An Analysis of Radicals-based Features in Subjectivity Classification on Simplified Chinese Sentences

Ge Xu
Minjiang University
xuge@pku.edu.cn

Chu-Ren Huang
Hong Kong Polytechnic University
Churen.huang@poly.edu.hk

Abstract

Chinese radicals are linguistic elements smaller than Chinese characters[1]. Normally, a radical is a semantic category and almost all characters contain radicals or are radicals themselves. In subjectivity classification on sentences, we can use radicals to represent characters, which reduce the scale of word space while keep the subjectivity information.

In this paper, we manually labeled a character set to build a high-quality radical-character mapping, and then the mapping is used to generalize character-based features with radicals. In experiments, we at first evaluated the performance when directly generalizing characters with radicals, and then offer a hypothesis that can reduce noises.

Experiments show that this approach based on our hypothesis can reduce feature space while keep or improve the performance, which is especially useful when the training samples are scarce.

keyword: sentiment analysis, subjectivity classification, radical, Chinese character

1 Introduction

In sentiment analysis, an important task is subjectivity classification on sentences, which means classify sentences as subjective or objective. This step's performance greatly affects the following processing that is related with polarity or emotion etc. Here we offer two sentences from NTCIR6 training corpus for subjectivity classification.

1. 邻近的永乐国小大部分的教室也被列为危楼。 Most classrooms of Yongle elementary school nearby are also classified as dangerous buildings.
2. 寄读生涯孩子累家长烦。 Boarding school life makes students tired and their parents irritable.

For the first sentence, although 危(dangerous) is a sentiment character, it is used to modify building, so the semantic emphasis of 危楼(dangerous building) is building, and also 危楼(dangerous building) is somewhat known as a term, so normally is regarded as objective, thus the whole sentence is labeled as objective.

For the second sentence, the subjectivity mainly comes from 累(tired) and 烦(irritable). These two characters are also with different level of subjectivity. "tired" is a physical experience, compared with 烦(irritable), it is somewhat "objective". But 烦(irritable) can surely make the sentence subjective. If we take a further step, we can see that 烦(irritable) has a Chinese radical 火(fire), which can derive concepts of sentiment from linguistic perspective.

In real system, to label subjectivity sentences is of high cost especially when high quality is required. As we know, the size of common Chinese characters is around several thousands, while the size of radicals in Chinese is only around several hundreds, if we use the radicals to generalize character, it may overcome to some extent the sparseness problem

[1]We use the terminology "character" for "Chinese character" in this paper.

28th Pacific Asia Conference on Language, Information and Computation pages 495–502

when training model and reduce the time and space required.

2 Related work

Sentiment classification on texts has been studied by many researchers, such as (Goldberg and Zhu, 2006; Pang and Lee, 2005) etc. Normally, machine learning-based methods dominate the filed, and much emphasis is put on polarity instead of subjectivity. Furthermore, compared with English, subjectivity classification on Chinese is relatively few. In the following, we pay more attention to work on Chinese, subjectivity and Chinese radicals.

Yao and Peng (2007) used 7 features to describe a text, which include "if a personal pronoun occurs in the sentence?", "if interjection occurs in the sentence?", etc. A SVM-based method offered the best performance (F-value 0.938) in their experiments. The work used a small corpus which includes 359 texts (191 subjective and 168 objective).

Li et al. (2006) made a detail comparison between words and character-bigrams when they are used to represent features in text classification, and concluded that Chinese character bigrams are better than words in feature representation for text classification. In our experiments, we followed some experimental configuration in (Li et al., 2006) and put more emphasis on evaluating the performance of subjectivity classification using radical representation

Qiu et al. (2009) presented an approach to guess word's sense by its components(characters), they used the LC(lexical compositionality) principle: "The words formed by similar constituents in the same mode fall into the same semantic category". When we use radicals to generalize characters, we are following the similar principle. If two characters share the same radical, they may fall into the same semantic category.

Huang et al. (2008) presented a qualia structure to analyze how characters derived from radicals, they classified the derived concepts of character radicals into 7 categories, expanded from the original four qualia aspects of Formal, Constitutive, Agentive, and Telic. This structure is useful when we label radicals for subjectivity classification, because characters derived by similar path may have similar concept, and when we use radicals to generalize characters, we can choose an accurate semantic category (finer than a radical) to avoid semantic roughness. For example, a frequent radical, such as 人(human), can derive many characters. In these characters, some are persons with certain identification，such as 仙(fairy),侠(swordsman) and 佛(Buddha); some are descriptive such as 仁(benevolent),俊(handsome) and 傻(stupid). Other possible concepts derived from 人(human) is not listed due to space limit. Considering this, we have to define finer semantic categories for the radical 人(human); Otherwise, different concepts will be grouped together, making the generalization in feature construction error-prone.

3 The basics of radicals

Chinese characters have a history of over 5000 years. They evolved from pictographs to nowadays characters after all sorts of unification and simplification. Basically, there are four ways to create a character: pictographs(象形) , ideographs(指示), logical aggregates(会意), phonograms(形声).

1. pictographs(象形): Character is similar with the entity in the world. Examples include 伞 for "umbrella", and 木 for "tree".

2. ideograph(指示): For instance, 刀 is "knife", and placing an indicator in the knife makes 刃, an ideograph for "blade". Other common examples are 上(up) and 下 (down).

3. logical aggregates(会意)：For instance, 木 (tree) is a pictograph of a tree, and putting two 木 together makes 林 , meaning forest. The difference between ideograph and logical aggregates lies in that the indicator for ideograph is normally not a radical, much more like a stroke of a Chinese character; while logical aggregate characters contain at least two radicals.

4. phonograms(形声)：It is also titled semantic-phonetic compounds, or phono-semantic compounds. According to (Xu, 121), approximately 82 percent of characters are classified into this category, and also the largest group of characters in modern Chinese. A phonogram character includes two parts: a pictograph,

which indicates the semantics of the character, and a phonetic part, which is a character itself and indicate how the phonogram character is pronounced. For example, 榕(banyan tree) contains two parts: 木(tree) and 容(pronounced as róng), 木 indicates that 榕 is a kind of tree, and 容 indicates that 榕 is pronounced as róng.

Roughly speaking, in Chinese, radicals are the minimum semantic units[2]. Normally, a character is composed of radical(s) or a radical itself. Let us check how four types of character-formations (pictograph,ideograph,logical aggregate,phonogram) are related with radicals.

1. For pictograph characters, normally they are radicals, such as 木(tree),鱼(fish),鹿(deer), 田(cropland) etc.

2. For ideograph characters, normally they are based on a pictograph, and add some stroke(s).

3. For logical aggregate characters, they contain two or more radicals.

4. For any phonogram character, one of two parts in the character is a radical and indicates the semantics of the character.

So we can see that radicals are closely related with character, we can know the rough semantic of a character by its radical(s). If the given NLP task required a semantic granularity coarser than radical-level, we can use radicals to assistant the task without sacrificing accuracy.

In "ShuoWenJieZi"(Xu, 121), all Chinese characters are classified as derived from 540 radicals. Nowadays, many of 540 radicals have been deprecated or are seldom used, so the size of common and active radicals is around 200. In (Zhou and Huang, 2005), ranked by how many characters a radical can derive, the top 20 radicals can cover 4425 of 9353 characters in (Xu, 121). Such radicals are closely related with human life, such as 水(water),艸(grass),木(tree), 手(hand),心(heart),言(speak) etc. When a radical can derive many characters, normally the semantics is derived into several categories, we will give more details in section 4.2 how we process this issue.

Of cause, there exists some case that radicals fail to indicate the semantics of characters. For example, 笨(stupid) contains two radicals: 竹(bamboo) above and 本(base) at the bottom which contains the radical 木(tree). Perhaps due to complicated evolution, it is hard to connect the semantics with either of the two radicals. By experience, such phenomenon is scarce, accounting for only a small portion in all Chinese characters. So in most cases, for a character, we can relate it to a radical which indicate its semantics.

4 Radical labeling on a Chinese character set

For subjectivity classification on Chinese sentences in our experiments, we manually created a radical-character mapping. For this task, two problems have to be considered:

- Choosing a Chinese character set
- Design a labeling schema

Furthermore, another important problem should be noticed. The corpus and the character set we used in experiments is simplified Chinese. However, in order to obtain high-quality radical-character mapping, we used traditional Chinese character to analyze radicals. For example, 云(cloud) is the simplified character of 雲(cloud) which has the radical 雨(rain),and we think that 云(cloud) has the radical 雨(rain) although this radical has been omitted after Chinese character simplification.

4.1 Choosing a Chinese character set

We have four choices for a Chinese character set, see table 1 for more details.

Note that, apart from "ShuoWenJieZi" character set which is traditional Chinese, other three character sets should use traditional Chinese character as a bridge to identify radicals in characters. In our experiments, we choose the first level character set of GB2312, which complies with national regulation of P.R.China and includes frequent (compared with the second level character set) Chinese characters which can cover most of Chinese conversation. We do not

[2]In our paper, we do not define radical strictly as in some linguistic literature As long as a element in character can be used to represent semantics and indivisible, we accept it as a radical

Table 1: Introduction of available Chinese character sets

1	XinHua Dictionary. The newest version contains 11200+ entries (characters). This dictionary accompanies almost every Chinese people from primary school, and each Chinese character in this dictionary is listed with all its senses.
2	GB2312 character set. The first level contains 3755 characters, and the second level contains 3008 characters. This Chinese character set complies with national regulation, which makes it easy to introduce and deliver. Most important of all, the first level is frequent Chinese character, which is naturally proper when processing big corpus and for obtaining comprehensive performance.
3	The character set in ShuoWenJieZi. It has 9353 characters. This dictionary is edited by radicals. But it suffers two several problems: This book's author didn't know oracle bone inscriptions, so many radical explanation is wrong; Many characters in this dictionary are deprecated.
4	Chinese character set from a given corpus. The size varies and although this is the minimum cost when develop a system on text processing, this set is too specific and is limited when transferred to other applications.

choose the character set in XinHua dictionary because it is a bit too large for one annotator to label.

4.2 Labeling schema

At first, we collected from internet all sorts of radical resources. We only consider those resources which list all Chinese characters sharing one radical in one line, and the first Chinese character is normally the radical. A clip of the finally collected radical resources is shown in figure 1.

Based on this resource, we will use the following labeling schema:

1. The first Chinese character should represent its original semantics of the radical that derives the characters in the line. For example, lot of Chinese characters with the radical 月(moon) is in fact related with 肉(meat), and 月(moon) is the pictograph of meat; so we put the 肉(meat) at the first of the line, which looks like:"肉肌肉肝肛肚肪肩股肢胚胎胃脆.....". The processing can make it more readable and understandable when we use radicals to replaces Chinese characters.

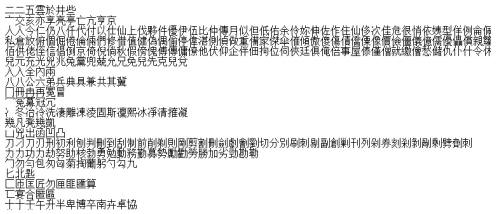

Figure 1: Collected radical resources

2. If some radical is only used for looking up in a dictionary, it is omitted. For example, a lot of Chinese characters are arranged in the line started with 一(one), "一丁丐万丈上下且.....", but in terms of semantics, 一(one) has little relationship with other Chinese character. So we omit the whole line.

3. If a Chinese character contains more than one radicals, choose the radical more similar in semantics with the character. For example, 娶(marry a woman) can be related to two parts, 取(fetch)and 女(female). Furthermore, 取(fetch) can be analyzed as 耳(ear) and 又(again), and 又(again) is the pictograph of 手(hand). So, in a more general level, 娶(marry a woman) contain the abstract semantics of the behavior of 手(hand), so the radical for 娶(marry a woman) should be 手(hand). In another situation, although a character contains more than one radical, but none is closely semantics-related. For example, 笨(stupid) includes two radicals 竹(bamboo) and 木(tree), but neither has the same semantic category with 笨(stupid), so we tend to regard such character as an independent character.

4. Since our main task is subjectivity classification, we require that each line is subjective or objective. For example, many characters share the radical 女(female), some characters are ob-

jective, they are mainly all sorts of female relatives such as “姐姑妈姨婆奶妹”; and some are subjective, such as “奴奸妒妓婀妙娟娥”.You also may note that the subjective ones contain both positive ones(婀妙娟娥) and negative ones(奴奸妒妓), since we do not distinguish polarities in our classification, we put both in one line.

5. Some radicals can derive too many character, and such radicals are normally closely related with human life, such as 人(human),口(mouth) and 手(hand) etc. In this situation, the radicals must be further divided.

In (Huang et al., 2008), the authors use Pustejovsky’s Qualia Structures base and observe the analysis on the definitions in “ShuoWenJieZi”, and then classify the derived concepts of character radicals into 7 categories , expanded from the original four qualia aspects of Formal, Constitutive, Agentive, and Telic, as shown in table 2.

We would refer to this schema in our labeling practicing while adjust and modify according to actual conditions.

4.3 Labeling practice

The labeling costs the first author approximately half a day with the help of a electrical dictionary[3]. Some radicals are easy to label. For example, all characters contain radical 父(father) are 父爸爹爺, which are fathers or grandfathers.

Once the size of the characters that a radical derived become large, it can derive different semantic categories. We used the Qualia Structures mentioned in (Huang et al., 2008) to create finer categories for a radicals. Several cases are listed as following:

1. Constitutive:鱗鰓鰾(various parts of a fish)

2. Formal-vision:鱼鲤鲸鲍.....(various types of fishes)

3. Descriptive:鲜(delicious)

The above is for the radical 鱼(fish).

[3]http://cn.bing.com/dict/

Table 2: Seven categories of derived concepts from radicals

Formal	This category can be further divided into 5 small categories: ”sense,” ”characteristic,” ”proper names,” and ”atypical.” The ”sense” categories can be further divided into 5 small categories: ”vision,” ”hearing,” ”smelling,” and ”taste.”
Constitutive	This category can be further divided into 3 small categories: ”part,””member,” and ”group.”
Telic	Concepts related to function or usage
Participant	Words are classified into this category when the definition in ’ShuoWenJieZi’ mentions the participant involved.
Participating	According to different events, concepts are divided into 6 small categories:”action,” ”state,” ”purpose,” ”function,” ”tool,” and ”others.”
Descriptive	This category can be further divided into two categories: ”active” and ”state.”
Agentive	The relationship between the radical and its meaning cluster coming from production or giving birth are classified in to agentive.

1. Constitutive:木 杈 本 末 林 根 梢 森 树 枝 果(various parts of a tree)

2. Formal-vision: 柳 栗 桑 桐 梨 棉 梅 枣 棕.....(various types of tree)

3. Telic/Agentive:梗 柯 柄 框 案 梁 梳 棋 棚.....(various types of components made by wood, various wood buildings)

The above is for the radical 木(tree).

1. Constitutive:叶 苗 蕊 蒂 茎 芯 藤 菁 苞 芽 茸(various parts of grass)

2. Formal-vision: 草 艾 芭 芥 芹 芝 茶 荔.....(various types of grass)

3. Descriptive:芬 芳 苦 茂 茫 芜 萧 菲 苍 蔼 萌(various characteristics of grass)

The above is for the radical 艸(grass).

The first-level character set of GB2312 contains 3755 characters, and some characters will be removed according to labeling schema in section 4.2, so the size of the final character set is smaller than 3755.

5 Experiment

In this section, we aim to evaluate how the generalization affects the subjectivity classification on Chinese simplified sentences when we use radicals to generalize characters.

5.1 The corpus

The NTCIR (NII-NACSIS Test Collection for Information Retrieval) workshops have been organized since 1999. In the sixth NTCIR Workshop (NTCIR6 for short), five subtasks are set in the evaluation, one of which is mandatory, which is to decide whether each sentence expresses an opinion or not. In another word, the subtask is a binary subjectivity classification on all sentences. The pilot task has tracks in three languages: Chinese, English, and Japanese. In this paper, we use its Chinese corpus for our experiments.

In our paper, the lenient evaluation metric is adopted, where two of the three annotators must agree for a value to be included in the gold standard. There are around 9000 sentences in the corpus, in which subjective sentences account for 60% roughly.

We use ICTCLAS[4] package to perform word segmentation and POS tagging, during which Specification for Corpus Processing at Peking University in (Shiwen Yu, 2003) is adopted.

5.2 Results of experiments

We used Weka[5] package for our experiments. According to research work on Chinese text classification(Li et al., 2006), SVM with linear kernel is a good classifier for such task, so we do not evaluate how various classifiers affect the performance, and put more emphasis on how feature are represented. Four-fold cross-validation is chosen.

Table 3: Comparison of different feature representations

Key_Dataset	Accuracy
radical_unigram	73.171%
radical_unibigram	75.033%
radical_bigram	75.303%
char_unigram	73.420%
char_unibigram	76.028%
char_bigram	76.050%
word_unigram	73.398%
word_unibigram	**76.548**%
word_bigram	74.026%
wordRadical_unigram	73.074%
wordRadical_unibigram	76.255%
wordRadical_bigram	73.745%
pos_unigram	73.117%
pos_unibigram	76.504%
pos_bigram	73.540%
posRadical_unigram	72.911%
posRadical_unibigram	76.310%
posRadical_bigram	73.788%

In the table 3, “unigram”,“bigram”,“unibigram” mean three types of n-gram; “char” means that a sentence is seen as sequence of characters, and “radical” means that each char is generalized to a radical or is kept if it contains no radical; “word” means we

[4]http://www.ictclas.cn

[5]http://www.cs.waikato.ac.nz/ml/weka/

see a sentence as a sequence of words after using word segmentation tools, and "wordRadical" means to generalize characters in words; 'pos' and "posRadical" are the POS version of "word" and "wordRadical".

According to table 3, for a char, a word and a word with tag, directly generalizing them by radicals will decrease the performance a little. Such phenomenon can be explained as that some noise will be incurred when generalize words or character-bigrams by radicals. For example, when using radicals, 抨击(denounce),提拔(promote),投掷(throw) are all generalized to 手手，because the three words are composed of two characters containing 手(hand). However, we know that these three words are of different semantic categories, of different subjectivity and even of different polarity.

A way to reduce such noise is based on a hypothesis in the next section.

5.3 A Radical-based Representation

Hypothesis: For two character bigrams, if they share a character in the same position and the other two character have the same radical, these two bigrams are in the same semantic category.

For example, 袜子(sock),袄子(a short Chinese-style coat),袖子(sleeve) have the same character 子(suffix for thing) in second position, and the first character 袜袄袖 are all derived from radical 衣(cloth). So, under our hypothesis, 袜子(sock),袄子(a short Chinese-style coat),袖子(sleeve) should fall into same semantic class, namely 'cloth'. Other examples are listed in table 4.

Of cause, there are counterexamples. When checking the corpus, we find that 应该(should) and 应试(take an examination) start by the same character 应(response) and the second character share the radical 言(speak). However, 应该(should) contain subjectivity to some extent, but 应试(take an examination) is an objective word. Such error comes from that derivation complexity of characters. The original meaning of 该(should) is a promise, but nowadays the meaning of 'promise' has been seldom used, and almost have no connection with 言(speak). Such error suggested that we should pay much attention on character's present usage when labeling radicals since most corpora given are not ancient.

We design an experiment to investigate how the hypothesis works and analyze the experimental results. Since 'unigram+bigram' performance best in table 4, it is used as default setting. The experimental result is shown in table 5.

Table 4: Examples of hypothesis

梨　花,杨花,杏　花,樱花,棉　花,梅花	the first characters all share a 木(tree) radical, the second character is the same. Each word is a kind of flower.
说话，讲话，训话，谈话	the first characters all share a 木(tree) radical, the second character is the same. Each word is a kind of speaking.
老爹，老爸，老爷，老父	the second characters all share a 父(father) radical, the first character is the same. Each word is "father" or "grandfather".

Table 5: Comparison on hypothesis and other generalizations

word	76.5476%
wordRadical	76.2554%
wordRadical_Hypothesis	**76.7424%**
pos	76.5043%
posRadical	76.3095%
posRadical_Hypothesis	76.6667%

"wordRadical_Hypothesis" and "posRadical_Hypothesis" mean processing the corpus using the hypothesis on "word" and "word with pos" representation respectively. Briefly speaking, based on the hypothesis, we at first find all the groups with same semantics, which means all words in a group should share one character and the other characters should contain the same radical. We can iterate this process from 2 character words to 3 characters, and so on. Finally, we got a set of groups, each group contain a set of words which belong to the same semantic category according to our hypothesis. In generalizing features, we use the first word in a group to label all the words in the group when processing the corpus.

The results show that such hypothesis can im-

prove the performance by a small margin. At first, the improvement is due to using the hypothesis, so some noises are removed. "posRadical_Hypothesis"is especially useful when part-of-speech tag can be used to reduce the generalization noise. For example, 下流(obscene) and 下海(go to sea, or go into business) is in the same group based on hypothesis, but they belong to different semantic categories and have different subjectivity. When POS is considered, 下流(obscene) is an adjective while 下海(go to sea, or go into business) is a verb, so they can be divided into different categories, which helps to reduce the noise when generalizing.

The improvement is not obvious enough because the words in groups is relatively small compared to the whole word space. In our experiments, there are 18099 words (without POS tag) in the corpus, but only 486 groups. Furthermore, most of the groups contain only two words or normally low-frequency words, so the impact is limited. Such a problem is supposed to be improved by labeling a bigger character set and by using other generalization strategies.

6 Conclusion and Future work

In this paper, we evaluate how subjectivity classification on Chinese sentences performs when radicals are used to generalize characters, and offer a hypothesis that can be used to find groups with the same semantic categories. All words in a group belong to the same semantic category, so the group ID can be used to label any word in it without decreasing the classification performance. Although the improvement on performance is not obvious enough, by manual checking the group, we find the quality is very high (which to some extent explains the amount of groups and amount of the words in groups are not very large.), which can guarantee that the improvement, although not obvious, is steady.

In the future, we will pay attention to two problems:1) label a larger character set with higher quality;2) explore new ways that can utilize radicals to obtain better performance.

Acknowledgements

References

Andrew B. Goldberg and Jerry Zhu. 2006. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *TextGraphs: HLT/NAACL Workshop on Graph-based Algorithms for Natural Language Processing*.

Chu-Ren Huang, Ya-Jun Yang, and Sheng-Yi Chen. 2008. An ontology of chinese radicals: Concept derivation and knowledge representation based on the semantic symbols of the four hoofed-mammals. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, pages 189–196, The University of the Philippines Visayas Cebu College, Cebu City, Philippines, November. De La Salle University, Manila, Philippines.

Jingyang Li, Maosong Sun, and Xian Zhang. 2006. A comparison and semi-quantitative analysis of words and character-bigrams as features in chinese text categorization. In *ACL*.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 115–124.

Likun Qiu, Kai Zhao, and Changjian Hu. 2009. A hybrid model for sense guessing of chinese unknown words. In *PACLIC*, pages 464–473.

Bin Swen Bao-Bao Chang Shiwen Yu, Huiming Duan. 2003. Specification for corpus processing at peking university: Word segmentation, pos tagging and phonetic notation. *Journal of Chinese Language and Computing*, 13.

Shen Xu. 121. 說文解字*ShuoWenJieZi*.

Tianfang Yao and Siwei Peng. 2007. 汉语主客观文本分类方法的研究. In 第三届全国信息检索与内容安全学术会议.

Yamin Zhou and Chu-Ren Huang. 2005. Construction of a knowledge structure based chinese radicals. In *The Sixth Chinese Lexical Semantics Workshop. Xiamen.*