OpinionIt: A Text Mining System for Cross-lingual Opinion Analysis

Honglei Guo, Huijia Zhu, Zhili Guo, Xiaoxun Zhang and Zhong Su
IBM Research –China
Beijing, China, 100193
{guohl, zhuhuij,guozhili, zhangxx, suzhong}@cn.ibm.com

ABSTRACT

Opinion mining focuses on extracting customers' opinions from the reviews and predicting their sentiment orientation. Reviewers usually praise a product in some aspects and bemoan it in other aspects. With the business globalization, it is very important for enterprises to extract the opinions toward different aspects and find out cross-lingual/cross-culture difference in opinions. Cross-lingual opinion mining is a very challenging task as amounts of opinions are written in different languages, and not well structured. Since people usually use different words to describe the same aspect in the reviews, product-feature (PF) categorization becomes very critical in cross-lingual opinion mining. Manual cross-lingual PF categorization is time consuming, and practically infeasible for the massive amount of data written in different languages. In order to effectively find out cross-lingual difference in opinions, we present an aspect-oriented opinion mining method with Cross-lingual Latent Semantic Association (CLaSA). We first construct CLaSA model to learn the cross-lingual latent semantic association among all the PFs from multi-dimension semantic clues in the review corpus. Then we employ CLaSA model to categorize all the multilingual PFs into semantic aspects, and summarize cross-lingual difference in opinions towards different aspects. Experimental results show that our method achieves better performance compared with the existing approaches. With CLaSA model, our text mining system OpinionIt can effectively discover cross-lingual difference in opinions.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*; I.2.7 [**Artificial Intelligence**]: Natural language processing—*text analysis*

General Terms

Algorithms, Experimentation

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

Keywords

cross-lingual opinion mining, product feature categorization, latent semantic association

1. INTRODUCTION

With the dramatic growth of web's popularity, amounts of people can post their reviews or comments for various products or services. These textual information exchanged on the Web is written in various languages. With the business globalization, it is very important for the enterprises to extract customers' opinions and find out the cross-lingual distribution difference in opinions. Cross-lingual opinion mining already becomes a critical task for business intelligence.

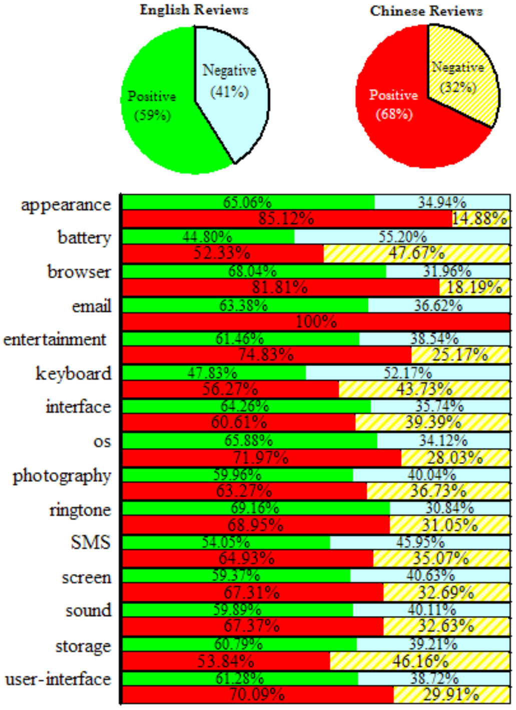


Figure 1: Cross-lingual Opinion Mining System OpinionIt. Pie charts compare the overall sentiment distribution. Bar charts compare the sentiment distribution on product aspects in Chinese and English reviews.

In the real application, we build a cross-lingual opinion mining system *OpinionIt*. Our goal is to facilitate enterprise to better understand customer concerns. *OpinionIt* can provide the overall sentiment distribution across languages, and find out fine-grained cross-lingual difference. As shown in Figure 1, the two pies compare the overall sentiment distribution in Chinese and English reviews. The bar charts compare the ratio of positive and negative opinions towards different aspects. From such comparison, one can easily find out the cross-lingual differences in opinions. Since reviewers usually praise a product in some aspects and bemoan it in other aspects, such aspect-oriented cross-lingual opinion mining provides more business insights for enterprises.

Cross-lingual opinion mining is a very challenging task since amounts of opinions are written in different languages, and not well structured. In *OpinionIt*, in order to effectively find out cross-lingual difference in opinions, we present an aspect-oriented opinion mining method with Cross-lingual Latent Semantic Association (CLaSA). The proposed method first constructs CLaSA model to categorize all the multilingual PFs into unified semantic categories, and then summarizes cross-lingual difference in opinions towards different aspects. From the viewpoint of text mining, we essentially use CLaSA model as a latent semantic association framework to find out cross-lingual difference in opinions.

Since people may use various linguistic representations to refer to the same PF in the multilingual reviews, PF categorization is always the key challenge for cross-lingual opinion mining in the real applications. Its quality directly impacts cross-lingual opinion mining. Although some product specifications are provided by the merchants, customers and merchants may use different words to refer to the same aspect. For example, English PFs "*photo*", "*picture*", "*image*", and Chinese PFs "*ZhaoXiang (i.e. photograph)*" and "*PaiZhao (i.e. photograph)*" all refer to the same aspect in digital camera reviews. It is tedious and time consuming to manually categorize all the PFs from the reviews in different languages. Hence, it raises the need for automatically categorizing all the multilingual PFs into a unified semantic categorization framework. Although some methods [14, 23, 24, 3, 27, 8] have been presented to build individual model for clustering PFs in one language. In order to create a unified cross-lingual PF categorization, all the PF clustering results in different languages need to be manually merged or aligned. Such manual alignment is practically infeasible for the massive amount of multilingual data. Hence, we present a CLaSA model for cross-lingual PF categorization. It categorizes all the multilingual PFs into unified semantic groups according to the multi-dimension cross-lingual latent semantic clues in the review corpus. These clues include: 1) the current PF term pf; 2)pf's machine translation pf^T; 3) component words in pf and pf^T; 4) word-level latent semantic topics for component words in pf and pf^T; 5) monolingual PF-level latent semantic of pf.

One important contribution of our method is that we create cross-lingual virtual context document for each PF with a bag of PF, its machine translation and mono-lingual latent semantic clues. These language-mixed virtual documents effectively characterize the latent association among PFs from different languages. The intuition behind our method is that words and terms in one concept set will have similar semantic features or latent semantic association, and share similar context in the corpus. They can be considered as behaving in the same way in the cross-lingual PF categorization. The proposed CLaSA model categorizes the multilingual PFs on a semantic level rather than by lexical comparison. It can better approximate the true underlying semantic category distribution in the domain without any labeled samples. Experimental results show that our method achieves better performance compared with the existing approaches. With CLaSA model, we can find out cross-lingual difference in opinions by effectively joining reviews between languages.

The remainder of the paper is organized as follows. Section 2 introduces some related work. Section 3 proposes cross-lingual latent semantic association method. Section 4 presents cross-lingual opinion mining with CLaSA model. The experimental results are discussed in Section 5. The conclusion is given in Section 6.

2. RELATED WORK

The task of opinion mining is to extract customers' opinions from the reviews and predict their sentiment orientation. Some papers [14, 10, 11, 12, 18, 20, 29, 22] focus on mining opinions from monolingual customer reviews. Many of these methods are dependent on the training data set, so are not generally applicable to cross-lingual opinion mining. Lu et al. [15] integrate monolingual expert reviews and customer reviews through semi-supervised topic modeling. Mei et al. [17] propose a mixture model to discover facets and opinions in the monolingual reviews at the same time. Several works also study cross-lingual analysis of sentiment and concerns [1, 19, 7, 5]. They analyze sentiment distribution in multilingual news and blogs collected with a topic keyword, and mine comparative difference of concerns. In this paper, we present a cross-lingual opinion mining method with CLaSA model. With CLaSA model, we can find out cross-lingual difference in opinions from customer reviews.

PF categorization is a key task for cross-lingual opinion mining. Some methods have been presented for monolingual PF categorization. Liu et al. [14] group PFs by the synonym set in WordNet and the semi-automated tagging of reviews. Titov and McDonald [23] employ rating information to identify coherent aspects in the reviews. They further propose a multi-grain topic model to cluster aspects [24], which considers the distributions of both global topics for a document and local topics for the local context of the word. Brananvan et al [3] cluster all the keyphrases in the monolingual reviews based on the lexical comparison and the distribution similarity. The lexical comparison is based on the cosine similarity among the phrase words while the distributional similarity is quantified in terms of the co-occurrence of keyphrases across review texts. Wong et al. [27] extract and normalize product attributes from the structured product descriptions provided by merchants in multiple web sites. However, merchants and customers may use different words to refer to the same feature. Guo et al. [8] present a monolingual PF categorization method, which constructs the categorization model according to the internal latent semantic and external context snippets of the PFs in the reviews. Different from these methods, in this paper, we present CLaSA model for cross-lingual PF categorization, which categorizes all PFs according to the multi-dimension cross-lingual latent semantic clues in the reviews.

Topic model has been successfully applied to mine topic patterns from text collections, and enhance document representations in text classification and information retrieval

[2, 9, 28, 13, 26]. Our work adds another novel use of such models for cross-lingual PF categorization.

3. CROSS-LINGUAL LATENT SEMANTIC ASSOCIATION METHOD

Since people may use various PF terms to refer the same aspect in the reviews, PF categorization is the key task in cross-lingual opinion mining. The challenge in PF categorization is how to capture latent semantic association among the PFs. In this section, we present CLaSA model to find out the cross-lingual association structures among the PFs.

3.1 Cross-lingual Latent Semantic Association

Each product aspect often has various terms. For example, English PFs "*photo*", "*picture*", "*image*", and Chinese PFs "*ZhaoXiang (i.e. photograph)*" and "*PaiZhao (i.e. photograph)*" all refer to the same aspect. Hence, it is a big challenge for us to effectively capture latent semantic association among multilingual PF terms from the reviews.

Cross-lingual PF categorization focuses on categorizing the PFs into a unified semantic category framework. Let X be a feature space to represent the multilingual PF instances, and let Y be the set of semantic category labels. Let $p_s(x, y)$ and $p_t(x, y)$ be the predicted underlying category distribution and the true underlying category distribution, respectively. In order to minimize the human efforts, we expect to $p_s(x, y)$ better approximate $p_t(x, y)$ without using any labeled data.

Cross-lingual PF categorization based on lexical comparison is usually not comprehensive enough to capture the underlying semantic distribution of various multilingual PFs. Many PF terms in the same aspect are not similar on the lexical level. However, these terms often appear in the similar semantic context. For instance, PF terms in the aspect "*appearance*" often occur around the indicates "*beautiful*", "*fashion*", "*popular*" etc. Such latent semantic association among words provides useful hints for capturing the underlying semantic distribution in the domain.

Hence, we present CLaSA model Φ to capture cross-lingual latent semantic association among the multilingual PFs. Φ is learned from the unlabeled review corpus. In the learning, each PF term is characterized by multi-dimension cross-lingual latent semantic clues. These clues include: 1) the current PF term pf; 2) pf's machine translation entry pf^T; 3) component words in pf and pf^T; 4) word-level latent semantic topics for component words in pf and pf^T; 5) monolingual PF-level latent semantic of pf. Semantic association feature in Φ is a hidden random variable that is inferred from data. Even though PF terms do not have the lexical similarity, but are in similar context, they still might have relatively high probability in the same semantic concept set. Obviously, Φ can better capture the cross-lingual latent semantic association among the PFs. With Φ, we may better approximate the real semantic category distribution $p_t(y|x; \Phi)$ without using any labeled data.

3.2 Learning CLaSA Model from Cross-lingual Virtual Context Documents

3.2.1 Cross-lingual Virtual Context Document

In order to learn latent relationships among multilingual terms, each PF term is characterized by a cross-lingual virtual context document. Given a PF term pf, its cross-lingual virtual context document cvd_{pf} is composed of the following multi-dimension latent semantic clues.

1. Current PF term pf and its machine translation result pf^T;
2. Component word sets in pf and pf^T, denoted as W_{pf} and W_{pf^T};
3. Word-level latent semantic topics of component words in pf and pf^T, denoted as S_{pf} and S_{pf^T};
4. Monolingual PF-level latent semantic of pf, denoted as MFS_{pf};

Hence,

$cvd_{pf} = \{pf, pf^T, W_{pf}, W_{pf^T}, S_{pf}, S_{pf^T}, MFS_{pf}\}$;

For example, given pf=*"screen resolution"*, Table 1 shows its cross-lingual virtual context document *cvd*("*screen resolution*") extracted from English and Chinese review corpus.

CLaSA Clues	*cvd*("*screen resolution*") (English and Chinese Reviews)
pf	"*screen resolution*" (English)
pf^T	"*Ping-Mu Fen-Bian-Lv*" (Chinese)
W_{pf}	"*screen*", "*resolution*"
W_{pf^T}	"*Ping-Mu*", "*Fen-Bian-Lv*"
S_{pf}	S("*screen*"), S("*resolution*")
S_{pf^T}	S("*Ping-Mu*"), S("*Fen-Bian-Lv*")
$MFS(pf)$	MFS("*screen resolution*")

Table 1: Cross-lingual virtual context document

In the cross-lingual virtual context document construction, we generate monolingual word-level and PF-level latent semantic topics using the algorithms presented in [8]. Component words are grouped into a set of latent topics according to their context in the monolingual corpus. Monolingual PF-level latent semantic topics are generated according to their latent semantic structures and context snippets in the corresponding corpus. Section 4.1 will illustrate how to construct the cross-lingual virtual context document in CLaSA-based PF categorization. Full-document machine translation is usually employed to capture the semantic association among PFs written in different languages. However, with the limitation of the existing machine translation techniques, data noise arose from full-document machine translation often make a bad impact on the cross-lingual semantic association. In order to reduce the noise arose from machine translation, the cross-lingual virtual context document employs PF term machine translation instead of full-review translation.

cvd_{pf} actually describes multi-dimension cross-lingual latent semantic features of pf in the reviews. We construct the feature vector of pf with all the observed features in cvd_{pf}. Given the feature vector of cvd_{pf}, $Vector$ (cvd_{pf}) = $\{x_1, ..., x_j, ..., x_m\}$, x_j denotes the jth context feature related to pf, m is the total number of features in cvd_{pf}. The weight of each context feature x_j in cvd_{pf} is calculated by Mutual Information [6] between pf and x_j.

$$Weight(x_j, pf) = \log_2 \frac{P(x_j, pf)}{P(x_j)P(pf)} \quad (1)$$

where, $P(x_j, pf)$ is the joint probability of pf and x_j cooccurred in the corpus, $P(x_j)$ is the probability of x_j occurred in the corpus. $P(pf)$ is the probability of pf occurred in the corpus. The weight is normalized to non-negative in the model training.

3.2.2 *Learning CLaSA Model*

CLaSA model actually can be considered as a general probabilistic topic model. It can be learned from the unlabeled review corpus using the hidden topic models such as Latent Dirichlet Allocation (LDA) [2] and probabilistic Latent Semantic Indexing (pLSI) [9].

Topic models are statistical models of text that posit a hidden space of topics in which the corpus is embedded [2]. LDA is a probabilistic model that can be used to model and discover underlying topic structures of documents. LDA assumes that there are K "topics", multinomial distributions over words, which describes a collection. Each document exhibits multiple topics, and each word in the document is associated with one of the topics. LDA imposes a Dirichlet distribution on the topic mixture weights corresponding to the documents in the corpus. The topics derived by LDA seem to possess semantic coherence. Those words with similar semantics are likely to occur in the same topic.

Algorithm 1: CLaSA Model Training

```
Inputs:
• $R_{l_1}$: customer review corpus written in language $l_1$;
• $R_{l_2}$: customer review corpus written in language $l_2$;
• $PFSet$: all the PFs used in $R_{l_1}$ and $R_{l_2}$;
• Monolingual Word-level latent topic models: $\theta^{l_1}_{wd}$, $\theta^{l_2}_{wd}$;
• Monolingual PF-level latent topic models: $\theta^{l_1}_{mp}$, $\theta^{l_2}_{mp}$;
Outputs:
• Cross-lingual PF categorization model: $\Phi$;
Initialization:
• Cross-lingual context document set: $CVDSet = \emptyset$;
Steps:
begin
    foreach $pf_i \in PFSet$ do
        $pf_i^T = MT(pf_i)$;
        $l_s \leftarrow$Language($pf_i$);
        $l_t \leftarrow$Language($pf_i^T$);
        $W_{pf_i} = GetComponentWords(pf_i)$;
        $W_{pf_i^T} = GetComponentWords(pf_i^T)$;
        foreach $w_j \in W_{pf_i}$ do
            $S_{w_j} = TP(w_j, \theta^{l_s}_{wd})$;
            $AddTo(S_{w_j}, S_{pf_i})$;
        foreach $w_k \in W_{pf_i^T}$ do
            $S_{w_k} = TP(w_k, \theta^{l_t}_{wd})$;
            $AddTo(S_{w_k}, S_{pf_i^T})$;
        $MFS_{pf_i} = TP(pf_i, \theta^{l_s}_{mp})$;
        $cvd_{pf_i} =$
        $\{pf_i, pf_i^T, W_{pf_i}, W_{pf_i^T}, S_{pf_i}, S_{pf_i^T}, MFS_{pf_i}\}$;
        $AddTo(cvd_{pf_i}, CVDSet)$;
    • Generate CLaSA model $\Phi$ with Dirichlet
    distribution on $CVDSet$.
end
```

In the following, we illustrate how to learn LDA-style CLaSA model Φ on the cross-lingual virtual semantic context documents. Given the unlabeled review corpus R_{l_1} and R_{l_2} written in the languages l_1 and l_2. The cross-lingual virtual context document of each PF is first constructed (see Section 3.2.1). Given a PF term pf written in Language l. In cvd_{pf} construction, latent topics of component words are generated using the monolingual word-level topic model θ^l_{wd}. Monolingual latent semantic of each PF is generated using the monolingual PF-level topic model θ^l_{mp} The weight of each feature in cvd_{pf_i} is computed using Mutual Information (see Equation 1 in Section 3.2.1). Then, CLaSA model Φ with Dirichlet distribution is generated on the cross-lingual virtual context document set $CVDSet$ using the algorithm presented by Blei et al [2]. In our experiments, α=0.1, and the number of iterations is 1000. Algorithm 1 describes the whole process in detail, where, Function $MT(pf_i)$ denotes that the machine translation result of pf_i. Function $TP(data, \theta)$ generates the latent topic for $data$ using the latent topic model θ; θ^l_{wd} denotes the monolingual word-level latent topic model for the given language l; θ^l_{mp} denotes the monolingual PF-level latent topic model for the given language l;

CLaSA model learns the posterior distribution to decompose multilingual PFs and their virtual context documents into topics. It extends the traditional bag-of-words topic models to context-dependence cross-lingual concept association model.

4. CROSS-LINGUAL OPINION MINING WITH CLASA MODEL

In *OpinionIt*, we present an aspect-based cross-lingual opinion mining framework with CLaSA model. The proposed method first constructs CLaSA model to categorize the multilingual PFs into semantic aspects, then summarizes cross-lingual difference in opinions toward different aspects.

4.1 CLaSA-based Cross-lingual PF Categorization

In the multilingual reviews, people often use various terms to refer to the same feature. It is very important to categorize PFs into semantic aspects in cross-lingual opinion mining. In order to capture latent semantic association among various opinions, we learn CLaSA-based cross-lingual PF categorization model Φ using Algorithm 1 in Section 3.2.2. Given the unlabeled review corpus R_{l_1} and R_{l_2}, the cross-lingual virtual context documents of all the PFs are first constructed from the review corpus. Then, PF categorization model Φ is trained on these virtual documents. It learns the posterior distribution to decompose multilingual PFs

In CLaSA-based PF categorization model learning, the cross-lingual virtual context document for each PF term is constructed from the review corpus as follows.

Component words in the PF term are important indicators for semantic categorization. Hence, in the cross-lingual virtual context document construction, given a PF term pf written in language l, we employ the word-level topic model θ^l_{wd} to generate the latent topics of component words in pf.

$$S(pf)= \sum_{w_j \in pf} TP(w_j, \theta^l_{wd});$$

where, $TP(w_j, \theta^l_{wd})$ generates word w_j's topic using θ^l_{wd}. Model θ^l_{wd} is learned from the monolingual review corpus R_l using the algorithm presented in [8]. It can effectively group the component words into a set of concepts according to their context in the corpus. In the model θ^l_{wd} construction, each word is characterized by all the context units around it in the corpus. For example, given $word_k$=*"screen"*,

sen_i=*“My new thinkpad is very good because its LCD screen is very large and nice”*. In the context window {-3, 3} (i.e. previous 3 words and next 3 words around $word_k$ in sen_i), the context feature set around *“screen”* in sen_i includes: 1) the current word: A_c(*screen*); 2) left/right opinion sets (i.e., two left/right adjacent adjective units): O_L(*good*), O_L(*new*), O_R(*large*), O_R(*nice*); 3) the nearest left/right adjacent units: A_L(*LCD*), A_R(*is*); 4) the other two left/right adjacent context unit sets: C_L(*its*), C_L (*because*), C_R(*very*), C_R(*large*).

Topic 4	Topic 8	Topic 12	Topic 16
design	sd	screen	video
looking	slot	display	multimedia
style	micro-sd	lcd	audio
appearance	expansion	touch screen	media
shape	storage	recorder	tv

Table 2: latent semantic topics of some component words in cell phone domain

Table 2 shows some latent topics of component words generated by the model θ^{l}_{wd}. As shown, words in the same topic actually are grouped into broad concept sets. For example, topic 4, 8, 12 and 16 are related to the concepts *appearance*, *storage*, *screen* and *entertainment*, respectively. Hence, in the cross-lingual virtual context document construction, we can effectively generate the latent semantic topics for the component words using the model θ^{l}_{wd}. For example, S(“*standard sd card*”) = {$Topic$32, $Topic$8, $Topic$2}.

In the cross-lingual virtual context document construction, given a PF term pf written in language l, we employ the monolingual PF-level topic model θ^{l}_{mp} to generate the monolingual latent semantic for pf, that is,

$$MFS_{pf}= TP(pf,\theta^{l}_{mp});$$

where, $TP(pf,\theta^{l}_{mp})$ generates pf's monolingual latent topic using θ^{l}_{mp}. Model θ^{l}_{mp} is learned from the review corpus R_l using the algorithm presented in [8]. Model θ^{l}_{mp} assigns a monolingual latent semantic topic to pf according to its internal latent topic structure and external context snippets in the corpus R_l. The internal latent semantic structures of pf (denoted by $ts(pf)$) is defined as the topic label sequence of all the words in pf, that is, “t_1 t_2 .... t_j .... t_n”, where, t_j denotes the jth word's latent topic assigned by the word-level latent topic model θ^{l}_{wd}. For example, ts(*“standard sd slot”*)= “*Topic32 Topic8 Topic8*”. The external context set of pf in a sentence sen_k is composed of a bag of all the non-stop words in sen_k. For example, given pf=*“LCD screen”*, sen_k=*“Its LCD screen is very nice”*, its context units consist of { “*LCD*”, “*screen*”, “*nice*”}.

In LDA-style CLaSA model Φ, the topic mixture is drawn from a conjugate Dirichlet prior that remains the same for all the cross-lingual virtual context documents. Hence, given a PF pf_i in the reviews, we may perform posterior inference to determine the conditional distribution of the hidden topic feature variables associated with pf_i. The semantic category of pf_i (denoted by $Category(pf_i)$) is generated using Algorithm 2. Here, Mult(Φ(cvd_i)) refers to sample from the posterior distribution over topics given a cross-lingual virtual document cvd_i.

With all the multi-dimension latent semantic clues extracted from the reviews, the CLaSA-based PF categorization model can better capture the cross-lingual semantic category distribution without any labeled data. Even though PF terms do not have the lexical similarity, but are in similar latent semantic context, they still might have relatively high probability in the same semantic category.

Algorithm 2: Generate semantic category for PF term pf_i Using K-topic CLaSA Model

```
Inputs:
• Φ: CLaSA-based Cross-lingual PF categorization
  model with multinomial distribution;
• $Dirichlet(\alpha)$: Dirichlet distribution with parameter $\alpha$;
• $pf_i$: PF term used in the multilingual reviews;
Outputs:
• $Category(pf_i)$: semantic category of $pf_i$;
Steps:
begin
    • Extract the cross-lingual virtual context document
      $cvd_i$ for $pf_i$ from the multilingual reviews;
    • Sample a topic distribution Φ($cvd_i$) from
      $Dirichlet(\alpha)$;
    • foreach $x_j \in cvd_i$ do
        sample topic $z_j \in$ { 1,...,$K$} from Mult(Φ($cvd_i$));
        $AddTo(z_j, Topics(cvd_i))$;
    • $Category(pf_i) \longleftarrow$ topic $z_k$ with the highest
      probability distribution in $Topics(cvd_i)$;
end
```

4.2 Aspect-oriented Opinion Integration with CLaSA Model

In *OpinionIt*, we focus on mining cross-lingual opinions towards different aspects. Given a collection of multilingual customer reviews, and CLaSA-based cross-lingual PF categorization model Φ. Our aspect-oriented opinion integration approach consists of the following stages.

1. Categorize all the multilingual PFs into semantic aspects {c_1;....; c_n} using CLaSA model Φ;

2. Extract the opinions towards different aspects.

3. Summarize cross-lingual difference in opinions towards different aspects.

In *OpinionIt*, all the multilingual PFs are first categorized into sematic aspects using CLaSA model (see Algorithm 2). Then, we extract the PF-Opinion pairs from the multilingual reviews and group them into the aspect-oriented opinion sets. We associate PF and opinions by their cooccurrence in the reviews. Given a PF term pf, we first find out its nearest neighboring opinion word within the clause using rule-based linguistic analysis. We attach this nearest adjacent opinion word to pf. Then we check a polarity lexicon for the sentiment polarity of the opinion word, and attach the sentiment polarity to the semantic aspect of pf. The sentiment strength for a product aspect c_i is obtained by summing up all the attached sentiment orientation with the aspect. Finally, we summarize the cross-lingual/cross-culture difference in opinions, such as PF aspects ranking and finer-grained sentiment difference on each aspect.

In the opinion integration, by projecting the multilingual PFs into a unified semantic category framework, we may well organize all the opinions towards different aspects, and find out cross-lingual difference in sentiment distribution.

5. EXPERIMENTS

In this section, we evaluate the proposed approach using both English and Chinese customer reviews in cell phone and laptop domains. We first evaluate the performance of cross-lingual PF categorization in both domains in detail. Then, we discuss the cross-lingual difference in opinions.

5.1 Data

We build English and Chinese customer review corpus for cell phone and laptop domain (see Table 3). The English data for cell phone and laptop domains respectively are collected from the popular product review web sites. The Chinese data for cell phone and laptop domains respectively are collected from Chinese review web sites. We automatically extracted all the multilingual PFs from these data sets using the statistical-based PF extraction method presented in [8]. In the preprocess, we employ a Maximum Entropy part-of-speech (POS) tagger to generate POS tagging for English data, and employ HMM to segment Chinese words and generate POS tagging for Chinese data. All the PFs are translated by the machine translation engines.

Data Set	Cell Phone Domain		Laptop Domain	
	Words (M)	Reviews	Words(M)	Reviews
English Corpus	1.36	2,528	0.74	2,348
Chinese Corpus	2.02	8,181	1.30	21,442
Total	3.38	10,709	2.04	23,790

Table 3: Multilingual review data sets

No.	Aspects	All PF terms	PF term Distribution	
			English	Chinese
1	entertainment	599	236	363
2	appearance	509	141	368
3	keyboard	494	206	288
4	photography	468	156	312
5	os	451	313	138
6	user-interface	413	131	282
7	interface	385	259	126
8	SMS	350	115	235
9	screen	335	171	164
10	sound	321	126	195
11	storage	278	135	143
12	ringtone	239	80	159
13	browser	228	171	57
14	battery	193	97	96
15	email	159	155	4
Total	–	5422	2492	2930

Table 4: Cross-lingual PF categorization evaluation sets for cell phone domain

We build evaluation sets on 15 PF aspects of cell phone domain and 10 aspects of laptop domain (see Table 4 and 5). These aspects are concerned by most reviewers. Here, "SMS" denotes "*short message*". All the PFs are checked manually. If a PF term satisfies the specification of PF aspects, we give it the relevant label. The quality is ensured by cross-validation checking.

5.2 Experimental Results

In the experiments, we categorize the multilingual PFs of each domain into semantic aspects, and mine aspect-oriented cross-lingual difference in opinions. In the following, we will analyze these experimental results in detail.

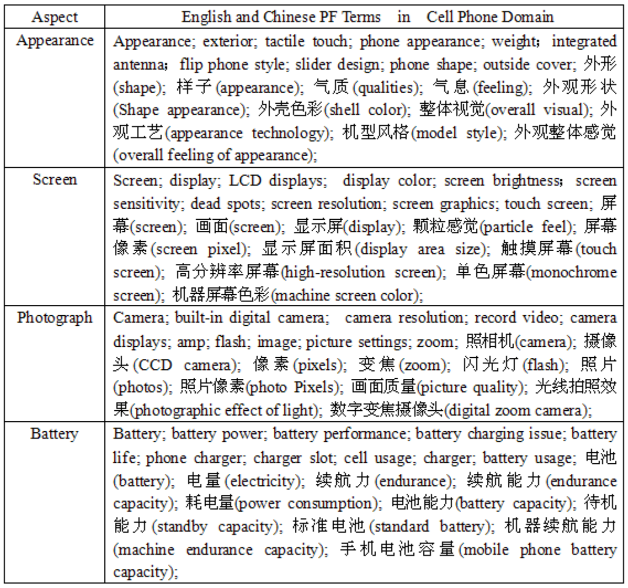
No.	Aspects	All PF terms	PF term distribution	
			English	Chinese
1	appearance	313	67	246
2	processor	178	125	53
3	keyboard	145	108	37
4	storage	142	99	43
5	screen	140	95	45
6	os	131	119	12
7	graphic	130	65	65
8	sound	97	42	55
9	battery	93	61	32
10	fan	37	11	26
Total	–	1406	792	614

Table 5: Cross-lingual PF categorization evaluation sets for laptop domain

5.2.1 *Cross-lingual PF Categorization*

In this section, we evaluate the performance of cross-lingual PF categorization in the given two domains. In the experiments, all the extracted multi-lingual PF terms in each domain are grouped into semantic aspects, respectively. The performance of PF categorization is evaluated using Rand Index [21], a measure of cluster similarity [3, 25, 4]. This measure varies from zero to one. Higher scores are better.

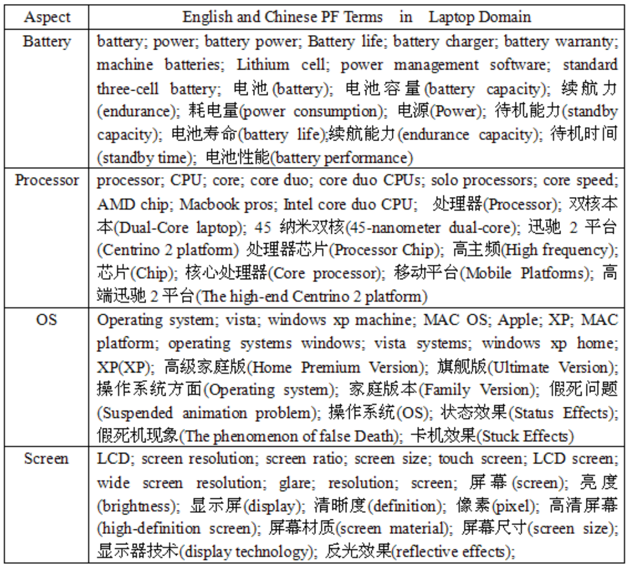
We compared our CLaSA method (see Algorithm 1 in Section 3.2.2) with *k-means* clustering [16] and LDA-based categorization method (denoted as *LDA-based*). In *k-means* clustering, each PF term is characterized by a bag of the current PF term pf, its machine translation pf^T, and their component words. In *LDA-based* categorization method, the virtual context document of each PF is also composed of pf, pf^T, and their component words. LDA-based PF categorization model is learned from the virtual context documents using Blei's algorithm [2]. Each PF is assigned to an aspect using this LDA-based model.



Aspect	English and Chinese PF Terms in Cell Phone Domain
Appearance	Appearance; exterior; tactile touch; phone appearance; weight； integrated antenna； flip phone style; slider design; phone shape; outside cover; 外形(shape); 样子(appearance); 气质(qualities); 气息(feeling); 外观形状(Shape appearance); 外壳色彩(shell color); 整体视觉(overall visual); 外观工艺(appearance technology); 机型风格(model style); 外观整体感觉(overall feeling of appearance);
Screen	Screen; display; LCD displays; display color; screen brightness； screen sensitivity; dead spots; screen resolution; screen graphics; touch screen; 屏幕(screen); 画面(screen); 显示屏(display); 颗粒感觉(particle feel); 屏幕像素(screen pixel); 显示屏面积(display area size); 触摸屏幕(touch screen); 高分辨率屏幕(high-resolution screen); 单色屏幕(monochrome screen); 机器屏幕色彩(machine screen color);
Photograph	Camera; built-in digital camera; camera resolution; record video; camera displays; amp; flash; image; picture settings; zoom; 照相机(camera); 摄像头(CCD camera); 像素(pixels); 变焦(zoom); 闪光灯(flash); 照片(photos); 照片像素(photo Pixels); 画面质量(picture quality); 光线拍照效果(photographic effect of light); 数字变焦摄像头(digital zoom camera);
Battery	Battery; battery power; battery performance; battery charging issue; battery life; phone charger; charger slot; cell usage; charger; battery usage; 电池(battery); 电量(electricity); 续航力(endurance); 续航能力(endurance capacity); 耗电量(power consumption); 电池能力(battery capacity); 待机能力(standby capacity); 标准电池(standard battery); 机器续航能力(machine endurance capacity); 手机电池容量(mobile phone battery capacity);

Figure 2: Cross-lingual PF categorization samples of *CLaSA* model in cell phone domain

In the experiments, all the multi-lingual PF terms in each domain are categorized using these methods, respectively. Experimental results show that *CLaSA* effectively groups the multi-lingual PFs into semantic aspects. Figure 2 and 3 list some categorization samples of *CLaSA* in cell phone and laptop domains, respectively. As shown, English and Chinese PF terms in each aspect are representative. Even though some PFs in the same aspect have little similarity on the lexical level, they are also correctly categorized. This shows that *CLaSA* model can effectively capture the cross-lingual latent semantic association among the PFs.



Aspect	English and Chinese PF Terms in Laptop Domain
Battery	battery; power; battery power; Battery life; battery charger; battery warranty; machine batteries; Lithium cell; power management software; standard three-cell battery; 电池(battery); 电池容量(battery capacity); 续航力(endurance); 耗电量(power consumption); 电源(Power); 待机能力(standby capacity); 电池寿命(battery life);续航能力(endurance capacity); 待机时间(standby time); 电池性能(battery performance)
Processor	processor; CPU; core; core duo; core duo CPUs; solo processors; core speed; AMD chip; Macbook pros; Intel core duo CPU; 处理器(Processor); 双核本本(Dual-Core laptop); 45 纳米双核(45-nanometer dual-core); 迅驰 2 平台(Centrino 2 platform) 处理器芯片(Processor Chip); 高主频(High frequency); 芯片(Chip); 核心处理器(Core processor); 移动平台(Mobile Platforms); 高端迅驰 2 平台(The high-end Centrino 2 platform)
OS	Operating system; vista; windows xp machine; MAC OS; Apple; XP; MAC platform; operating systems windows; vista systems; windows xp home; XP(XP); 高级家庭版(Home Premium Version); 旗舰版(Ultimate Version); 操作系统方面(Operating system); 家庭版本(Family Version); 假死问题(Suspended animation problem); 操作系统(OS); 状态效果(Status Effects); 假死机现象(The phenomenon of false Death); 卡机效果(Stuck Effects)
Screen	LCD; screen resolution; screen ratio; screen size; touch screen; LCD screen; wide screen resolution; glare; resolution; screen; 屏幕(screen); 亮度(brightness); 显示屏(display); 清晰度(definition); 像素(pixel); 高清屏幕(high-definition screen); 屏幕材质(screen material); 屏幕尺寸(screen size); 显示器技术(display technology); 反光效果(reflective effects);

Figure 3: Cross-lingual PF categorization samples of *CLaSA* model in laptop domain

Selecting the right number of topics is also an important problem in the PF categorization. A range of 50 to 300 topics is typically used in the topic modeling literature. 50 topics are often used for small collections and 300 for relatively large collections [26]. However, in the cross-lingual PF categorization, the number of topics might be set in a different range. It is confirmed here by our experiments with different values of K (10, 20, ...,100) in the two domains.

We evaluate the accuracy curves of these methods with different number of topics K in each domain. Experimental results show that *CLaSA* and *LDA-based* methods achieve better accuracy than *k-means* in each domain, as shown in Figure 4 and 5. The major reason for significant performance enhancement is that *CLaSA* and *LDA-based* methods effectively capture the latent semantic association clues from the reviews. Moreover, *CLaSA* also outperforms *LDA-based* method at all the operation points in each domain by using multi-dimension cross-lingual latent semantic clues. In the real applications, in order to provide non-trivial opinion mining and summarization, the number of the final product aspects is usually less than 20 topics. Experimental results show, at K=20 , the score of *CLaSA* method achieves 0.9150 and 0.8723 in cell phone and laptop domains, respectively. Compared with *K-means*, *CLaSA* significantly enhances the accuracy by 3.56% and 4.28% respectively. Compared with *LDA-based* method, *CLaSA* improves the accuracy by 1.98% and 0.88%, respectively. We perform t-tests on all the comparison experiments in all the domains (see Figure 4 and 5). p-value < 0.01 on both 20 comparison experiments of *k-means* and *CLaSA*, and 20 comparison experiments of *LDA-based* method and *CLaSA*. This shows that the enhancement is statistically significant.

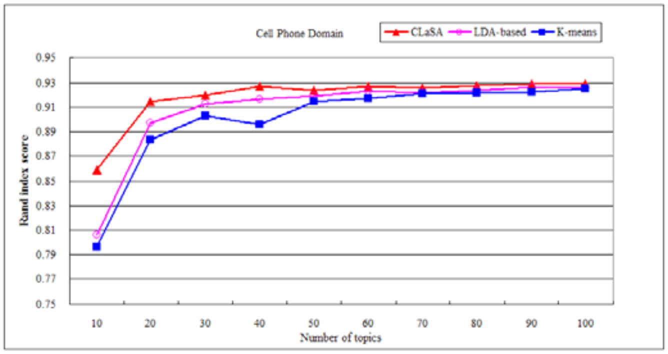


Figure 4: Cross-lingual PF categorization evaluation on cell phone domain with different number of topics

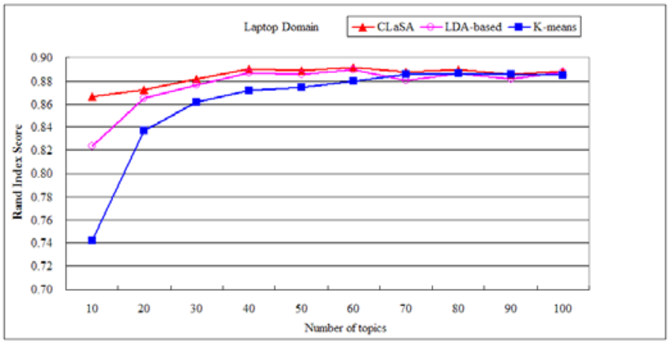


Figure 5: Cross-lingual PF categorization evaluation on laptop domain with different number of topics

Table 6 shows the performance comparison of these methods with different number of topics (K). The evaluation measure is average rand index score of each method on all the given domains with a predefined K. *CLaSA* always obtains much better average score than *k-means* and *LDA-based* methods at each predefined K. Moreover, the average score of *CLaSA* method is above 0.8937 at K=20.

We also evaluate the impact of CLaSA model on the monolingual PF categorization in cell phone and laptop domains. In this experiment, we respectively do English PF categorization and Chinese PF categorization using LDA, k-means and CLaSA. Both LDA and k-means construct the feature vector for each monolingual PF with a bag of the current PF and its component words. CLaSA model constructs the feature vector for each PF with the language-mixed virtual context document which consists of the current PF, its machine translation and latent semantic clues. Experimental results show that monolingual PF categorization using language-mixed virtual context documents also outperforms doing it in one language alone. Figure 6 shows the average performance of English PF categorization in both domains. The

Topics (K)	Average Rand Index Score			Δ P (%) over	
	K-means	LDA-based	CLaSA	(k-means)	(LDA-based)
10	0.7692	0.8148	0.8629	+12.18	+5.90
20	0.8601	0.881	0.8937	+3.91	+1.44
30	0.8825	0.8945	0.9008	+2.07	+0.70
40	0.8838	0.9017	0.9088	+2.83	+0.79
50	0.8946	0.9025	0.9065	+1.33	+0.44
60	0.8988	0.9063	0.9098	+1.22	+ 0.39
70	0.9034	0.9014	0.9068	+0.38	+0.60
80	0.904	0.9054	0.9089	+0.54	+0.39
90	0.9041	0.9043	0.9076	+0.39	+0.36
100	0.9051	0.9071	0.9088	+0.41	+0.19

Table 6: Comparison of *k-means* clustering, *LDA-based* method and *CLaSA* with different number of topics (K). The evaluation measure is the average rand index score of each method on all the domains with a predefined number-of-topics (K). ΔP denotes the percentage change in performance (measured in average rand index score) of *CLaSA* over *k-means* or *LDA-based* method.

evaluation measure is the average rand index score of each method on all the domains with a predefined number-of-topics (K). As shown, at each operation point, the average score of English PF categorization is significantly enhanced by CLaSA model. The similar trends are observed on Chinese PF categorization (see Figure 7). We perform t-tests on all the comparison experiments on mono-lingual PF categorization in both domains, p-value < 0.001 on both 40 comparison experiments of *k-means* and *CLaSA*. And p-value < 0.01 on 40 comparison experiments of *LDA* method and *CLaSA*. Both p-value show that the performance is significantly improved by CLaSA model.

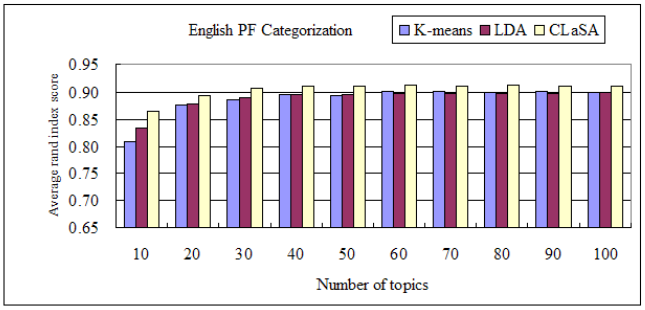


Figure 6: The impact of CLaSA model on English PF categorization. The evaluation measure is the average rand index score of each method on all the domains with a predefined number-of-topics (K).

All the above experimental results show that *CLaSA* produces better cross-lingual PF categorizations than *LDA-based* method and *K-means*. The major reason for the significant enhancement is that *CLaSA* better captures deeper cross-lingual latent semantic association among the PFs. *CLaSA* better approximates the underlying semantic distribution without using any labeled data.

5.2.2 Cross-lingual Difference in Opinions

In this section, we discuss the cross-lingual opinion mining results of *OpinionIt*. In the experiments, we use the above CLaSA model to categorize all the multilingual PFs. Based on the CLaSA-based PF categorization, *OpinionIt* further integrates opinions towards different aspects, and summarizes cross-lingual PF aspects and the finer-grained sentiment distribution difference in both cell phone and laptop review corpus.

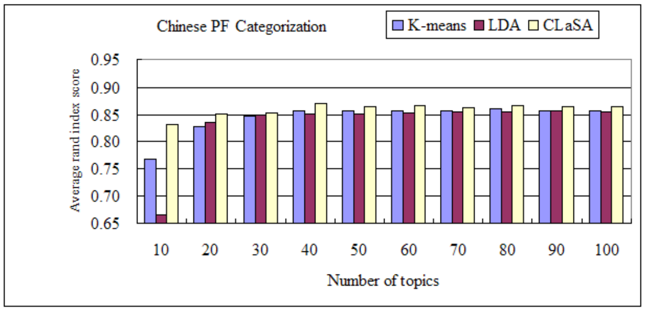


Figure 7: The impact of CLaSA model on Chinese PF categorization. The evaluation measure is the average rand index score of each method on all the domains with a predefined number-of-topics (K).

Table 7 shows the ranking lists of the PF aspects in cell phone and laptop domains. All the PF aspects are ranked according to their distribution ratio in the review corpus. Given an aspect c_i, the distribution ratio of c_i in the monolingual review corpus is defined as $\frac{n}{N}$, where, n is the number of the reviews in which c_i is reviewed; N is the total number of the reviews. Experimental results show that the ranking is varied with the languages. For example, in cell phone domain, *Appearance* is the top 1 aspect concerned in Chinese reviews while *OS* is the top 1 aspect concerned in English reviews. The different ranking in English and Chinese reviews indicates the cross-culture difference in the aspects concerned by people.

No.	Cell Phone Domain		Laptop Domain	
	English	Chinese	English	Chinese
1	os	appearance	screen	appearance
2	battery	screen	appearance	storage
3	screen	sound	keyboard	graphic
4	keyboard	battery	os	sound
5	appearance	keyboard	battery	keyboard
6	entertainment	entertainment	processor	screen
7	photography	photography	storage	fan
8	sound	SMS	graphic	battery
9	interface	ringtone	sound	processor
10	browser	os	fan	os
11	storage	storage	–	–
12	SMS	user-interface	–	–
13	email	interface	–	–
14	user-interface	browser	–	–
15	ringtone	email	–	–

Table 7: Ranking list of the PF aspects in English and Chinese reviews for cell phone and laptop domains

Figure 8 and 9 respectively demonstrate the distribution ratio of PF aspects in English and Chinese reviews for cell phone and laptop domains. From these experimental results, we may have the following observations. 1) In cell phone domain (see Figure 8), the advanced applications *Web Browser* and *Email* are more concerned by English customers than Chinese customers. The basic aspects *Appearance*, *Screen*, *Sound*, *Ringtone* and *SMS* are more concerned by Chinese customers. Only aspect *user-interface* has the similar distribution in both English and Chinese reviews. 2) In laptop domain (see Figure 9), the aspects *screen*, *os*, *processor*, *battery* and *keyboard* are more concerned by English reviewers. Only the aspects *appearance* and *fan* are more concerned by Chinese reviewers. From all the above finer-grained statistical distribution, one can easily find out the cross-lingual difference in customer concerned aspects.

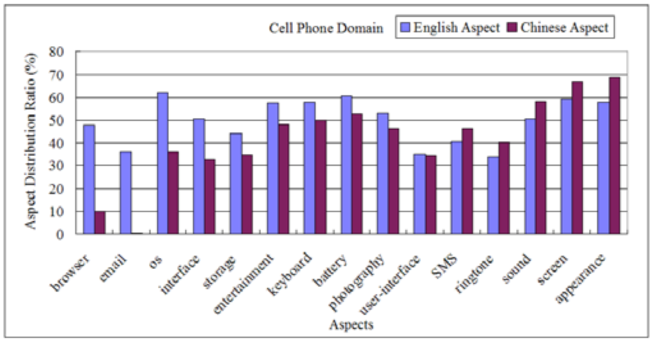


Figure 8: Cross-lingual difference in the distribution ratio of PF aspects in the cell phone reviews

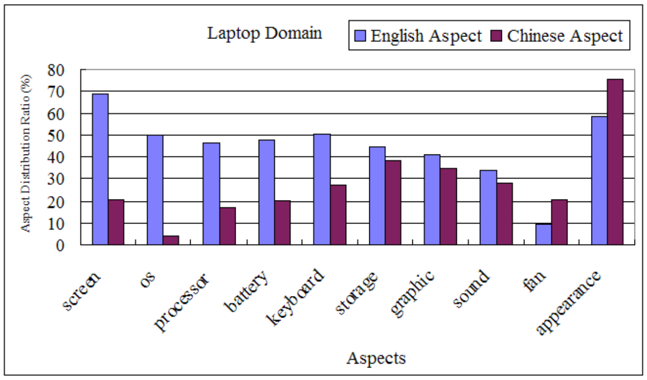


Figure 9: Cross-lingual difference in the distribution ratio of PF aspects in the laptop reviews

Figure 10 and 11 respectively demonstrate the sentiment distribution of each aspect in English and Chinese reviews for two popular cell phone brands. "*en*" and "*zh*" denote English and Chinese. Each bar in both figures shows the percentage of positive (above x-axis) and negative (below x-axis) opinions on an aspect in the domain. Given an aspect c_i, the overall distribution ratio of the positive (or negative) opinions on c_i is defined as $\frac{m}{M}$, where, m is the number of the positive (or negative) opinions on c_i, M is the total number of all the opinions in the monolingual corpus. The overall ratio is further normalized for cross-lingual comparison. These figures enable the user to clearly see cross-lingual difference in opinions. For example, for both brands, the advanced applications *browser* and *email* are more concerned by English customers while the basic aspects *appearance*, *SMS* and *sound* are more concerned by Chinese customers. Especially, for brand 2, Aspect *os* is more concerned by English customers (see Figure 11) while aspect *ringtone* is more concerned by Chinese customers. Moreover, English customers express more negative opinions about aspects *battery* and *keyboard* while more Chinese customers have positive opinions about both aspects. Regarding aspects *screen* and *storage* of both brands, there are similar sentiment distribution in both English and Chinese reviews.

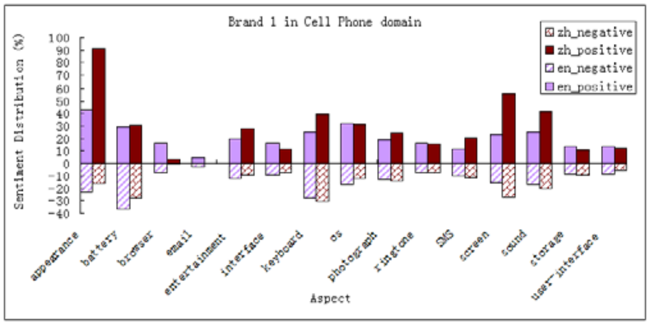


Figure 10: The sentiment distribution on PF aspects in English and Chinese reviews for brand 1

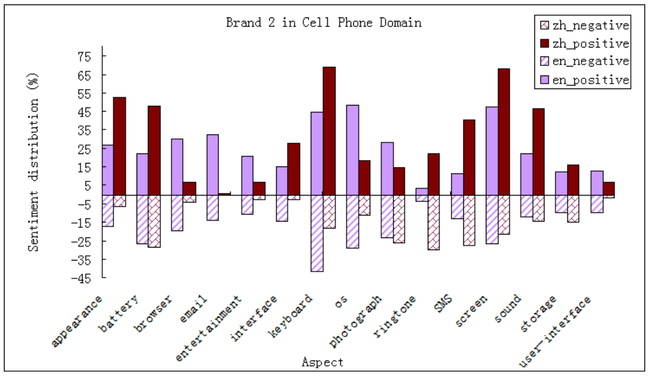


Figure 11: The sentiment distribution on PF aspects in English and Chinese reviews for brand 2

6. CONCLUSION

Cross-lingual opinion mining is a very challenging task since various opinions are written in different languages, and not well structured. Since people usually use different words to describe the same aspect in the multilingual reviews, PF categorization is critical for high-quality cross-lingual opinion mining. In our text mining system *OpinionIt*, we present an aspect-oriented cross-lingual opinion mining method with CLaSA model. The proposed method first constructs CLaSA model to categorize the PFs into semantic categories, and then summarizes cross-lingual difference in opinions toward different product aspects. One important contribution of our work is that we present a CLaSA-based unified cross-lingual PF categorization framework. With language-mixed virtual context documents, the proposed unsupervised CLaSA model can better capture the semantic association between different languages. Experimental results show that our method achieves better performance compared with the existing approaches. With the CLaSA model, *OpinionIt* can find out finer-grained cross-lingual/cross-culture difference in opinions by effectively joining reviews between languages. Our proposed method is quite general, and can be applied to integrate cross-lingual opinions about any product in any domain, thus potentially has many interesting applications.

7. REFERENCES

[1] M. Bautin, L. Vijayarenu, and S. Skiena. International sentiment analysis for news and blogs. In *Proceedings of 23rd AAAI Conference on Artificial Intelligence (AAAI'08)*, pages 19–26, 2008.

[2] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(5):993–1022, 2003.

[3] S. Branavan, H. Chen, J. Eisenstein, and R. Barzilay. Learning document-level semantic properties from free-text annotations. In *46th Annual Meeting of the Association for Computational Linguisticsm (ACL'08)*, 2008.

[4] C. Cardie and K. Wagstaff. Noun phrase coreference as clustering. In *Proceedings of the 1999 Conference on Empirical Methods in Natural Language Process (EMNLP'99)*, pages 82–89, 1999.

[5] C. Cesarano, A. Picariello, D. Reforgiato, and V. Subrahmanian. The oasys 2.0 opinion analysis system. In *Proceedings of 2007 International AAAI Conference on Weblogs and Social Media(ICWSM'07)*, pages 313–314, 2007.

[6] K. W. Church and P. Hanks. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[7] T. Fukuhara, T. Utsuro, and H. Nakagawa. Cross-lingual concern analysis from multilingual weblog articles. In *Proceedings of 6th Inter. Workshop on Social Intelligence Design*, pages 55–64, 2007.

[8] H. Guo, H. Zhu, Z. Guo, X. Zhang, and Z. Su. Product feature categorization with multi-level latent semantic association. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*, pages 1087–1096, 2009.

[9] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22th Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, 1999.

[10] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD-2004)*, 2004.

[11] M. Hu and B. Liu. Mining opinion features in customer reviews. In *Proceedings of AAAI-2004*, 2004.

[12] W. Jin, H. H. Ho, and R. K. Srihari. Opinionminer: A novel machine learning system for web opinion mining and extraction. In *Proceedings of KDD'09*, 2009.

[13] W. Li and A. McCallum. Pachinko allocation:dag-structured mixture models of topic correlations. In *Proceedings of the 2006 IEEE International Conference on Data Mining (ICDM'06)*, 2006.

[14] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of WWW'05*, pages 1024–1025, 2005.

[15] Y. Lu and C. Zhai. Opinion integration through semi-supervised topic modeling. In *Proceedings of WWW'08*, pages 121–130, 2008.

[16] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[17] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of WWW'07*, 2007.

[18] P. Melville, W. Gryc, and R. D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of KDD'09*, 2009.

[19] H. Nakasaki, M. Kawaba, T. Utsuro, and T. Fukuhara. Mining cross-lingual/cross-cultural differences in concerns and opinions in blogs. In *LNAI 5459*, pages 213–224, 2009.

[20] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL'05*, 2005.

[21] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[22] 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. In *Proceedings of the 17th international conference on World Wide Web (WWW'08)*, pages 959–968, 2008.

[23] I. Titov and R. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL'08*, pages 308–316, 2008.

[24] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of WWW'08*, 2008.

[25] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained k-means clustering with background knowledge. In *Proceedings of ICDM'01*, 2001.

[26] X. Wei and B. Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR06)*, 2006.

[27] T.-L. Wong, W. Lam, and T.-S. Wong. An unsupervised framework for extracting and normalizing product attributes from multiple web sites. In *Proceedings of SIGIR'08*, pages 35–41, 2008.

[28] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-2004)*, pages 743–748, 2004.

[29] L. Zhuang, F. Jing, and X. Zhu. Movie review mining and summarization. In *Proceedings of the 15th ACM Conference on Information and Knowledge Management (CIKM'06)*, pages 43–50, 2006.