstructured textual reviews.

A good many approaches have been proposed to extract opinion features in opinion mining. Supervised learning model may be tuned to work well in a given domain, but the model must be retrained if it is applied to different domains [2], [3]. Unsupervised *natural language processing* (NLP) approaches [4], [5], [6] identify opinion features by defining domain-independent syntactic templates or rules that capture the dependence roles and local context of the feature terms. However, rules do not work well on colloquial real-life reviews, which lack formal structure. Topic modeling approaches can mine coarse-grained and generic topics or aspects, which are actually semantic feature clusters or aspects of the specific features commented on explicitly in reviews [7], [8]. Existing corpus statistics approaches try to extract opinion features by mining statistical patterns of feature terms only in the given review corpus, without considering their distributional characteristics in another different corpus [10], [11].

One key finding of our work is that the distributional structure of an opinion feature in a given domain-dependent

- *Z. Hai, K. Chang, and J.J. Kim are with the School of Computer Engineering, Nanyang Technological University, N4-B3C-14 DISCO lab, 50 Nanyang Avenue, Singapore 639798, Singapore.*
  E-mail: {haiz0001, askychang, jungjae.kim}@ntu.edu.sg.
- *C.C. Yang is with the College of Information Science and Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104-2875.*
  E-mail: Chris.Yang@drexel.edu.

Manuscript received 6 June 2012; revised 30 Nov. 2012; accepted 13 Jan. 2013; published online 31 Jan. 2013.
Recommended for acceptance by H. Wang.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2012-06-0395.
Digital Object Identifier no. 10.1109/TKDE.2013.26.

review corpus,[1] for example, cellphone reviews, is different from that in a domain-independent corpus. For instance, the opinion feature "battery" tends to be mentioned quite frequently in the domain of *cellphone* reviews, but not as frequently in the domain-irrelevant *Culture* article collection. This leads us to propose a novel method to identify opinion features by exploiting their distribution disparities across different corpora. Specifically, we proposed and evaluated the *domain relevance* (DR) of an opinion feature across two corpora. The DR criterion measures how well a term is statistically associated with a corpus.

Our method is summarized as follows: First, several syntactic dependence rules are used to generate a list of candidate features from the given domain review corpus, for example, cellphone or hotel reviews. Next, for each recognized feature candidate, its domain relevance score with respect to the domain-specific and domain-independent corpora is computed, which we termed the *intrinsic-domain relevance* (IDR) score, and the *extrinsic-domain relevance* (EDR) score, respectively. In the final step, candidate features with low IDR scores and high EDR scores are pruned. We, thus, call this interval thresholding the *intrinsic and extrinsic domain relevance* (IEDR) criterion. Evaluations conducted on two real-world review domains demonstrate the effectiveness of our proposed IEDR approach in identifying opinion features.

2 Related Work

2.1 Opinion Mining

Opinions and sentiments expressed in text reviews can be generally analyzed at the document, sentence, or even phrase (word) levels. The purpose of document-level (sentence-level) opinion mining is to classify the overall subjectivity or sentiment expressed in an individual review document (sentence).

Hatzivassiloglou and Wiebe [12] studied the effects of dynamic adjectives, semantically oriented adjectives, and gradable adjectives on predicting subjectivity; they proposed a supervised classification method to predict sentence subjectivity. Pang et al. [13] proposed three machine learning methods, naive Bayes, maximum entropy, and support vector machines, to classify whole movie reviews into positive or negative sentiments. They found that standard machine learning techniques produced good results in comparison to human-generated baselines. Moreover, machine learning methods did not perform as well on sentiment classification as on traditional topic-based categorization. To prevent a sentiment classifier from considering irrelevant or even potentially misleading text, Pang and Lee [14] proposed to first employ a sentence-level subjectivity detector to identify the sentences in a document as either subjective or objective, and subsequently discarding the objective ones. They then applied the sentiment classifier to the resulting subjectivity extract, with improved results.

Mcdonald et al. [15] investigated the use of a global structured model that learns to predict sentiments on different levels of granularity for a textual review. The primary advantage of the proposed model is that it allows classification decisions from one level in the text to influence decisions at another. A regression method based on the bag-of-opinions model was proposed for review rating prediction from sparse text patterns [16]. Review rating estimation is a much more complicated problem compared to binary sentiment classification. Generally, sentiments are expressed differently in different domains. The sentiment classification methods discussed above can be tuned to work very well on a given domain; however, they may fail in classifying sentiments in a different domain. Bollegala et al. [17] proposed a cross-domain sentiment classifier using an automatically extracted sentiment thesaurus.

An unsupervised learning method was proposed to classify review documents as thumbs up (positive) or thumbs down (negative) in [18]. The sentiment of each review document is predicted by the average sentiment orientations of phrases in the review. Domain-dependent contextual information is also considered for better estimation of the phrase sentiments. One limitation of this work is its reliance on an external search engine. Zhang et al. [19] proposed a rule-based semantic analysis approach to classify sentiments for text reviews. They used word dependence structures to classify the sentiment of a sentence, and predicted document-level sentiments via aggregating the sentence sentiments. Rule-based approaches like this typically suffer from poor coverage due to the lack of comprehensiveness in their rules. In addition, Maas et al. [20] presented an approach to document-level and sentence-level sentiment classification tasks, which uses a mix of unsupervised and supervised techniques to learn word vectors by capturing semantic term-document information as well as rich sentiment content.

Differently, sentiment analysis at the phrase (word) level mainly focuses on classifying sentiment polarities of opinion phrases (words). Generally, the sentiment polarity of an opinion word is usually context-dependent as well as domain-specific. Wilson et al. [21] presented an approach to predicting contextual sentiments at the phrase level by applying machine learning techniques on a variety of feature factors. Yessenalina and Cardie [22] presented a compositional matrix-space model for phrase-level sentiment analysis. One of the benefits of the proposed approach is that by learning matrices for words, the model can handle unseen word compositions (e.g., unseen bigrams) as long as the component unigrams have been learned. A two-level affective reasoning method was proposed to mimic the integration of conscious and unconscious reasoning to address word-level sentiment analysis tasks [23].

Note that opinion mining at the document, sentence, or phrase (word) level does not discover what exactly people liked and disliked in reviews. In other words, it fails to associate the identified sentiments to the corresponding features commented on in the reviews. Clearly, an extracted opinion without the corresponding feature (opinionated target) is of limited value in reality [1]. Next, we survey existing work on extracting opinion features.

2.2 Opinion Feature Extraction

Opinion feature extraction is a subproblem of opinion mining, with the vast majority of existing work done in the product review domain. Previous approaches can be

[1] Note that in this work the given review corpus is termed to be domain-dependent, while any other corpus irrelevant to the review domain is called domain-independent.

roughly classified into two categories, namely, supervised and unsupervised.

By formulating opinion mining as a joint structural tagging problem, supervised learning models including hidden Markov models and conditional random fields have been used to tag features or aspects of commented entities [2], [24]. Supervised models may be carefully tuned to perform well on a given domain, but need extensive retraining when applied to a different domain, unless transfer learning is adopted [25]. In addition, a decent-sized set of labeled data is generally needed for model learning on every domain.

Unsupervised NLP approaches extract opinion features by mining syntactic patterns of features implied in review sentences. In particular, the approaches attempt to discover syntactic relations among feature terms and opinion words in sentences by using carefully crafted syntactic rules [5], [6] or semantic role labeling [4]. Syntactic relations identified by the methods help locate features associated with opinion words, but could also inadvertently extract large number of invalid features due to the colloquial nature of online reviews.

Unsupervised corpus statistics approaches use the results of statistical analysis on a given corpus to understand the distributional characteristics of opinion features. The approaches are somewhat resistance to the colloquial nature of online reviews given a suitably large review corpus. Hu and Liu [10] proposed an *association rule mining* (ARM) approach to mine frequent itemsets as potential opinion features, which are nouns and noun phrases with high sentence-level frequency (or support). However, ARM, which relies on the frequency of itemsets, has the following limitations for the task of feature identification, 1) frequent but invalid features are extracted incorrectly, and 2) rare but valid features may be overlooked.

To address feature-based opinion mining problems, Su et al. [34] introduced a *mutual reinforcement clustering* (MRC) approach to mine the associations between feature categories and opinion word groups, based on a cooccurrence weight matrix generated from the given review corpus. Unlike several other corpus statistics methods, MRC is able to extract infrequent features, provided that the mutual relationships between feature and opinion groups found during the clustering phase is accurate. However, MRC's precision is low due to the difficulty in obtaining good clusters on real-life reviews.

Yu et al. [26] proposed an aspect ranking algorithm based on the probabilistic regression model to identify important product aspects from online consumer reviews. Moreover, their focus is not on extracting feature terms commented on explicitly in reviews, but rather on ranking product aspects that are actually coarse-grained clusters of specific features.

Unsupervised topic modeling approaches, such as *latent Dirichlet allocation* (LDA) [7], which is a generative three-way (term-topic-document) probabilistic model, have been used to solve aspect-based opinion mining tasks. The models are developed primarily for mining latent topics or aspects, which actually correspond to distinguishing properties or concepts of the commented entities, and may not necessarily be opinion features expressed explicitly in reviews [7], [8], [27], [9]. For example, "where" could be a valid LDA topic concept associated with cellphone reviews, since some users like to discuss about cellphone vendors, but it is not a specific opinionated cellphone feature. Therefore, though the approaches are effective in discovering latent structures of review data, they may be less successful in dealing with identifying specific feature terms commented on explicitly in reviews. In our experiments, we find LDA to perform reasonably well as an unsupervised method, but it still fall short of our proposed approach.

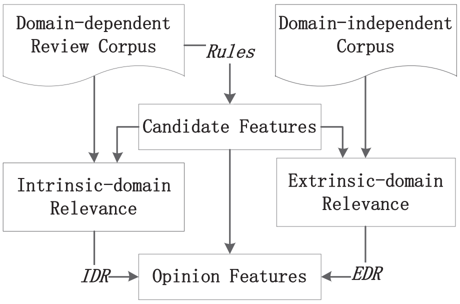

Domain-dependent Review Corpus
Rules
Domain-independent Corpus
Candidate Features
Intrinsic-domain Relevance
Extrinsic-domain Relevance
IDR
Opinion Features
EDR

Fig. 1. IEDR workflow.

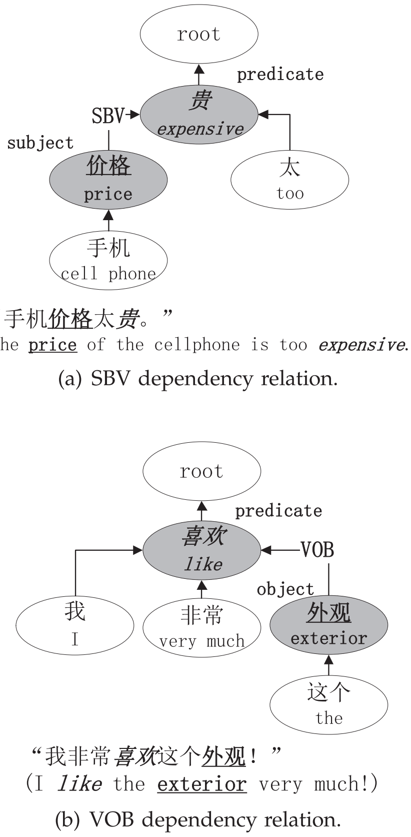
As discussed, existing approaches to feature extraction typically only use the knowledge or patterns mined from a given single review corpus, while completely ignoring the possible variations present in a different domain-independent corpus, for example, a corpus of documents on *Culture*. Our proposed IEDR approach utilizes the fact that word distribution characteristics vary across different types of corpora, in particular domain-specific versus domain-independent, to derive powerful hints that help discriminate valid features from the invalid ones. In the first step of our approach, we define some syntactic dependence rules to extract candidate features, similar to NLP approaches. In the second step, we employ the IEDR measures to identify the desired domain-specific opinion features. The key difference of IEDR compared to existing methods lies in its smart fusion of domain-dependent and domain-independent information sources.

3 Methodology

3.1 Overview

An opinion feature such as "screen" in cellphone reviews is typically domain-specific. That is, the feature appears frequently in the given review domain, and rarely outside the domain such as in a domain-independent corpus about *Culture*. As such, domain-specific opinion features will be mentioned more frequently in the domain corpus of reviews, compared to a domain-independent corpus.

Fig. 1 shows the workflow of our proposed method. Given a domain-dependent review corpus and a domain-independent corpus, we first extract a list of candidate features from the review corpus via manually defined syntactic rules (denoted "Rules" in the figure). For each extracted candidate feature, we estimate its IDR, which represents the statistical association of the candidate to the given domain corpus, and *extrinsic-domain relevance*, which reflects the statistical relevance of the candidate to the domain-independent corpus. Only candidates with IDR scores exceeding a predefined intrinsic relevance threshold and EDR scores less than another extrinsic relevance



(a) SBV dependency relation.

(b) VOB dependency relation.

Fig. 2. Sample dependence trees for (a) SBV and, (b) VOB syntactic relations. Highlighted (underlines) nouns are validated opinion features.

threshold are confirmed as valid opinion features. In short, we identify opinion features that are domain-specific and at the same time not overly generic (domain-independent) via the intercorpus statistics IEDR criterion.

3.2 Candidate Feature Extraction

Intuitively, opinion features are generally nouns or noun phrases, which typically appear as the subject or object of a review sentence.

In the case of dependence grammar [28], the subject opinion feature has a syntactic relationship of type *subject-verb* (SBV) with the sentence predicate (usually adjective or verb). The object opinion feature has a dependence relationship of *verb-object* (VOB) on the predicate. In addition, it also has a dependence relationship of *preposition-object* (POB) on the prepositional word in the sentence. Some syntactic relation examples in Chinese are listed in Figs. 2a and 2b, with their corresponding dependence trees. The letter "V" in both SBV and VOB in the figure indicates the predicate of a review sentence.

In particular, as shown in the dependence tree in Fig. 2a, the opinion feature "price" (underline), which is associated with the adjective "expensive" (italic), is the subject of the sentence. It has a dependence relation of SBV with the adjective predicate. In Fig. 2b, the noun feature "exterior" is the object of the verb predicate "like," and thus has a VOB dependence relation with the predicate.

From the aforementioned dependence relations, i.e., SBV, VOB, and POB, we present three syntactic rules in Table 1, where "NN" and "CF" denote nouns (noun phrases) and candidate features, respectively. For example, by employing the first rule in Table 1 to the example, we can extract the noun "price" as a candidate feature, as shown in Fig. 2a, which has an SBV relation with the adjective predicate "expensive."

TABLE 1
Syntactic Rules

Rules	Interpretation
$NN + SBV \rightarrow CF$	Identify NN as a CF, if NN has a SBV dependency relation
$NN + VOB \rightarrow CF$	Identify NN as a CF, if NN has a VOB dependency relation
$NN + POB \rightarrow CF$	Identify NN as a CF, if NN has a POB dependency relation

The candidate feature extraction process works in the following steps: 1) Dependence parsing (DP) is first employed to identify the syntactic structure of each sentence in the given review corpus; 2) the three rules in Table 1 are applied to the identified dependence structures, and the corresponding nouns or noun phrases are extracted as candidate features whenever a rule is fired. Our candidate feature extraction method is language dependent, in this case it is based on the Chinese language. But it is not a serious problem, since we can similarly define such simple extraction rules in other different languages.

There could be many invalid features in the extracted candidate feature list, the next step is to prune the list via the proposed IEDR criterion.

3.3 Opinion Feature Identification

3.3.1 Domain Relevance

Domain relevance characterizes how much a term is related to a particular corpus (i.e., a domain) based on two kinds of statistics, namely, dispersion and deviation.

Dispersion quantifies how significantly a term is mentioned across all documents by measuring the distributional significance of the term across different documents in the entire corpus (horizontal significance).

Deviation reflects how frequently a term is mentioned in a particular document by measuring its distributional significance in the document (vertical significance).

Both dispersion and deviation are calculated using the well-known *term frequency-inverse document frequency* (TF-IDF) term weights. Each term T_i has a term frequency TF_{ij} in a document D_j, and a global document frequency DF_i. The weight w_{ij} of term T_i in document D_j is then calculated as follows:

$$w_{ij} = \begin{cases} (1 + log\,TF_{ij}) \times log\frac{N}{DF_i} & \text{if } TF_{ij} > 0, \\ 0, & otherwise, \end{cases} \tag{1}$$

where $i = 1, \ldots, M$ for a total number of M terms, and $j = 1, \ldots, N$ for a total number of N documents in the corpus.

The standard variance s_i for term T_i is calculated as follows:

$$s_i = \sqrt{\frac{\sum_{j=1}^{N}(w_{ij} - \bar{w}_i)^2}{N}}, \tag{2}$$

where the average weight $\bar{w}_i$ of term T_i across all documents is calculated by

$$\bar{w}_i = \frac{1}{N}\sum_{j=1}^{N} w_{ij}.$$

The dispersion $disp_i$ of each term T_i in the corpus is defined as follows:

$$disp_i = \frac{\bar{w}_i}{s_i}. \tag{3}$$

Dispersion thus measures the normalized average weight of term T_i. It is high for terms that appears frequently across a large number of documents in the entire corpus.

The deviation $devi_{ij}$ of term T_i in document D_j is given by

$$devi_{ij} = w_{ij} - \bar{w}_j, \tag{4}$$

where the average weight $\bar{w}_j$ in the document D_j is calculated over all M terms as follows:

$$\bar{w}_j = \frac{1}{M}\sum_{i=1}^{M} w_{ij}.$$

Deviation $devi_{ij}$ indicates the degree in which the weight w_{ij} of the term T_i deviates from the average $\bar{w}_j$ in the document D_j. The deviation thus characterizes how significantly a term is mentioned in each particular document in the corpus.

The domain relevance dr_i for term T_i in the corpus is finally defined as follows:

$$dr_i = disp_i \times \sum_{j=1}^{N} devi_{ij}. \tag{5}$$

Clearly, the domain relevance dr_i incorporates both horizontal (dispersion $disp_i$) and vertical (deviation $devi_{ij}$) distributional significance of term T_i in the corpus. The domain relevance score thus reflects the ranking and distributional characteristics of a term in the entire corpus. Note that the domain relevance scores for some terms can be negative, which indicates a relatively weaker association.

Our domain relevance criterion is actually inspired by the domain dependence measure used for event tracking in the topic detection and tracking (TDT) field [29]. However, our domain dependence criterion is different on two counts as follows:

- Domain dependence was introduced to track news events by discriminating topic words from event words expressed in the news stories. In our work, we do not distinguish between topic and event words. Instead, we simply employ the proposed domain relevance as a measure to identify opinion features from unstructured text reviews.
- Our domain relevance formula is tailored for measuring inter-corpus statistics disparity; Specifically, it is tuned to capture the distributional disparities of opinion features across two corpora.

3.3.2 Intrinsic and Extrinsic Domain Relevance

The domain relevance of an opinion feature, which is computed on a domain-specific review corpus, is called *intrinsic-domain relevance*. Likewise, the domain relevance of the same opinion feature computed on a domain-independent corpus is called *extrinsic-domain relevance*. IDR reflects the specificity of the feature to the domain review corpus (e.g., cellphone reviews), while EDR characterizes the statistical association of the feature to the domain-independent or generic corpus. Intuitively, a candidate term is relevant to either one or the other, but not both. As such, EDR also characterizes the irrelevance of a feature to the given domain review corpus.

Granted, there do exist some relatively common terms that are used almost everywhere and also in a review corpus as features. For example, the term "price" usually appears as a feature in many review domains, such as cellphone and hotel reviews. Therefore, the success of our approach boils down to the careful selection of a domain-independent corpus that is as distinct from the domain-specific review corpus as possible. Section 4.5 provides some guidelines and experimental results on good domain-independent corpus selection.

The procedure for computing the domain relevance is the same regardless of the corpus, as summarized in Algorithm 1. When the procedure is applied to the domain-specific review corpus, the scores are called IDR, otherwise they are called EDR.

Algorithm 1: Calculating Intrinsic/Extrinsic Domain Relevance (IDR/EDR)

```
Input: A domain specific/independent corpus $C$
Output: Domain relevance scores (IDR or EDR)

for each candidate feature $CF_i$ do
    for each document $D_j$ in the corpus $C$ do
        Calculate weight $w_{ij}$ by (1);
    Calculate standard deviation $s_i$ by (2);
    Calculate dispersion $disp_i$ by (3);
    for each document $D_j$ in the corpus $C$ do
        Calculate deviation $devi_{ij}$ by (4);
    Compute domain relevance $dr_i$ by (5);
return A list of domain relevance (IDR/EDR)
scores for all candidate features;
```

Candidate features with overly high EDR scores or miserably low IDR scores are pruned using the intercorpus criterion of IEDR. Algorithm 2 summarizes the proposed IEDR approach, where the minimum IDR threshold *ith* and maximum EDR threshold *eth* can be determined experimentally. A sample run of the IEDR algorithm on a toy example is given in Example 3.1 and Table 2.

Example 3.1. "The screen of Iphone5 looks really beautiful, and its battery is okay for me. I am one of its many fans and I really want to have one, but it is too expensive, and I have no money now!"

Example 3.1 shows a sample product review on iPhone 5. Here both nouns "screen" and "battery" are annotated as true opinion features (with associated opinions). Applying Algorithm 2 on the example as follows: First, apply the syntactic rules (Rules in short) defined in Table 1 to extract a list of candidate features (nouns): "screen," "battery," "fans," and "money," as shown in line 1 of Table 2. Next, filter the four candidates using IEDR, to obtain the final

TABLE 2
Extracted Opinion Features via Different Methods

Rules	screen	battery	fans	money
IEDR	**screen**	**battery**		
IDR	screen	battery		money
EDR	screen	battery	fans	

TABLE 3
Domain Review Corpora Statistics

Corpus/Topic	# Reviews	# Sentences
Cellphone	10,073	15,771
Hotel	6,313	23,636

TABLE 4
Domain-Independent Corpora Statistics

Corpus/Topic	# Documents	# Sentences
Culture	8,000	490,354
Sports	8,000	301,653
Tourism	8,000	291,313
Finance	8,000	323,065
Employment	8,000	359,617
Education	8,000	483,271
IT	8,000	201,944
Health	8,000	329,957
Military	8,000	245,539
Automobile	8,000	253,554

confirmed set of opinion features: "screen" and "battery," as shown in line 2 of Table 2.

Algorithm 2: Identifying Opinion Features via IEDR

```
Input: Domain review corpus $R$ and
       domain-independent corpus $D$
Output: A validated list of opinion features

Extract candidates from the review corpus $R$;
for each candidate feature $CF_i$ do
    Compute IDR score $idr_i$ via Algorithm 1 on
    the review corpus $R$;
    Compute EDR score $edr_i$ via Algorithm 1 on
    the domain-independent corpus $D$;
    if $(idr_i \geq ith)$ $AND$ $(edr_i \leq eth)$ then
        Confirm candidate $CF_i$ as a feature;
return A validated set of opinion features;
```

For comparison, we also listed the extracted opinion features when only one of the two measures is used, as shown in line 3 (IDR only) and 4 (EDR only) of Table 2. Using IDR, "fans" is not domain-specific enough, so "fans" is pruned. Using EDR, "money" is too generic, so "money" is pruned. IEDR combines both thresholds to prune both "fans" and "money," resulting in two correct features.

4 Experiments

We have incorporated the IEDR feature extraction into an existing opinion mining system named iMiner [30], and thus far evaluated its performance using real-world Chinese reviews from two different domains, i.e., cellphones and hotels.

4.1 Corpus Description

The cellphone review corpus contains 10,073 real-life text reviews collected from a major Chinese forum website.[2] The hotel review corpus contains 6,313 reviews crawled from a famous Chinese travel portal.[3] Summary statistics of the two domain review corpora are shown in Table 3. Hotel reviews are twice as long as cellphone reviews on average.

We randomly selected 508 documents from the cellphone review corpus for annotation. Two persons manually annotated opinion feature(s) expressed in every review sentence in each of the 508 documents. An annotated opinion feature is considered valid if and only if both annotators highlight it. If only one of the annotators mark an opinion feature, then a third person has a final say on whether to keep or reject it. A total of 995 opinion features were annotated from the 508 cellphone review documents. Using the same method, we annotated 1,013 opinion features from 206 randomly selected hotel review documents. The Kappa coefficients, a quantitative measure of the magnitude of interannotator agreement, are 0.66 and 0.62 for cellphone and hotel reviews, respectively. Kappa values within 0.6 to 0.8 generally indicate substantial agreement [32].

We also collected 10 domain-independent (generic) corpora from a Chinese website,[4] with each corpus containing 8,000 documents. The collected corpora cover domain-irrelevant heterogeneous topics spanning *Culture*, *Sports*, *Tourism* and so on. Summary statistics of the 10 domain-independent corpora are shown in Table 4.

All documents from the domain review corpora as well as the domain-independent corpora were parsed using the *language technology platform* (LTP) [31], a Chinese natural language analyzer.

4.2 Experiment Design

We conducted various experiments to comprehensively evaluate the IEDR performance on two real-world review domains, cellphone and hotel reviews. We first evaluated IEDR performance against the competition using precision versus recall curves. We then measured the effect of domain-independent corpus size and topic. Since the selection of IDR and EDR thresholds is important, we measure IEDR performance versus various thresholds. Finally, we plugged features extracted via IEDR into a sentiment classifier to see how our extracted features can improve the overall performance of feature-based opinion classification.

We compared the proposed IEDR to several opponent methods as follows:

1. *Intrinsic-domain relevance (IDR)*, which uses only the given review corpus to extract opinion features,

[2] product.tech.163.com.

[3] www.lvping.com/hotels.

[4] www.sogou.com/labs.

TABLE 5
Summary of Evaluated Methods

Method	Characteristics	Corpus
IEDR	Intrinsic and extrinsic domain relevance criterion	Review and domain-independent
IDR	Intrinsic-domain relevance	Review
EDR	Extrinsic-domain relevance	Domain-independent
LDA	Topic modeling	Review
ARM	Frequent itemset mining	Review
MRC	Mutual reinforcement principle	Review
DP	Dependency parsing	Review

2. *Extrinsic-domain relevance (EDR)*, which uses only the domain-independent corpus to extract opinion features,
3. Latent Dirichlet allocation (LDA) [7], which is a generative probabilistic graphical topic model,
4. *Association rule mining (ARM)* [33], which mainly discovers frequent nouns or noun phrases as opinion features,
5. Mutual reinforcement clustering (MRC) [34], and
6. Dependency parsing (DP) [5], which uses synthetic rules to extract features.

Table 5 summarizes the methods we evaluated.

4.3 Precision versus Recall

We first extracted candidate features from the given review domains, i.e., cellphone and hotel reviews, using the syntactic rules defined in Table 1.

Based on the same set of candidates, we compared IEDR to both IDR and EDR on the cellphone review domain. The precision-recall curves for IEDR, IDR, and EDR are plotted as solid lines in Fig. 3. Note that the best performing *Culture* corpus was selected as the domain-independent corpus for both IEDR and EDR.

In Fig. 3, the IEDR curve lies well above the IDR curve for all but the two lowest recall levels. This is perfectly acceptable since precision values at high recall levels are more practical. Across all recall levels, the largest precision gap of IEDR over IDR is 11.90 percent (located at 0.55 recall). At recall rates larger than 0.5, the best IEDR precision is 91.67 percent, which is 10.68 percent higher than the best IDR precision. The IEDR curve lies largely above that of the EDR for all recall levels, and the best IEDR precision is 23.42 percent higher than that of EDR, for recall rates larger than 0.5. The Proposed IEDR thus achieved a significant improvement over either IDR or EDR.

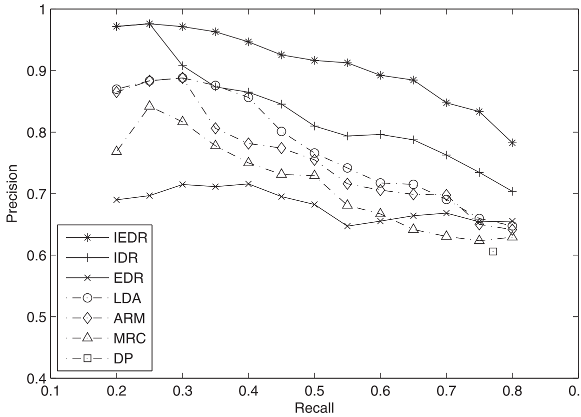


Fig. 3. Precision-recall curves for cellphone feature extraction. Results are generated by plotting precision at each of 13 recall levels (approximately 5 percent apart). Only one precision-recall value is available for DP, which has no tuning parameters.

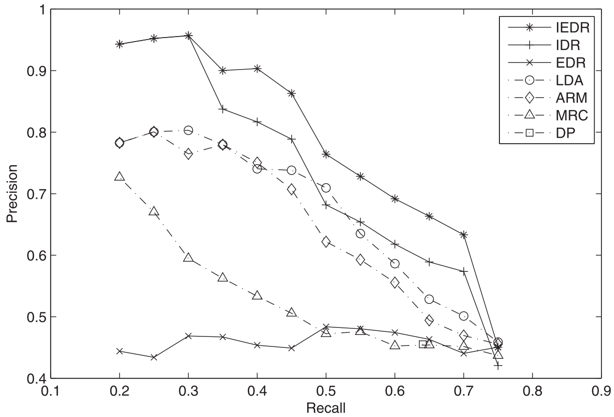


Fig. 4. Precision-recall curves for hotel feature extraction. Results are obtained by estimating the precision at each of the 12 recall levels (approximately 5 percent apart). One precision-recall pair is shown for DP.

Next, we benchmarked IEDR against four existing major methods LDA, ARM, MRC, and DP (see details in Table 5). The full precision-recall curves of LDA, ARM, and MRC are shown in dashed lines in Fig. 3, and only one precision-recall value, indicated by the square, is available for DP.

In Fig. 3, the IEDR curve lies very well above LDA, ARM, and MRC curves as well as the DP point for all recall levels. The best IEDR precision is 91.67 percent for recall rates higher than 0.5, which is 15.06, 16.18, 18.76, and 31.08 percent better than the best precision for LDA, ARM, MRC, and DP methods, respectively. The experimental results demonstrated the effectiveness of our proposed IEDR approach on the cellphone review domain.

We further evaluated the IEDR feature extraction performance on a different domain, hotel reviews. The precision-recall curves of IEDR, IDR, EDR, LDA, ARM, and MRC are shown in Fig. 4. Again, there is only one precision-recall pair shown for the DP method. Here we used the same domain-independent corpus of *Culture* for IEDR and EDR.

In Fig. 4, IEDR precision-recall curve again lies well above other curves for almost all recall levels. The largest precision gain of IEDR over IDR is 8.68 percent (at 0.40 recall) across all recall levels. For recall rates larger than 0.5, the best IEDR precision is 76.41 percent, which is 8.22, 28.05, 5.47, 14.23, 28.84, and 30.90 percent better than that of IDR, EDR, LDA, ARM, MRC, and DP, respectively. IEDR again achieved much improved feature extraction performance compared to all other well-established competing methods on the hotel domain.

LDA, as an effective topic modeling technique, is designed mainly for mining latent topics or structures of a corpus of documents. It is, thus, less successful in dealing with the opinion feature extraction task. ARM may be effective in discovering frequent opinion features.

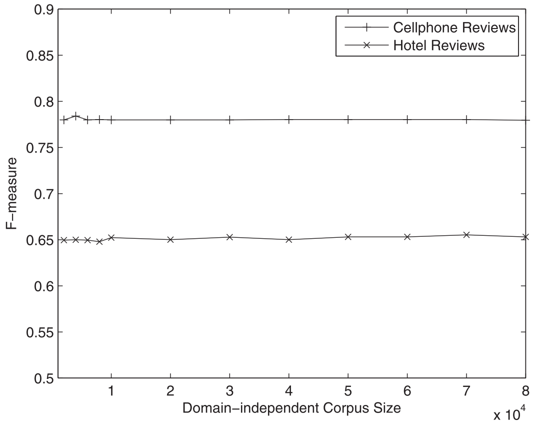


Fig. 5. IEDR feature extraction results versus domain-independent corpus size.

However, the method is less effective in pruning frequent but invalid candidate features, as well as identifying infrequent but valid opinion features. MRC may be effective in grouping similar features (or opinions) together into clusters, but ironically it depends on good clustering for good results, which is hard to optimize in practice. DP is domain-independent as it uses syntactic rules to extract opinion features, but it cannot deal with colloquial languages expressed in reviews.

Most importantly, LDA, ARM, MRC, DP, as well as both IDR and EDR[5] all operate on a single corpus. In contrast, our IEDR approach makes full use of the distributional disparities of features across different corpora, to achieve significantly better feature extraction results.

In practice, given a domain review corpus, for example, cellphone reviews, it is nontrivial to collect a high-quality domain-independent corpus to achieve a decent IEDR feature extraction performance. Specifically, How big should the domain-independent corpus be? Which domain-independent topic(s) can lead to improved performance? How many topics should it contain? In the following sections, we investigate the effects of size and topic of the domain-independent corpus on feature identification, in an attempt to shine some light on the above questions.

4.4 Size of Domain-Independent Corpus

To evaluate the effect of (domain-independent) corpus size on IEDR feature extraction, we generated 12 new corpora using stratified sampling from the 10 original domain-independent corpora (as shown in Table 4). The sizes (in thousands) evaluated are 2, 4, 6, 8, 10, 20, 30, 40, 50, 60, 70, 80. Each corpus contains the same distribution characteristics of the 10 domain-independent topics.

Fig. 5 gives the IEDR F-measure performance versus the size of each of the 12 new domain-independent corpus. As shown in the curve on the cellphone review domain, there are no big changes in F-measure performance across all 12 sizes. Similarly, the curve on the hotel reviews demonstrates that the IEDR F-measure performance does not vary significantly from 2,000 to 80,000 number of generic documents.

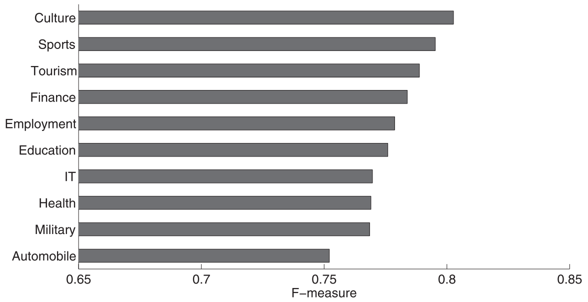


Fig. 6. IEDR performance on cellphone reviews versus choice of domain-independent corpus/topic. Topics are ranked in descending order of F-measure.

Therefore, it can be concluded that the size of the domain-independent corpus does not significantly affect IEDR performance for both cellphone and hotel review domains. In practice, using a domain-independent corpus with more than 10,000 documents is sufficient to reap the benefits of the IEDR approach.

4.5 Choice of Domain-Independent Corpus

4.5.1 Single Topic in Domain-Independent Corpus

We first evaluated IEDR performance in terms of F-measure against each of the 10 original domain-independent corpora (each corpus covers only one topic). The 10 domain-independent topics are ranked in descending order of their F-measures on the two review domains, respectively.

As shown in Fig. 6 on the cellphone domain, IEDR performance varies across different domain-independent topics. Specifically, IEDR based on the *Culture* corpus achieved the best F-measure of 80.27 percent, which is 5.06 percent better than the worst performer, which is based on *Automobile*.

Interestingly, the *IT* corpus performed badly (at seventh place). This is expected because *IT* is closely related to the given cellphone review domain.

From Fig. 7, the *Culture* corpus again attained the best F-measure of 66.23 percent, which is almost 3 percent better than the worst *Automobile* corpus (63.36 percent F-measure).

The *Culture* corpus performed the best in both cases because it is the most topically-distinct from both cellphone and hotel domains. The *Sports* and Tourism topics ranked second and third for cellphone reviews, but ranked

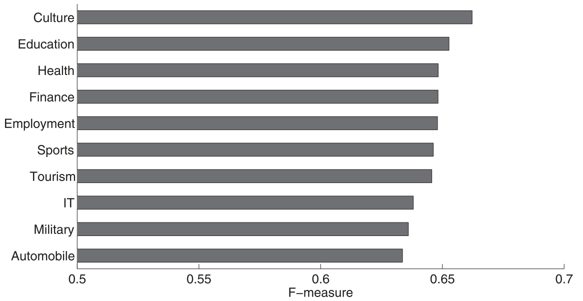


Fig. 7. IEDR performance on hotel reviews versus choice of domain-independent corpus/topic. Topics are ranked in descending order of F-measure.

[5] EDR utilizes only one corpus, namely the domain-independent corpus, for pruning. Therefore, it qualifies as a single-corpus approach.

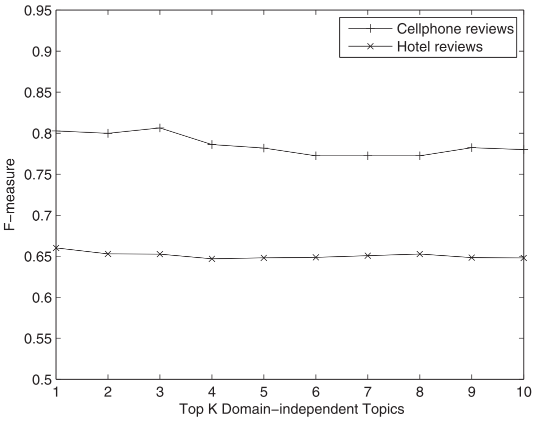


Fig. 8. IEDR performance for cellphone and hotel reviews versus top-K domain-independent topics.

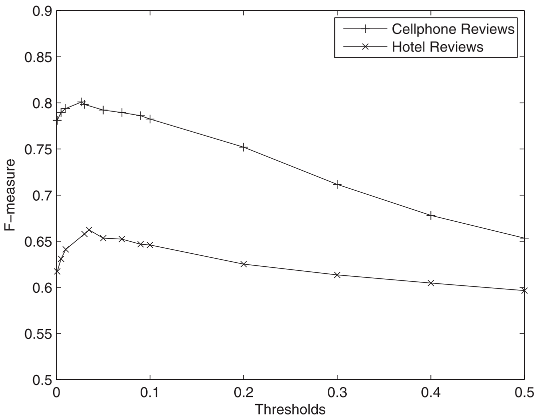


Fig. 9. IEDR F-measure performance versus intrinsic-domain relevance threshold ith. Extrinsic-domain relevance threshold was fixed at $\mathrm{eth} = 0.54$ for both cellphone and hotel review domains.

relatively lower (sixth and seventh places) for hotel reviews. This agrees with intuition because sports and tourism are much more related to hotels than to cellphones. Finally, *Military*, *Automobile*, as well as *IT* consistently ranked last for both hotel and cellphone feature extraction. This is due to the review-oriented nature of these topics, which include a lot of discussions about products or service, including their features, attributes, specifications, and performances. As a result, these poor performers do not differ too much from the domain-specific cellphone and hotel review corpora, rendering them the worst performers for IEDR feature extraction.

We, thus, conclude that IETR performed the best when the domain-independent corpus is most distinct from the domain review corpus. Consequentially, employing a completely different corpus like *Culture* leads to the best feature extraction results for both cellphone and hotel review domains.

4.5.2 Top-K Topics in Domain-Independent Corpus

Given the good performance of IEDR on individual domain-independent corpus (topic) like *Culture*, we would like to find out if combining more topics for the domain-independent corpus would further help IEDR performance.

We newly sampled 10 equal-sized corpora of 8,000 documents each from the original 10 domain-independent corpora. For example, the top-1 domain-independent corpus will contain 8,000 *Culture* documents for both cellphone and hotel domains. The top-2 domain-independent corpus for cellphone includes 4,000 *Culture* documents and 4,000 *Sports* documents. The top-3 domain-independent corpus for cellphone includes 2,667 *Culture*, 2,667 *Sports*, and 2,666 *Tourism* documents and so forth. Similarly, the top-2 domain-independent corpus for hotel includes 4,000 *Culture* and 4,000 *Education* documents. The top-3 domain-independent corpus for hotel includes 2,667 *Culture*, 2,667 *Education*, and 2,666 *Health* documents and so forth.

Fig. 8 shows the IEDR F-measure results for cellphone and hotel reviews versus top-K topics covered in each domain-independent corpus.

The IEDR performance on the cellphone domain shows relatively large variance from top 1 to top 10. In particular, the top-3 domain-independent corpus achieved the best F-measure of 80.64 percent, which is slightly better than the 80.27 percent F-measure achieved by the top 1 topic. However, adding more topics decreased performance, with the lowest F-measure of 77.24 percent at the top 6, 7, and 8 topics. Later, despite the slight improvement from top-8 to top-9 at 78.23 percent, it is still not comparable to the top-3 result. Intuitively, since the "bad" topics are similar to the domain review corpus, adding more of them actually displaces the "good" documents from the domain-independent corpus, thereby leading to poorer IEDR performance.

IEDR performed consistently for the hotel domain, with the best F-measure of 66.01 percent achieved at the top 1 *Culture* topic. The lowest performance (64.68 percent F-measure) was obtained on the domain-independent corpus containing top 4 topics, which is similar to the performance (64.79 percent F-measure) obtained for the top 10 topics. The smaller variation with respect to the topic numbers is attributed to the more generic nature of hotel reviews, i.e., the disparity between the 10 domain-independent topics and hotel reviews is smaller compared to cellphone reviews.

For both hotel and cellphone domains, the results show that a bag of widely diverse topics did not perform as well as a handpicked set of topics. In fact, using the top-1 domain-independent corpus, i.e., *Culture*, suffices to achieve good IEDR performance in practice.

We conclude with the IEDR best-practice of using a domain-independent corpus of around 10,000 documents, which should be drawn from one topic that is markedly distinct from the given review domain.

4.6 Domain Relevance Thresholds

In practice, it is very important to select appropriate domain relevance thresholds for the proposed IEDR method, which may vary across domains. We evaluate the IEDR performance against the two intrinsic and extrinsic relevance thresholds ith and eth, as shown in Fig. 9. The domain-independent corpus *Culture* was selected for both cellphone and hotel review domains.

In particular, given a selected extrinsic relevance threshold $\mathrm{eth} = 0.54$ (which can give relatively better performance), we first evaluate the F-measure versus the intrinsic

TABLE 6
Feature-Specific Opinion Mining Results on Cellphone Reviews

Methods	Precision	Recall	F-measure
FOM-IEDR	**65.60%**	61.71%	**63.60%**
FOM-IDR	57.97%	62.51%	60.15%
FOM-EDR	52.15%	**65.83**%	58.20%
FOM-LDA	51.54 %	60.42 %	55.62%
FOM-ARM	55.02 %	56.18 %	55.60%

TABLE 7
Feature-Specific Opinion Mining Results on Hotel Reviews

Methods	Precision	Recall	F-measure
FOM-IEDR	**50.37%**	54.30%	**52.26%**
FOM-IDR	45.23%	55.28%	49.76%
FOM-EDR	35.82%	**58.34%**	44.39%
FOM-LDA	46.84%	47.48%	47.16%
FOM-ARM	46.90%	47.09%	47.0%

relevance threshold ith on the cellphone reviews. The performance initially improved as ith increases from 0.001, achieving the best F-measure of 80.10 percent at $ith = 0.027$, and follows a declining trend thereafter all the way till $ith = 0.5$. This reasonable since a relatively larger intrinsic relevance threshold will prune many noisy features, however, growing it beyond a certain point will filter out some valid opinion features as well.

A similar trend can be observed on the hotel reviews. The best F-measure of 66.22 percent was achieved at the $ith = 0.035$ threshold, given the same extrinsic relevance threshold $eth = 0.54$.

Clearly, the choice of IEDR relevance thresholds can affect performance by more than 10 percent, as shown in the cellphone reviews. In practice, suitable thresholds can be determined by cross-validation on a small labeled set of data.

4.7 Feature-Based Opinion Mining Application

To appreciate the importance of well-extracted opinion features, we further evaluated IEDR in an opinion analysis setting. Specifically, we fed IEDR feature extraction results to an actual opinion mining system called iMiner [30], in which associated opinion words are recognized by using the IEDR identified features, after which the polarities of the opinion words are inferred by employing a likelihood ratio test [35] based semantic association method.

Feature-based opinion mining (FOM) results based on the opinion features identified via IEDR as well as IDR, EDR, LDA, and ARM are shown in Table 6, respectively. Note that the opinion mining results for MRC or DP cannot be evaluated because these two methods cannot isolate the feature extraction task from opinion mining.

As shown in Table 6, FOM-IEDR achieved the best F-measure of 63.60 percent which is 3.45, 5.4, 7.98, and 8.0 percent higher than that of FOM-IDR, FOM-EDR, FOM-LDA, and FOM-ARM, respectively. Note that FOM-EDR yields the highest recall, which benefits from EDR feature extraction, as EDR tends to discard overly generic terms, while keeping the domain-dependent terms. This allows EDR to find most of the domain features, but at the expense of including lots of noisy features.

The opinion mining results based on features extracted by IEDR, IDR, EDR, LDA, and ARM methods on hotel reviews are shown in Table 7. FOM-IEDR achieved a F-measure of 52.26 percent, which is 2.5, 7.87, 5.1, and 5.26 percent higher than that of FOM-IDR, FOM-EDR, FOM-LDA, and FOM-ARM, respectively. Similar to the opinion mining results on the cellphone review domain, we see that FOM-EDR yields the best recall due to the larger number of features that EDR finds.

The evaluation results on both hotel and cellphone domain not only show the effectiveness and robustness of IEDR in identifying opinion features across particular review domains, but also demonstrate that the improved feature extraction via IEDR can significantly boost the performance of feature-based opinion mining.

5 Discussion

The opinion feature extraction performance of IEDR (as well as all the competing methods) on the hotel reviews is not as good as that on the cellphone reviews. This is because hotel reviews

- are longer and more complicated, which makes feature extraction much more challenging, and
- contain large number of noisy domain-irrelevant user anecdotes or stories, which overlap with documents in the domain-independent corpus. In other words, the distinction between the intrinsic-domain relevance and the extrinsic-domain relevance of a feature is not that clear-cut for the hotel reviews.

As a result, IEDR (and all other competitors) is less successful in identifying features on the hotel domain compared to the cellphone domain.

Given a particular review domain, IEDR benefits greatly from a highly distinct domain-independent corpus for opinion feature extraction. It is, thus, very crucial to choose the right domain-independent corpus. According to our experiments, neither corpus size nor topic number in corpus has a large effect on IEDR feature extraction, but the distinct nature of the domain-independent corpus/topic from the given review domain makes a big difference.

The F-measure of opinion mining based on IEDR feature extraction is not great, achieving a mere 63.60 and 52.26 percent on cellphone and hotel domains, respectively. One explanation is that feature-based opinion mining depends not only on the feature identification results, but also relies on the methods used to classify opinion polarities. Moreover, the main focus of this paper is on feature extraction, not on opinion classification. We do not claim that feature extraction alone will lead to fantastic opinion orientation classification performance, but merely improves it.

The evaluations using real-world Chinese reviews have demonstrated the effectiveness of the proposed IEDR. Note that our proposed IEDR method is largely language-independent except for the candidate feature extraction part, which can be easily addressed via defining similar simple synthetic rules or common NLP noun-extraction tools in other different languages.

Though the proposed IEDR approach has resulted in improved performance compared to several existing mainstream methods, several weaknesses still exist:

- As a corpus statistics approach, IEDR is less successful in dealing with the extraction of infrequent features. For instance, "FM" (which is the shortform for "radio") in cellphone reviews and "French window" in hotel reviews could not be extracted via IEDR, due to their relatively low occurrence frequencies. Feature clustering can be used to address this problem in future.
- IEDR does not currently extract non-noun opinion features due to the limitation of only considering nouns (noun phrases) for candidate feature extraction in the dependence parsing phase. According to our experimental results, the three syntactic rules yield a recall of 75.22 percent for feature extraction on the hotel domain, which is not as good as that in cellphone domain (where an 83.82 percent recall is achieved). This is because hotel reviews contain much more verbal features, compared to cellphone reviews, which cannot be recognized via our noun-based syntactic rules.
- IEDR is at the mercy of other errors in the POS-tagging and syntactic parsing stages, due to the informal or grammatically incorrect sentences contained in real-life reviews.

6 Conclusions

In this paper, we proposed a novel intercorpus statistics approach to opinion feature extraction based on the IEDR feature-filtering criterion, which utilizes the disparities in distributional characteristics of features across two corpora, one domain-specific and one domain-independent. IEDR identifies candidate features that are specific to the given review domain and yet not overly generic (domain-independent). Experimental results demonstrate that the proposed IEDR not only leads to noticeable improvement over either IDR or EDR, but also outperforms four mainstream methods, namely, LDA, ARM, MRC, and DP, in terms of feature extraction performance as well as feature-based opinion mining results. In addition, since a good-quality domain-independent corpus is quite important for the proposed approach, we evaluated the influence of corpus size and topic selection on feature extraction performance. We found that using a domain-independent corpus of a similar size as but topically different from the given review domain will yield good opinion feature extraction results.

For future work, we will employ fine-grained topic modeling approach to jointly identify opinion features, including non-noun features, infrequent features, as well as implicit features. We plan to further test the IEDR opinion feature extraction in several other opinion mining systems. In addition, neutral opinions will be considered; currently only positive and negative opinions are considered. We will also investigate new opinion mining algorithms that make good use of the IEDR extracted opinion features to summarize online reviews of products or services. A new two-stage feature-based opinion mining approach will be developed, where opinion features are first identified, followed by the discovery of the associated opinion polarities. Finally, we will also evaluate the IEDR approach on reviews in other languages. We have had some preliminary success in applying IEDR to extract English opinion features from hotel reviews.

Acknowledgments

The authors would like to thank the Center for Information Retrieval at Harbin Institute of Technology for providing the LTP software. They would also like to thank the anonymous reviewers for their valuable comments. This research was supported in part by Singapore Ministry of Education's Academic Research Fund Tier 2 grant ARC 9/12 (MOE2011-T2-2-056).

References

[1] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies,* vol. 5, no. 1, pp. 1-167, May 2012.

[2] W. Jin and H.H. Ho, "A Novel Lexicalized HMM-Based Learning Framework for Web Opinion Mining," *Proc. 26th Ann. Int'l Conf. Machine Learning,* pp. 465-472, 2009.

[3] N. Jakob and I. Gurevych, "Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields," *Proc. Conf. Empirical Methods in Natural Language Processing,* pp. 1035-1045, 2010.

[4] S.-M. Kim and E. Hovy, "Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text," *Proc. ACL/COLING Workshop Sentiment and Subjectivity in Text,* 2006.

[5] G. Qiu, C. Wang, J. Bu, K. Liu, and C. Chen, "Incorporate the Syntactic Knowledge in Opinion Mining in User-Generated Content," *Proc. WWW 2008 Workshop NLP Challenges in the Information Explosion Era,* 2008.

[6] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion Word Expansion and Target Extraction through Double Propagation," *Computational Linguistics,* vol. 37, pp. 9-27, 2011.

[7] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Research,* vol. 3, pp. 993-1022, Mar. 2003.

[8] I. Titov and R. McDonald, "Modeling Online Reviews with Multi-Grain Topic Models," *Proc. 17th Int'l Conf. World Wide Web,* pp. 111-120, 2008.

[9] Y. Jo and A.H. Oh, "Aspect and Sentiment Unification Model for Online Review Analysis," *Proc. Fourth ACM Int'l Conf. Web Search and Data Mining,* pp. 815-824, 2011.

[10] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," *Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* pp. 168-177, 2004.

[11] A. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," *Proc. Human Language Technology Conf. and Conf. Empirical Methods in Natural Language Processing,* pp. 339-346, 2005.

[12] V. Hatzivassiloglou and J.M. Wiebe, "Effects of Adjective Orientation and Gradability on Sentence Subjectivity," *Proc. 18th Conf. Computational Linguistics,* pp. 299-305, 2000.

[13] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment Classification Using Machine Learning Techniques," *Proc. Conf. Empirical Methods in Natural Language Processing,* pp. 79-86, 2002.

[14] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," *Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics,* 2004.

[15] R. Mcdonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar, "Structured Models for Fine-to-Coarse Sentiment Analysis," *Proc. 45th Ann. Meeting of the Assoc. of Computational Linguistics,* pp. 432-439, 2007.

[16] L. Qu, G. Ifrim, and G. Weikum, "The Bag-of-Opinions Method for Review Rating Prediction from Sparse Text Patterns," *Proc. 23rd Int'l Conf. Computational Linguistics,* pp. 913-921, 2010.

[17] D. Bollegala, D. Weir, and J. Carroll, "Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus," *IEEE Trans. Knowledge and Data Eng.,* vol. 25, no. 8, pp. 1719-1731, Aug. 2013.

[18] P.D. Turney, "Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proc. 40th Ann. Meeting on Assoc. for Computational Linguistics,* pp. 417-424, 2002.

[19] C. Zhang, D. Zeng, J. Li, F.-Y. Wang, and W. Zuo, "Sentiment Analysis of Chinese Documents: From Sentence to Document Level," *J. Am. Soc. Information Science and Technology,* vol. 60, no. 12, pp. 2474-2487, Dec. 2009.

[20] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis," *Proc. 49th Ann. Meeting of the Assoc. for Computational Linguistics: Human Language Technologies,* pp. 142-150, 2011.

[21] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," *Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing,* pp. 347-354, 2005.

[22] A. Yessenalina and C. Cardie, "Compositional Matrix-Space Models for Sentiment Analysis," *Proc. Conf. Empirical Methods in Natural Language Processing,* pp. 172-182, 2011.

[23] E. Cambria, D. Olsher, and K. Kwok, "Sentic Activation: A Two-Level Affective Common Sense Reasoning Framework," *Proc. 26th AAAI Conf. Artificial Intelligence,* pp. 186-192, 2012.

[24] F. Li, C. Han, M. Huang, X. Zhu, Y.-J. Xia, S. Zhang, and H. Yu, "Structure-Aware Review Mining and Summarization," *Proc. 23rd Int'l Conf. Computational Linguistics,* pp. 653-661, 2010.

[25] S.J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Trans. Knowledge and Data Eng.,* vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[26] J. Yu, Z.-J. Zha, M. Wang, and T.-S. Chua, "Aspect Ranking: Identifying Important Product Aspects from Online Consumer Reviews," *Proc. 49th Ann. Meeting of the Assoc. for Computational Linguistics: Human Language Technologies,* pp. 1496-1505, 2011.

[27] W.X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly Modeling Aspects and Opinions with a Maxent-Lda Hybrid," *Proc. Conf. Empirical Methods in Natural Language Processing,* pp. 56-65, 2010.

[28] L. Tesniere, *Elements de la syntaxe structurale.* Librairie C. Klincksieck, 1959.

[29] F. Fukumoto and Y. Suzuki, "Event Tracking Based on Domain Dependency," *Proc. 23rd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval,* pp. 57-64, 2000.

[30] Z. Hai, K. Chang, Q. Song, and J.-J. Kim, "A Statistical Nlp Approach for Feature and Sentiment Identification from Chinese Reviews," *Proc. CIPS-SIGHAN Joint Conf. Chinese Language Processing,* pp. 105-112, 2010.

[31] W. Che, Z. Li, and T. Liu, "LTP: A Chinese Language Technology Platform," *Proc. 23rd Int'l Conf. Computational Linguistics,* pp. 13-16, 2010.

[32] A.J. Viera and J.M. Garrett, "Understanding Interobserver Agreement: The Kappa Statistic." *Family Medicine,* vol. 37, no. 5, pp. 360-363, May 2005.

[33] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," *Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* pp. 342-351, 2004.

[34] Q. Su, X. Xu, H. Guo, Z. Guo, X. Wu, X. Zhang, B. Swen, and Z. Su, "Hidden Sentiment Association in Chinese Web Opinion Mining," *Proc. 17th Int'l Conf. World Wide Web,* pp. 959-968, 2008.

[35] T. Dunning, "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics,* vol. 19, no. 1, pp. 61-74, Mar. 1993.

Zhen Hai is currently working toward the PhD degree in the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include opinion mining (or sentiment analysis), information retrieval, and topic models. He is currently working on probabilistic graphical models and their application to problems like helpful review selection and aspect-based review summarization.

Kuiyu Chang received the BS degree from National Taiwan University, the MS degree from the University of Hawaii at Manoa, and the PhD degree from the University of Texas at Austin, all in electrical/computer engineering. He is an assistant professor of computer engineering at Nanyang Technological University. Prior to that, he served as senior risk management analyst for ClearCommerce (now eFunds). From 2000 to 2002, he was a member of technical staff at Interwoven (now HP). He has served as program cochair for PAISI (2006-2008), and publications chair for PAKDD 2006. He has published more than 60 papers in the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *IEEE Transactions on Knowledge and Data Engineering*, SIGIR, IJCAI, ICDE, ICDM, and SDM. He is also a corecipient of three international best paper awards (PAKDD 2012, ISI 2005, and Motorola 1996) and one merit paper award (IAENG 2012). He consults regularly for IT companies in Singapore, Malaysia, and China.

Jung-Jae Kim received the BSc, MS, and PhD degrees from Korea Advanced Institute of Science and Technology, Korea, in 1998, 2000, and 2006, respectively. He is currently an assistant professor in the School of Computer Engineering at Nanyang Technological University in Singapore. He has worked as a postdoctoral researcher for the Text Mining Group of the European Bioinformatics Institute from 2006 to 2009.

Christopher C. Yang is an associate professor in the College of Information Science and Technology at Drexel University. His recent research interests include social intelligence and technology, healthcare informatics, security informatics, web search and mining, and knowledge management. He has published more than 200 referred journal and conference papers in the *ACM Transactions on Intelligent Systems and Technology*, the *Journal of the American Society for Information Science and Technology* (JASIST), the *IEEE Transactions on Systems, Man, and Cybernetics*, the *IEEE Transactions on Image Processing*, the *IEEE Transactions on Robotics and Automation*, *Computer*, *IEEE Intelligent Systems*, Decision Support Systems (DSS), Information Processing and Management (IPM), Graphical Models and Image Processing, Optical Engineering, Pattern Recognition, the *International Journal of Electronic Commerce,* and more. He is currently serving as the associate editor-in-chief of *Security Informatics* published by Springer and the coeditor of *Electronic Commerce Research and Applications* published by Elsevier. He has edited several special issues on social media, healthcare informatics, security informatics, web mining, multilingual information systems, knowledge management, and electronic commerce in IEEE Transactions, ACM Transactions, *IEEE Intelligent Systems*, the *Journal of the American Society for Information Science and Technology*, DSS, IPM, and others. He has chaired many international conferences and workshops such as the IEEE International Conference on Healthcare Informatics, the International Conference on Social Intelligence and Technology, the ACM SIGHIT International Health Informatics Symposium, the IEEE International Conference on Intelligence and Security Informatics, the ACM International Conference on Information and Knowledge Management, the International Conference on Electronic Commerce, and more. In his recent work on healthcare informatics and security informatics, he closely collaborated with the Children's Hospital of Philadelphia, UPenn Medical School, Johnson & Johnson, Philadelphia Police Department, Temple University, and the Canadian Institute of Urban Research Studies at Simon Fraser University. His work has been supported by the US National Science Foundation, Pennsylvania Department of Health, Children's Hospital of Philadelphia, Hong Kong Research Grant Council, Hong Kong Innovation and Technology Fund, and Hong Kong SAR Government.