2014 IEEE International Conference on Data Mining Workshop

A Localization Toolkit for SenticNet

Yunqing Xia*, Xiaoyu Li*†, Erik Cambria‡ and Amir Hussain§

*Department of Computer Science and Technology, TNList, Tsinghua University, Beijing 100084, China
Email: yqxia@tsinghua.edu.cn
†School of Communication, Beijing University of Post and Telecommunication, Beijing 10000, China
Email: 2011212836@bupt.edu.cn
‡School of Computer Engineering, Nanyang Technological University, 639798 Singapore
Email: cambria@ntu.edu.sg
§School of Natural Sciences, University of Stirling, Stirling FK9 4LA Scotland, UK
Email: ahu@cs.stir.ac.uk

***Abstract*—SenticNet is a popular resource for concept-level sentiment analysis. Because SenticNet was created specifically for opinion mining in English language, however, its localization can be very laborious. In this work, a toolkit for creating non-English versions of SenticNet in a time- and cost-effective way is proposed. This is achieved by exploiting online facilities such as Web dictionaries and translation engines. The challenging issues are three: firstly, when a Web lexicon is used, one sentiment concept in English can usually be mapped to multiple concepts in the local language. In this work, we develop a concept disambiguation algorithm to discover context within texts in the target language. Secondly, the polarity of some concepts in the local language may be different from the counterpart in English, which is referred to as language-dependent sentiment concepts. An algorithm is developed to detect sentiment conflict using sentiment annotation corpora in the two languages. Lastly, some sentiment concepts are not included in the local language after dictionary consulting and online translation. In this work, we develop a tool to extract these concepts from sentiment dictionary in the local language. Our practice and evaluation in constructing the Chinese version of SenticNet indicate that the proposed algorithms represent an effective toolkit for localizing SenticNet.**

I. Introduction

Popular approaches to sentiment analysis can be grouped into four main categories: keyword spotting, lexical affinity, statistical methods, and concept-based techniques [1]. While keyword spotting, lexical affinity and statistical methods have been thoroughly investigated by the natural language processing (NLP) community, concept-based techniques have gained increasing popularity in recent years.

Such methods exploit Web ontologies or semantic networks to accomplish concept-level text analysis, which enables systems to better grasp the conceptual and affective information associated with natural language opinions. Superior to purely syntactical techniques, in fact, concept-based approaches can detect subtly expressed sentiments [2].

A popular resource for concept-level sentiment analysis is SenticNet [3], an affective common-sense knowledge base that exploits an ensemble of graph-mining and dimensionality-reduction techniques to bridge the conceptual and affective gap between word-level natural language data and the concept-level sentiments conveyed by them.

SenticNet has been exploited for the development of several applications in many different fields including big social data analysis [4], human-computer interaction [5], pattern recognition [6], and e-health [7]. SenticNet, however, was created specifically for opinion mining in English language.

Its localization, i.e., its transposition into a different language, can be very laborious as it usually requires the manual translation of each concept into the local language. Automatic localization of sentiment analysis resources, in fact, is a very complex task, which involves the following key challenges:

1) Using a dictionary to directly map English terms into a target language is wrong as the correspondence between English and non-English words is usually one-to-many;
2) The polarity of some concepts, e.g., cultural-dependent multi-word expressions, in the local language may be opposite to the polarity of their counterparts in English;
3) After translation, some sentiment expressions, i.e., language-dependent concepts, may still be left untranslated either because they are out-of-vocabulary (OOV) concepts or because they are untranslatable.

In a Web where the proportion of non-English speakers is growing exponentially, the automatic localization of NLP resources is becoming increasingly important. Chinese, in particular, is poised to outpace English as the dominant language online in a few years' time.

So far, just a few isolated research endeavors have been undertaken to meet the demands of real-life Chinese web environments. NLP research endeavors, in fact, primarily depend on the availability of resources like lexicons and corpora, which are still very limited for sentiment analysis research in Chinese language.

To this end, we developed a Chinese version of SenticNet. Although specifically developed for Chinese language, the localization toolkit can potentially be applied to construct SenticNet in any existing language. In order to address the above-mentioned challenges, a concept disambiguation algorithm based on topic models is firstly developed so as to discover context within texts in the local language.

978-1-4799-4274-9/14 $31.00 © 2014 IEEE
DOI 10.1109/ICDMW.2014.179

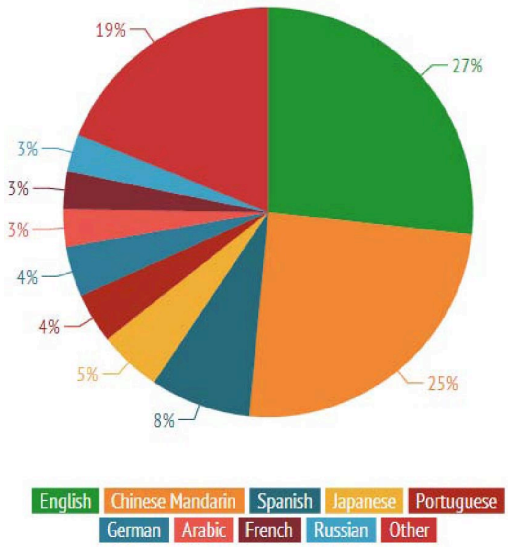

Fig. 1. Distribution of languages on the Web as of 2013

Additionally, a sentiment conflict detection algorithm is trained on annotated sentiment corpora in the two languages. Finally, a tool for extracting language-specific sentiment concepts from sentiment dictionaries in the local language is also developed.

We exploit Bing Online dictionary[1] and Google Translate[2], to create a first prototype of Chinese SenticNet. We then exploit Chinese Gigaword 2nd Edition[3] to build topic models so as to resolve ambiguity. Finally, OPINMINE [8] is used to discover language-specific sentiment concepts.

The remainder of this paper is as follows: Section II illustrates the architecture of the toolkit; Section III and IV describe the algorithms for single-word concept mapping and multi-word concept translation, respectively; Section V presents the polarity prediction algorithm; Section VI presents the tool for OOV sentiment concept detection; Section VII and VIII report the methodology adopted for constructing and evaluating Chinese SenticNet, respectively; finally, Section IX concludes the paper.

II. Architecture

As shown in Fig.2, the localization toolkit includes four main modules: a single-word concept localization tool, a multi-word concept localization tool, a concept polarity prediction tool and an OOV concept detection tool (implementation details are provided in the following four sections, respectively). For presentation convenience, we first give an example of a SenticNet concept in Fig.3.

As shown in Fig.3, SenticNet concepts consist of a concept string, five semantics, four sentics (i.e., Pleasantness, Attention, Sensitivity, and Aptitude), and polarity. The polarity value is calculated in terms of the four sentics according to the following formula:

$$p = \sum_{i=1}^{N} \frac{Plsnt(c_i) + |Attnt(c_i)| - |Snst(c_i)| + Aptit(c_i)}{3N}$$

where N is the number of concepts in a clause.

The concept string plays an identification role. The five semantics are concepts that are semantically relevant to the current concept. Such semantics can be viewed as context of the concept in resolving ambiguity.

This information is built through a semi-supervised approach that leverages on affective common-sense knowledge collected by means of crowdsourcing techniques and games with a purpose (GWAPs) [9]. More details on how SenticNet is built are provided in [3]. SenticNet is also accessible through an API[4].

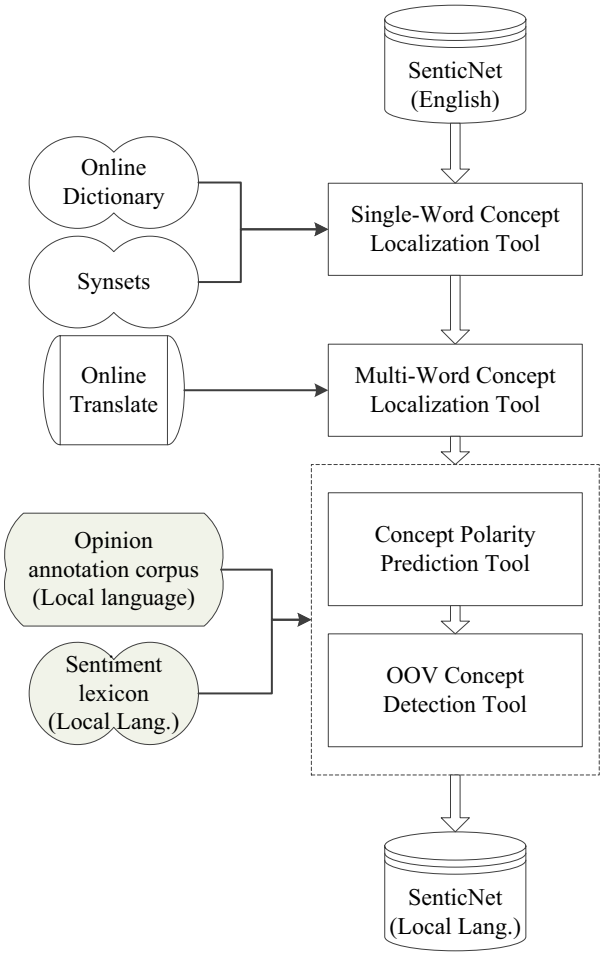

Fig. 2. SenticNet localization toolkit architecture

The localization task includes the following four steps:

[1]http://cn.bing.com/dict
[2]http://translate.google.com
[3]http://catalog.ldc.upenn.edu/LDC2005T14
[4]http://www.sentic.net/api

```
<?xml version="1.0" encoding="UTF-8"?>
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
    - <rdf:Description rdf:about="http://sentic.net/api/en/concept/strong">
          <rdf:type rdf:resource="http://sentic.net/api/concept"/>
          <text xmlns="http://sentic.net/api">strong</text>
          <semantics rdf:resource="http://sentic.net/api/en/concept/durable" xmlns="http://sentic.net/api"/>
          <semantics rdf:resource="http://sentic.net/api/en/concept/rigid" xmlns="http://sentic.net/api"/>
          <semantics rdf:resource="http://sentic.net/api/en/concept/sturdy" xmlns="http://sentic.net/api"/>
          <semantics rdf:resource="http://sentic.net/api/en/concept/hard" xmlns="http://sentic.net/api"/>
          <semantics rdf:resource="http://sentic.net/api/en/concept/heavy" xmlns="http://sentic.net/api"/>
          <pleasantness xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0</pleasantness>
          <attention xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.069</attention>
          <sensitivity xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0</sensitivity>
          <aptitude xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0</aptitude>
          <polarity xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.023</polarity>
    </rdf:Description>
</rdf:RDF>
```

Fig. 3. SenticNet entry for the single-word concept *strong*

Step 1: The concept (i.e., *strong* in Fig.3) is converted to the counterpart concepts in the local language.

Step 2: The five semantics are learned from the text corpus in the local language.

Step 3: The four sentics are transferred to the counterpart concepts that belong to the same context.

Step 4: For the OOV sentiment concepts in the local language, new concept nodes need to be created in SenticNet. The addition of new nodes, however, is quite time-consuming as it requires to define the full set of semantics and sentics associated with the OOV concept.

III. Single-Word Concept Mapping

For the single-word concepts, we rely on a Web dictionary to find the counterpart concepts in the local language. Consulting the dictionary is an easy task. If we use the concept word as input to the dictionary, however, we usually obtain a group of words in the local language. Such words may be related to more than one entry, which makes it impossible to perform a one-to-one mapping. Similarly to word sense disambiguation (WSD), concept ambiguity occurs constantly in dictionary consulting. A disambiguation algorithm is thus required to differentiate concepts.

A. Online Dictionary Consulting

Web dictionaries are available for many languages. For example, Bing provides a free English-Chinese dictionary through an API. Using the example of Fig.3, we obtain the following output for the concept string *strong*: {强壮;强健;有力;强劲;强烈;猛烈;坚固;巩固;坚牢;强效;烈性;厉害;浓烈;刺鼻;富有;有财力;资力雄厚;有势力;强大;优势}.

As seen from the above output, not all the counterpart Chinese words are sentiment words. Moreover, most of these words are shared by different concepts, after mapping is performed.

B. Concept Disambiguation

The previously obtained set of Chinese sentiment words can be remapped to the following concepts:

- *Powerful*: {强壮;强健;有力}
- *Intensive*: {强烈;强劲;猛烈}
- *Solid*: {坚固;巩固;坚牢}
- *Effective*: {强效;烈性;厉害;浓烈;刺鼻}
- *Wealthy*: {富有;有财力;资力雄厚}
- *Advantageous*: {有势力;强大;优势}

For each concept set, the listed words are synonymous. Such grouping is performed by exploiting an extended version of HIT IR-Lab Tongyici Cilin [10]. Cilin is a free online resource that contains 77,343 Chinese words within 17,817 synsets. For the synonymous words belonging to a concept, we select the first word as string of the Chinese concept.

Anther important task is to find the five semantics associated with each concept. To this end, we apply the LDA topic model algorithm [11] on the Chinese Gigaword 2nd Edition[5] in order to obtain semantic contexts. In particular, we first use all the words in Chinese as keywords to retrieve relevant sentences. We then run LDA on the sentences to obtain topics. According to the topic-word distribution matrix, we select the top 5 words that are most likely associated to each topic.

To improve retrieval efficiency, we adopt an information-retrieval solution. We first split articles into sentences based on punctuation stop marks, then segment all sentences into words with Stanford segmentor[6], and use Solr[7] to index the sentences. The standard BM25 algorithm[8] is used to calculate query-document relevance score.

[5]http://catalog.ldc.upenn.edu/LDC2005T14
[6]http://nlp.stanford.edu/software/segmenter.shtml
[7]http://lucene.apache.org/solr
[8]http://en.wikipedia.org/wiki/Okapi_BM25

```
<?xml version="1.0" encoding="UTF-8"?>
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
    - <rdf:Description rdf:about="http://sentic.net/api/en/concept/mathematical_skill">
        <rdf:type rdf:resource="http://sentic.net/api/concept"/><text xmlns="http://sentic.net/api">mathematical skill</text>
        <semantics rdf:resource="http://sentic.net/api/en/concept/remember" xmlns="http://sentic.net/api"/>
        <semantics rdf:resource="http://sentic.net/api/en/concept/cogitate" xmlns="http://sentic.net/api"/>
        <semantics rdf:resource="http://sentic.net/api/en/concept/contemplate" xmlns="http://sentic.net/api"/>
        <semantics rdf:resource="http://sentic.net/api/en/concept/remember_phone_number" xmlns="http://sentic.net/api"/>
        <semantics rdf:resource="http://sentic.net/api/en/concept/think" xmlns="http://sentic.net/api"/>
        <pleasantness xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">-0.083</pleasantness>
        <attention xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.146</attention>
        <sensitivity xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.056</sensitivity>
        <aptitude xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.07</aptitude>
        <polarity xmlns="http://sentic.net/api" rdf:datatype="http://www.w3.org/2001/XMLSchema#float">0.026</polarity>
    </rdf:Description>
</rdf:RDF>
```

Fig. 4. SenticNet entry for the multi-word concept *mathematical skill*

C. Concept Merging

It is very common that the same concept is associated with different English words. For example, both *strong* and *firm* are mapped to the Chinese concept 坚固(jian1 gu4, strong). In order to avoid the presence of duplicates, the inclusion of any of the two English words should be omitted. To this end, we exploit HIT IR-Lab Tongyici Cilin [10] again to check whether two or more English terms are synonyms in Chinese. When this happens, we select the English concept with higher polarity in absolute value in order to build a more information-rich knowledge base.

IV. Multi-Word Concept Translation

There are a lot of multi-word concepts in English SenticNet (Fig.4). These are much easier to handle as, unlike single words, they are unambiguous because they already carry some context.

We simply use Google Translate[9] to achieve the mapping task. Besides solving the problem of ambiguity, multi-word expressions also enable a better translation because of the way Google Translate is designed.

To obtain the five semantics, we adopt an approach which is almost identical to the one used for the single-word concepts. The only difference intact, lies in that the query contains only the multi-word concept string. As we assume the multi-word concepts are unambiguous, we simply run TextRank [12] to obtain the top 5 keywords from the relevant sentences.

V. Polarity of Concepts

In the previous steps, sentiment polarity assignment was never addressed. In general, the polarity of a translated sentiment concept is identical in two languages. However, there are a few exceptions. For example, *dragon* is a cruel animal in English while in Chinese, it is a lucky animal. This causes polarity conflict in the two languages. We call these concepts language-dependent concepts.

[9]https://translate.google.com

We make use of the opinion annotation corpus in the local language to resolve polarity of the concepts in the local language. For each concept, we search within the concept string in the corpus. If an opinion is matched, we check annotation of the opinion and obtain the polarity. If no opinion is matched, we then make use of point-wise mutual informal (PMI) equation [13] to calculate its polarity tendency. The algorithm is elaborated below.

For two given words w_1 and w_2, PMI value is calculated as follows:

$$PMI(w_1, w_2) = log_2 \frac{p(w_1 \& w_2)}{p(w_1)p(w_2)} \quad (1)$$

where $p(w_1)$ represents the probability that word w_1 occurs, $p(w_2)$ the probability that word w_2 occurs, and $p(w_1 \& w_2)$ represent the probability that word w_1 and w_2 co-occur in one context.

Viewing the polarity-known concepts as seeds, we obtain a group of positive concepts $\{c_i^+\}_{i=1,\ldots,N}$ and another group of negative concepts $\{c_j^-\}_{j=1,\ldots,K}$. For a polarity-unknown concept c, we calculate sentiment orientation SO value of concept c with the following equation.

$$SO(c) = \frac{\sum_{i=1}^{N} PMI(c, c_i^+)}{N} - \frac{\sum_{j=1}^{K} PMI(c, c_j^-)}{K} \quad (2)$$

We assign the SO value as polarity of the polarity-unknown concept.

VI. OOV Sentiment Concepts

No matter how complete the English SenticNet is, there are inevitably OOV sentiment concepts in the local language due to cultural differences. Localized versions of SenticNet cannot achieve a sound coverage unless some OOV sentiment concepts are discovered and inserted in SenticNet.

As sentiment analysis research rapidly evolves, sentiment lexicons has been constructed for many different languages.

TABLE I. STATISTICS OF CHINESE SENTICNET.

Item	Statistics
# of concepts	32,478
# of concepts from the English SenticNet	29,543
# of OOV concepts in Chinese	2,935
# of language-dependent concepts	1,426
# of single-word concepts	18,439
# of multi-word concepts	14,039

HowNet, for example, is a sentiment lexicon[10] for Chinese, which we make use of for discovering OOV sentiment concepts. We adopt a simple matching strategy. For each word in the local language sentiment lexicon, we first match the synsets to find synonyms by using IR-Lab Tongyici Cilin [10]. The synonyms are then matched against SenticNet entries. If a match is obtained, this word is not considered as an OOV sentiment word; otherwise, an OOV sentiment concept is detected. We adopt a similar approach to discover the five semantics and predict polarity of the OOV sentiment concept.

VII. PRACTICE IN CONSTRUCTING CHINESE SENTICNET

In this section, we report our practice in constructing Chinese SenticNet with the localization toolkit, and evaluate its quality. We start from SenticNet [3], which contains 30,000 affect-driven concepts. We work on the RDF format data directly. The following resources are used in the localization task:

- Bing online English-Chinese dictionary
- IR-Lab Tongyici Cilin (extended) in Chinese
- Google English-Chinese Translate online
- OPINMINE Chinese opinion annotation corpus
- HowNet Chinese sentiment lexicon

Statistics of Chinese SenticNet are presented in Table I. Firstly, we found the volume of Chinese SenticNet to be bigger than English SenticNet. We obtained 32,478 concepts in Chinese SenticNet. This is mainly due to the fact that Bing dictionary and HowNet lexicon added concepts. We also found that 1,426 concepts are language-dependent within the multilingual opinion corpora.

Secondly, we found out that 457 concepts from English SenticNet are not included in the Chinese version. This is because of the PMI-based polarity detection algorithm: the Chinese counterparts of these English concepts were discarded because they carry very little polarity orientation.

Thirdly, we discover only 2,935 OOV Chinese concepts, which are not included in the English SenticNet. This indicates that our OOV sentiment concept detection algorithm can be further improved. We leave this to future work.

Lastly, we found out that single-word concepts are still the majority. However, multi-word concepts may carry much clearer sentiment. We plan to revise the sentiment detection algorithm to discover more sentiment concepts in Chinese.

[10]http://www.keenage.com

TABLE II. EVALUATION RESULTS OF CHINESE SENTICNET.

Metric	Results
Relevance	0.892
Accuracy of English concepts	0.993
Accuracy of OOV concepts	0.862

VIII. EVALUATION

A. Setup

Metrics: Because there is no gold standard for the automatic evaluation of the localization toolkit, we employ students to review every concept in Chinese SenticNet according to the following aspects:

- Relevance: Relevance measures whether the discovered semantics are relevant to the root concept. We then calculate the overall relevance which indicates the percentage of relevant semantics in all the semantics. Relevance of a concept is judged by two postgraduate students separately. The agreement is 0.945.
- Accuracy: Accuracy measures correctness of the predicted polarity of the root concept. We then calculate the overall accuracy which indicates the percentage of the concepts which are assigned correct polarity. The judgment on sentiment accuracy is made by two postgraduate students separately. The agreement is 0.901. Note that the accuracy is calculated on English concepts and OOV concepts separately.

We found no prior work attempting a concept-level localization in the literature. Hence, we hereby only report the qualitative evaluation of Chinese SenticNet as a standalone tool.

B. Results and Discussions

Evaluation results are presented in Table II.

Firstly, relevance of Chinese SenticNet is pretty high. This indicates that the semantics obtained by the localization tool are reasonable.

Secondly, machine-predicted polarity is also acceptable. For the English concepts, the error rate is 0.007, which comes from the language-dependent concepts. This indicates that the localization toolkit is reliable in producing the initial version of SenticNet in the local language.

While any language resource released to the public should be almost flawless, our localization toolkit can save a great deal of time. With this toolkit, researchers can concentrate on more challenging tasks such as language-specific sentiment concept detection.

IX. CONCLUSION

This paper presents a localization toolkit which convert English SenticNet to multiple languages automatically. The contributions of the paper are summarized as follows: firstly, online facilities are exploited to translate English SenticNet to the local language; secondly, powerful algorithms such as topic modeling and sentiment orientation prediction are integrated to help resolve concept ambiguity automatically.

The main aim of the proposed toolkit, however, is just the construction of a SenticNet prototype in the local language. Before releasing the resource to the public, intensive review is necessary. Meanwhile, some continuous expansion work is key to make SenticNet as complete as possible. In the future, we plan to conduct more research on the automatic detection of language-specific sentiment concept.

ACKNOWLEDGMENT

This paper is partially supported by National Natural Science Foundation of China (NSFC: 61272233, 6141101023). We thank the anonymous reviewers for the valuable comments.

REFERENCES

[1] E. Cambria, B. Schuller, B. Liu, H. Wang, and C. Havasi, "Knowledge-based approaches to concept-level sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 12–14, 2013.

[2] S. Poria, E. Cambria, G. Winterstein, and G.-B. Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis," *Knowledge-Based Systems*, vol. 69, pp. 45–63, 2014.

[3] E. Cambria, D. Olsher, and D. Rajagopal, "SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis," in *AAAI*, Quebec City, 2014, pp. 1515–1521.

[4] E. Cambria, H. Wang, and B. White, "Guest editorial: Big social data analysis," *Knowledge-Based Systems*, vol. 69, pp. 1–2, 2014.

[5] E. Cambria and A. Hussain, "Sentic album: Content-, concept-, and context-based online personal photo management system," *Cognitive Computation*, vol. 4, no. 4, pp. 477–496, 2012.

[6] Q. Wang, E. Cambria, C. Liu, and A. Hussain, "Common sense knowledge for handwritten chinese recognition," *Cognitive Computation*, vol. 5, no. 2, pp. 234–242, 2013.

[7] E. Cambria, A. Hussain, T. Durrani, C. Havasi, C. Eckl, and J. Munro, "Sentic computing for patient centered application," in *IEEE ICSP*, Beijing, 2010, pp. 1279–1282.

[8] R. Xu, Y. Xia, K.-F. Wong, and W. Li, "Opinion annotation in on-line chinese product reviews," in *LREC*, 2008.

[9] E. Cambria, Y. Xia, and A. Hussain, "Affective common sense knowledge acquisition for sentiment analysis," in *LREC*, Istanbul, 2012, pp. 3580–3585.

[10] W. Che, Z. Li, and T. Liu, "Ltp: A chinese language technology platform," in *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, ser. COLING '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 13–16.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[12] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *EMNLP*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 404–411.

[13] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 417–424.