Mining for Opinions Across Domains: A Cross-Language Study[⋆]

Anna Kravchenko

Higher School of Economics,
Research and Educational Center of Information Management Technologies
Kirpichnaya ul. 33/4, 105679 Moscow, Russia

Abstract. An important task in opinion mining is detecting subjective expressions in texts and distinguishing them from factual information. High lexicon diversity between different domains excludes the possibility of formulating universal rules that would work for any area of knowledge. In this article we suggest a solution for this problem. We define the features that most opinionated sentences share and propose a cross-language classification of subjective expressions, illustrated by examples in Russian, English and Chinese. We also propose an algorithm based on this classification that generates a set of extraction patterns for any domain from a corpus of untagged texts. The corpus requires no additional preparation except for POS-tagging. The effectiveness of the proposed approach is evaluated for English and Russian on collections of approximately 300 000 sentences each, gathered from three different domains: user reviews on movies, headphones and photo cameras.

Keywords: opinion mining, sentiment analysis

1 Introduction

One of the main challenges of opinion mining is that subjective expressions vary profoundly, depending on the domain. The exact same word or phrase may or nay not be considered opinionated in different contexts. For example, "short battery life" is clearly a negative opinion, and "short article" is simply a literature genre.

There is a current trend to focus only on machine learning techniques as a workaround for this problem, entirely dismissing the underlying linguistic structure, but we strongly believe it is essential to take it into account as well.

It is easy to see that there are properties that subjective sentences share across domains:

- Syntactic structure of subjective expressions is similar and domain-independent,

[⋆] This work was conducted with financial support from the Government of the Russian Federation (Russian Ministry of Science and Education) under contract № 13.G25.31.0096 on "Creating high-tech production of cross-platform systems for processing unstructured information based on open source software to improve management innovation in companies in modern Russia".

– Some subjective words and expressions are domain-independent ("fine", "terrific", "happy"),
– Opinionated sentences usually contain more than one subjective expression and often occur next to each other in texts.

Those properties allow us to find opinionated sentences in the text, using domain-independent words as pivots, and then to extract domain-specific expressions from them, based on the expected syntactic structure of opinion. Those expressions can be further transformed into extraction patterns and used in mining algorithms.

Therefore it allows us to extract a lexicon of opinionated expressions for such diverse areas as, for example, political news, reviews on movies and reviews on cameras.

It is important that no manual tagging of the processed texts is reqired, which minimizes the need for human participation.

2 Types of opinionated sentences

As it has been mentioned earlier, opinionated expressions share syntactic structure between domains, but there are different ways of expressing an opinion. We propose a classification of subjective expressions, based on study of English, Russian and Chinese. Those languages were chosen as the most heterogeneous examples - Russian has free word order and a well-developed morphology, in English the word order is fixed and morphology is significantly less complex, and Chinese also has fixed word order and very poor morphology. It is also important that Chinese is a Non-Indo-European language.

We will use the term **object** to denote the target entity that has been commented on. We will also use the notions of proposition. A proposition is the semantic core of a clause or a sentence that is constant, despite changes in such things as the voice or illocutionary force of the clause.

Subjective expressions can be divided into the following classes:

Class 1 **Explicit opinion:**

Feature or characteristics with evaluative meaning is directly attributed to the object.
It can be expressed with the following types of propositions:

1.a. **Complete proposition, opinion is expressed by a noun phrase.**

Examples:
English:
the situation seems bad, John is an outstanding painter
Russian:
наушники дрянь, Иванов плохой руководитель
earphones garbage, *Ivanov bad manager*
Chinese (Pinyin):

Ěrj shì fèiwù, Zhang shì ygè huài de jngl.
earphones is garbage, Zhang is single bad of manager

The proposition can be used as a complete sentence, a copula verb is used (zero copula in Russian, "shì" in Chinese).
Not all examples are consistent for better illustration of material.

1.b. **Complete proposition, opinion is expressed by an adjective phrase.**

Examples:
English:
the sound is terrific
Russian:
звук замечательный
sound terrific
Chinese:
Shngyn hěn dà
sound very good

2.a. **Incomplete proposition, opinion is convoluted with the noun phrase. Genitive case.**

Examples:
English:
photo's great quality, great quality of photos
Russian:
высокое качество фотографий.
hight quality-GEN photo-GEN-PL
Chinese:
doesn't occur in Chinese

It is important that the word "quality" itself bears affective connotation in this case. For example, "big TV screen" would be of type 3.

2.b. **Incomplete proposition, opinion is convoluted with the noun phrase pointing to the object. Nominative case.**

Examples:
English:
high quality photos
Russian:
известный преступник Петров
known criminal Petrov
Chinese:
doesn't occur in Chinese

3. **Incomplete proposition, opinion is expressed by adjective.**

   Examples:
   English:
   amazing design
   Russian:
   замечательный дизайн
   amazing design
   Chinese:
   Yuzhì de yìngxiàng
   high quality image

4. **Complete proposition, describing a situation involving the object. Opinion is expressed by a verb phrase.**

   Examples:
   English:
   the phone broke quickly
   Russian:
   телефон быстро сломался
   phone quickly broke-PAST
   Chinese:
   Diànhuà hěn kuài dpòle
   phone very quickly broke

Class 2. **Direct affective connotation.**

Object is characterised by its relation to entities with strong affective connotations. For example, "The President fights against corruption", "people criticize the government".

This type is expressed by a complete proposition, the semantic orientation of opinion is formed by the semantic orientation of the predicate and the associated entity.

Class 3. **Associated affective connotation.**

Object is characterised by a class of situations appearing in the same text or sentence, but not related directly to the object. For example, "Impoverishment risks", "bought a new pair".

This type can be expressed by a proposition of any form and is extremely hard to detect with natural language processing methods.

Analysis of classes 2 and 3 is very complex and requires deep syntactic analysis, for this reason we will only focus on expressions of class 1. We will also show that this is sufficient in most applications.

3 Extracting subjective expressions

A lot of opinionated sentences contain more than one subjective expression, and some of those are may be domain-independent. Therefore, if a sentence contains a domain-independent subjective expression and a new domain-specific one, we can extract the latter, if it matches any of the syntactic patterns that we expect.

Most of the time we can also 'guess' its semantic orientation, using the semantic orientatin of the pivot word. It can act as an additional argument when dealing with ambiguity.

This tagging methods was first proposed by Ellen Rillof [11], though she used shallow parsing instead of syntactic patterns. It requires a large enough corpus to process, but unannotated texts are easy to come by, so even if the classifier can label only 30% of the sentences as subjective, it will still produce a large collection of labeled sentences.

Example (known and new words are highlighted in bold and italic respectively):

These are the **best** closed-back headphones I've heard at this price, *bass is intense*, *highs are not shrill*, *no sound leak*, *comfortable design.*

We can extract the following expressions from the example:

bass is intense, [N, V, Adj], type 1.b
highs are not shrill, [N, V, Adj], type 1.b
no sound leak, [Part, N, N], type 2b
comfortable design, [Adj, N], type 3

Those expressions (we will call them **segments**) then can be further transfomed into lexical patterns. Some segments may contain name of the exact model they are describing, while they can in fact be applied to any other model or it's feature. It is important to replace those names with some universal label.

The method is as follows: Each segment is first converted to a sequence. Each sequence element is a word, which is represented by both the word itself and its POS tag in a set. In the training data, all object features or objects' names are labeled and replaced by the label $feature according to the original segment syntactic structure.

For example, the sentence segment, "Included memory is stingy", is turned into the sequence:

{included, Adv}{memory, N}{is, V}{stingy, Adj}.

After labeling, it becomes an extraction pattern (note that "memory" is an object feature):

{included, Adv}{$feature, N}{is, V}{stingy, Adj},

Feature extraction is performed by matching the patterns with each sentence segment in a new review to extract object features. That is, the word in the sentence segment that matches $feature in a pattern is extracted

A similar method is described by Bing Liu [8].

4 Filtering

Not all extracted patterns will indeed be subjective. Choosing which patterns to keep usually requires a human expert. Riloff also proposed a method for minimizing human participation. The method is based on a tendency of subjective expressions to reappear in multiple subjective sentences in the text more often than the expressions extracted by mistake.

All extraction patterns are ranked using a conditional probability measure: the probability that a sentence is subjective given that a specific extraction pattern appears in it.

The exact formula is:

$$P(subjective/pattern_i) = subjfreq(pattern_i)/freq(pattern_i),$$

where $subjfreq(pattern_i)$ is the frequency of $pattern_i$ in subjective training sentences, and $freq(pattern_i)$ is the frequency of $pattern_i$ in all training sequences.

A thresholds are used to select extraction patterns. We choose extraction patterns for which $pr(subkective/pattern_i) > \theta$. The threshold is chosen manually.

5 Evaluation

Currently the methods of testing sentiment analysis systems are not fully developed. For this reason we use a method based on subjective evaluations of small text collection by an expert.

Expert marks each pattern as evaluative or extracted by mistake, and precision is then calculated using the following formula:

$$P = N_{subj}/N_{all},$$

where N_{subj} is the number of correctly extracted patterns and N_{all} is the number of all extracted patterns. We do not evaluate recall, because of the amount of expert work it requires and because this method does not provide high recall by design, which can be compensated by the corpus size and automatic tagging.

For the evaluation process two corpora of approximately 300,000 sentences each were collected for three languages. All three consisted of three parts - reviews on photo cameras, earphones and movies.

Results show that the algorithm manages to extract opinionated phrases from texts of all three domains, though the accuracy differs. For domains with objective evaluation criteria and relatively low lexical variability (for example, reviews on earphones and photo cameras) shows good precision: 52% before filtering and 80% after filtering for Russian and 67% and 83% acorrdingly for English, with $\theta = 0,9$. For movie reviews precision was much lower, 29% before filtering and 64,3% after filtering for Russian and 31% and 68% for English with $\theta = 0, 6$.

Precision can be made higher by increasing θ, but it lowers recall valye and requires a significantly larger corpus,

6 Conclusion

As results show, the proposed method achieves high enough precision for texts on certain subjects and can be further used as a component of opinion extraction system. Presicion can be enhanced by improving the size and quality of the training corpora.

It is important that the method requires almost no manual preparation, and a collection of texts on certain topic can be easily acquired for example by searching for specific category in online stores.

The direction of further work is creating a full opinion mining based on the proposed algorithm and classification.

References

1. Carenini G., Pauls A.: Multi-Document Summarization of Evaluative Text in Proceedings of EACL 2006, Trento, Italy, pp. 305-312. (2006)
2. Ermakov A.E., Kiselev S.L.: Linguistic Model For Computational Sentiment Analysis of Media, Proceedings of the International Conference Dialog 2005, Moscow. (2005)
3. Hatzivassiloglou V., McKeown K.: Predicting the semantic orientation of adjectives in Proceedings of ACL/EACL 1997, Madrid, Spain, Complutense University of Madrid, pp. 174-181. (1997)
4. Hu M. and Liu B.: Mining and Summarizing Customer Reviews in Proceedings of KDD-2004, Seattle, WA, 2004, pp. 168-177. (2004)
5. Hu M. and Liu B.: Mining Opinion features in Customer Reviews in Proceedings of AAAI'04, Boston, Massachusetts, USA: AAAI Press, pp. 755-760. (2004)
6. Kobayashi N., Inui K., Tateishi K., Fukushima T.: Collecting Evaluative Expressions for Opinion Extraction in Proceedings of IJCNLP-2004, Berlin, Germany: Springer, pp. 596-605. (2004)
7. Liu, B.: Web Data Mining, Springer, Berlin. (2007)
8. Liu, B.: Sentiment Analysis and Subjectivity in Handbook of Natural Language Processing, Second Edition, Chapman and Hall/CRC, NY, USA, pp. 257-282. (2010)
9. Nozhov I. (2003) Morphologic and Syntactic Text Processing (Models and Computations), available at http://www.aot.ru/docs/Nozhov/msot.pdf
10. Popescu A.M., Etzioni O.: Product features and Opinions from Reviews in Proceedings of HLT-EMNLP 2005. Vancouver, Canada: ACL, pp. 339-346. (2005)
11. Riloff E, Wiebe J.: Learning Extraction Patterns for Subjective Expressions, in Proceedings of EMNLP-03, Sapporo, Japan, pp. 97-104. (2003)
12. Turney P.: Inference of Semantic Orientation from Association, available at `http://cogprints.org/3164/01/turney-littman-acm.pdf` (2003)
13. Turney P.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews in Proceedings of ACL 2002, Philadelphia, PA, U.S.A., pp. 417-424. (2002)