Multilingual Opinion Holder and Target Extraction using Knowledge-Poor Techniques

Taras Zagibalov, John Carroll

University of Sussex
UK
{T.Zagibalov,J.A.Carroll}@sussex.ac.uk

Abstract

We describe an approach to multilingual sentiment analysis, in particular opinion holder and opinion target extraction, which requires no annotated data and minimal language-specific input. The approach is based on unsupervised, knowledge-poor techniques which facilitate adaptation to new languages and domains. The system's results are comparable to those of supervised, language-specific systems previously applied to the NTCIR-7 MOAT evaluation data.

Keywords: opinion mining, multilingual natural language processing

1. Introduction

1.1 Sentiment analysis

Sentiment analysis (or opinion mining) is concerned not with the topic or factual content in a document, but rather with the opinion expressed in a document. Sentiment analysis has often been broken down into a set of sub-tasks, such as subjectivity classification, opinion orientation detection, opinion holder and opinion target extraction, and feature-based opinion mining. Opinion orientation is usually a three-way classification of positive, negative or neutral, and can be applied to different levels of the text: phrases, sentences, documents or collections of documents. An opinion may have a holder (a person or a group that expresses an opinion) and a target (an object which is being discussed or evaluated). Feature-based opinion mining tries to find opinions about particular features of a product or service (as opposed to an overall opinion about something).

1.2 Motivation

Ways in which opinions are expressed can vary not only between languages, but also within languages (so-called "domain-dependency"). A major current challenge is to be able to automatically extract sentiment information from a variety of documents in different languages and from different domains. Most existent approaches are based on adapting systems designed for one language (or domain) to another. Obviously, there are differences between cultures, languages and even within a language (consider the difference between evaluations of company financial prospects in a business newspaper and reviews of a hard-rock festival in a participant's blog). Such differences make adaptation problematic. Porting to new languages is even more difficult. To address these issues we describe a novel, knowledge-poor unsupervised method for opinion mining. The underlying idea of the approach is to extract all required information from the text which needs to be processed. We describe our approach in terms of an implemented system which extracts opinion holders and targets from documents in English, Chinese and Japanese[1].

The first author was supported by the Ford Foundation International Fellowships Program. We also wish to particulary aknowledge the thoughtful contributions of the anonymous rviewers, who made a lot of helpful suggestions and additions to this paper.

2. Related Work

2.1 Opinion Holder Extraction

Choi et al. (2005) consider opinion holder extraction to be an information extraction task and use a combination of two techniques: named entity recognition (Conditional random Fields) and information extraction (AutoSlog). The former models source identification as a sequence tagging task, the latter learns extraction patterns.

Bloom et al (2006) describes an opinion holder extraction approach based on a hand-built lexicon, a combination of heuristic shallow parsing and dependency parsing, and expectation-maximization word sense disambiguation; they match phrases in the text with domain-dependent holder type taxonomies.

Kim and Hovy (2006) used machine learning technique for opinion holder extraction. As features for their Maximum Entropy classifier they used selected structural features from a deep parse, based on a frame representation of opinionated expressions. The frame was built around an opinion word, and semantic relations between it and opinion holder and target were investigated. Such relations used semantic role labelling within the frames.

Kim et al. (2008) exploited a set of communication and appraisal verbs, SentiWordNet, a named entity recognizer, and a syntactic parser for opinion holder extraction. In each sentence they looked for the most opinionated word and then ascended the tree to its first ancestor node with verb part of speech, and looked for its subject (a noun phrase) that may contain opinion holder candidates. If a subject was not found, then "author" was set as the opinion holder of the sentence. If a subject was found, then from the NP chunk, any named entities or opinion holder candidates were extracted as the opinion holder. If no named entity or opinion holder candidate was found, then the holder was set as the "author" of the

[1]As a further test of the adaptivity of our approach, we note that none of the developers of our system knows Japanese.

document. Regardless of the previous step, if a sentence included quotation marks, then the speaker of the quote was extracted as the opinion holder.

Seki (2008) used an author and authority classification approach as a basis for holder detection. Seki's system was based on the features selected from the significance of frequency in training corpora, and classified sentences into opinionated sentences expressed from an author viewpoint or from an authority viewpoint. These differentiations were passed into opinion holder identification system, which treated the two kinds in different ways: author opinion holder and authority opinion holders were extracted with different sets of rules.

2.2 Opinion Target Extraction

For opinion target extraction, Kim and Hovy (2006) used the same approach as they used for the opinion holder extraction: semantic role labelling.

Bloom et al (2006) also used a similar technique for both tasks: their manually created taxonomies also included opinion targets.

Reasoning that an opinion target shares similar features with opinion holder (a noun phrase, but acting as object rather than subject), Kim et al. (2008) used a technique similar to the one used by Kim and Hovy (2006) for opinion holder extraction: they used structural features for machine learning. More specifically, they proposed a statistical machine learning technique based on syntactic features (syntactic path and dependency) and other heuristic features, such as topic words and named entity.

Gamon et al. (2005) used a clustering technique to find a product feature taxonomy. The algorithm used a stop-word list, which should not be used for building clusters, and 'go-words' known to be salient in the domain. Once sentences had been clustered according to the product feature taxonomy, they were processed by a sentiment classifier trained on a corpus bootstrapped from a small manually-created corpus.

3. Data

For our experiments we used the NTCIR-7 MOAT (Multilingual Opinion Analysis Task) English, Chinese and Japanese test data collections. The English data runs from 1998 to 2001 with texts from the Mainichi Daily News, Korea Times, Xinhua News, Hong Kong Standard, and the Straits Times. It consists of 142 documents split into 14 topics (4312 sentences). The Simplified Chinese data contains documents from Xinhua News and Lianhe Zaobao from 1998 to 2001, it consists of 252 documents in 14 topics (4877 sentences). The Japanese data consists of 249 Japanese news items from 1998 to 2001 from the Mainichi newspapers split into 18 topics (5885 sentences). All documents in the test corpus in each language were annotated using a pool of six annotators (Seki et al., 2008).

4. Opinion Holder and Opinion Target Extraction

We use a knowledge-poor language independent approach with some simple linguistic typology, similar to the one described by Bender (2009).

Our opinion holder and opinion target extraction system consists of two major parts: a core system implementing a general approach to the extraction task, and a small set of language-specific extensions. The basic idea of this approach is based on the assumption that opinion holders and opinion targets are words or phrases which are topic-related and tend not to occur in other topics. A further assumption is that each language has markers of subjectivity, and surface clues which can be used to find syntactic subjects. This set of assumptions together with a small amount of language-specific information constitutes the minimal language description.

4.1 Overview of the Approach

Our first assumption is that opinion holders and opinion targets are topic-related (with the exception of pronouns and generic phrases like *our correspondent*)[2]. So to implement our approach we first find topical words – words that are strongly related to the topic of a given text. The problem we immediately face is that there is no common language-independent notion of what a word is, which is compounded by the fact that in many Asian languages one cannot find even a graphical word (a sequence of characters / letters separated by whitespace or a punctuation mark). In order to minimise language-specific input (such as word lists or automatic segmenters), we have to find sequences that could be used as 'basic units'. This step is done by finding longest common strings amongst all text in the document collection being processed, with punctuation marks serving as delimiters. For example in the following two sentences, the underlined part (translated as *US Federal Reserve*) is the longest common string:

但该行预期，美国联邦储备局将会通过积极减息挽救经济，布什政府预料也将会增加支出，这对经济将起稳定作用。

美国联邦储备委员会 6 日再次减息 0.5 个百分点,将联邦基金利率和贴现率分别下调到 2%和 1.5%,以刺激经济复苏。

Of course, the resulting list contains a lot of noise. This problem is dealt with by filtering out those items that occur in too many different topics. Such items are filtered out on the basis of the number of different topics they occur in. For the experiments described here only one threshold was used: a lexical item is regarded a topical word if it is used in no more than 50% of the topics. This technique filters out most topic-irrelevant units. Preliminary investigation with lower thresholds showed that some potential holders may occur in many different topics (e.g. *President Bush*) so a higher threshold would significantly reduce coverage.

The next step is to find only those sentences that are subjective. The easiest way to do this is to use a lexical subjectivity marker (e.g. the word *said* in English). We experimented with automatically finding such markers (usually they are words that introduce indirect speech), but although we had some success, the system turned out to be very complex and not particularly reliable, while making a list of such words (and extending it) is a very

[2]This is a purely empirical assumption. A better version of it could be defining a topic by 'holder – target' pairs, but it would be too restrictive for the relatively small corpus we used.

trivial task even for a person who does not know the language well.

Having a list of topic-relevant lexical units and a set of sentences that have been identified as subjective, we then find out which topic-relevant lexical items in these sentences are opinion holders and which are targets. To do this we use a 'subject marker', a word that denotes a subject in a sentence. This marker is language-dependent and for English and Chinese it is the same as a subjectivity marker, but for Japanese it is not. The relative position of a holder (subject) and a marker (predicate) is also a language-dependent feature which we use for finding holders.

After opinion holders are found, we exclude these lexical items from the list of found topic-related lexical items and use the remainder of them to find opinion targets in the sentences. We make the assumption here that documents (news items) should be consistent on what a holder and a target are.

Having found the lists of opinion holders and opinion targets, it is likely that there are other subjective sentences that were not found with the subjectivity marker, so we use the newly found holders and targets as a further set of subjectivity markers. Thus all sentences that contain any of these words are assumed to be subjective, and opinion holders and targets are extracted from all of them. If a sentence contains a target, but a holder was not found, then the holder is tagged as 'AUTHOR'.

4.2 Language-specific Adjustment

The system described above cannot be used without any adjustment to the language being processed. First of all, to find noun phrases that could be holders or targets, we need to have well-formed lexical units, which implies finding word delimiters (such as *space* in English). This can be done automatically by counting the relative number of space symbols in the document collection: for English documents the number of space symbols will be very high, whereas it will be close to zero in Chinese and Japanese. Once we have such a delimiter we can form proper lexical items for English: meaningless sequences like *prose, rosec, cutor* and alike are eliminated, but a valid *prosecutor* is preserved as it occurs with delimiters (space or punctuation) at both sides. This task is more difficult for the Chinese and Japanese languages (it may require trimming out function words that 'stick' to the lexical items). But for further processing it is more important to find if there is such a delimiter as space to avoid malformed phrases in English (or any other languages that where words are separated by space).

Another piece of language-specific information is the minimal lexical item length. This is not a particularly important parameter, but since we do not want our system to waste time filtering out 1-letter 'word-candidates' from an English list of lexical units, we set the minimal word length to 4 letters. This variable was set to 2 for Chinese, and 3 for Japanese[3].

[3]These values are empirical trade-offs between the average length of words in a language and the number of candidate lexical items that could potentially be extracted.

As outlined above, we need a list of subjectivity markers to find subjective sentences. We use the word *said* for English, the unit 说 (*say, says, said*) for Chinese, and for Japanese we use と言う, という, 言, 話, and 話し (which are equivalents of the English *said*). We use only one word for English and Chinese because in preliminary experiments we found that adding synonyms did not seem to improve performance for either of these languages: the synonyms are too infrequent, as are modal verbs. But since we do not know Japanese, we could not decide which of the words is the most important and left all of them in the list as they were found in an electronic dictionary.

Once subjective sentences are found, we need to find an opinion holder which is assumed to be the subject of a sentence. Fortunately, the subjectivity markers for English and Chinese are verbs, and verbs in these languages are usually quite close to nouns denoting subjects. This enables us to reuse these words as subject markers. To find the opinion holder we find the lexical item closest to the marker. We also consider the relative position of a holder: in English, the subject denoting the speaker can usually be found before the verb (as in *John said ...*), but the inverted construction (*..., said John*) can also be found in some genres. In Chinese, the corresponding verb-noun construction is almost impossible, so we had to adjust the extraction rule accordingly:

布什说，政府可能还会采取更广泛的振兴经济措施。

(*(President) Bush said that ...*)

The Japanese language is quite different from the two mentioned above in its syntactic structure: it is a SOV (subject-object-verb) language. This means that the Japanese marker (equivalent of *said*) cannot be near a holder (which is assumed to be a subject). But there is a special function word in Japanese (は– *wa*) that denotes a topic of a sentence which in conjunction with equivalents of *said* may often be an opinion holder. So we used a simple rule to find a holder near and before this marker:

長崎大の谷川教授は支配層がコントロール能力を失えば「最悪の場合、スリランカのような内乱状態にならない保証はない」と話す。

(*Prof. Tanikawa from Nagasaki (University) said that ...*)

4.3 System Summary

To summarise, the system performs the following steps:

1. Find lexical items.
2. Filter out noisy (not topic-relevant) lexical items
3. Find all subjective sentences.
4. Find opinion holder near subject marker.
5. Find opinion targets.
6. Extract all found holders and targets from all sentences.

Language-specific information that is required is:

1. Word delimiter (can be found automatically)
2. Word-length (not critical, mostly for better performance)
3. Subjectivity marker (word *said* and its equivalents, they also can be found (semi-) automatically)

4. Subject marker (same as p. 3 for English and Chinese, function word *wa* for Japanese)[4]
5. The relative position of a subject (usually before the marker in English, and always before in Chinese and Japanese).
6. As can be seen from this summary, our approach requires little language-specific information.

5. Experiments

5.1 Gold Standard

For holder and target extraction experiments we used the NTCIR-7 MOAT test data collections: English, Simplified Chinese and Japanese. The Simplified Chinese data as supplied by the task organisers had been annotated by twelve annotators, and all topics were annotated by three of them. The English data was annotated using a pool of six annotators. The same approach was taken for Japanese data annotation. The gold standard authors provided two versions of the data: strict and lenient. The strict gold standard contains only those opinion holders and opinion targets that all annotators agreed on. The lenient version has all variants of holders/targets that the annotators came up with (Seki et al, 2008). In this paper we report only lenient evaluation results; strict results follow a very similar pattern.

5.2 Approximate Matches

For each test we used the standard NTCIR-7 MOAT evaluation metrics, consisting of precision, recall and F-measure (F1). In every test we calculated a number of correct matches, when a string (holder or target) extracted by the system exactly matches the one stored in the gold standard file. But since it is not always possible even for a human annotator to establish exact boundaries of a string expressing target or holder, the evaluation script additionally counts all approximate matches. There are three kinds of such matches: **superstring**, **substring** and **overlap**. A **superstring** is a string which is longer than the gold standard string and incorporates the latter entirely, for example:

Gold standard: *"don rodbell"*
System proposed: *"mr don rodbell"*

A **substring** is a shorter string that exactly matches a part of a gold standard string:

Gold standard: *"former nuremberg prosecutor said"*
System proposed: *"former nuremberg prosecutor"*

An **overlap** of two strings is a substring that is present in both strings, but is not an exact match of either:

Gold standard: *"igor ivanov"*
System proposed: *"mr ivanov"*

The approximate matches described above may produce a lot of noise, matching for example short function words or phrases with a long string from the gold standard that also contains such words. To avoid this and to reduce the number of false positives we set a limit of how different in length matching strings can be. For superstring and substring the shorter one should be at least half of the length of the longer one. For overlapping strings, the length of the shared part should at least 1/3 of the combined length of the two strings. For example: for overlapping strings ABCD and BCDY the overlapping part should be at least 2.6 characters long: (ABCD.length + BCDY.length) / 3 = (4 + 4) / 3 = 2.6, so since BCD.length = 4, ABCD and BCDY is a valid approximate match.

Manual inspection of the approximate matches indicated that the vast majority of approximate match strings are valid opinion targets or opinion holders.

[4]This is a language dependent information: for some languages (Slavic, Turkic) it could be morphological units, rather than lexical ones.

5.3 Results

We obtained the results summarised in Table 1, for holder and target identification in each of the three languages, English, Simplified Chinese and Japanese. Figures in brackets are results for approximate matches, which we argue above are reliable indicators of system performance.

The low performance is rather typical for the task (see 5.4) even for supervised monolingual systems. Nonetheless our approach may form the basis for applications in web-based information retrieval where results can be aggregated and ranked.

	Lang	P	R	F1
holder	Eng	0.19 (0.28)	0.09 (0.13)	0.12 (0.18)
holder	Ch	0.18 (0.24)	0.17 (0.22)	0.17 (0.23)
holder	Jap	0.16 (0.16)	0.56 (0.56)	0.25 (0.25)
target	Eng	0.02 (0.16)	0.01(0.06)	0.01 (0.09)
target	Ch	0.03(0.08)	0.03 (0.07)	0.03 (0.07)
target	Jap	0.03 (0.08)	0.10 (0.25)	0.05 (0.13)

Table 1: Performance on the NTCIR-7 MOAT test sets.

5.4 Comparison

These results are numerically fairly low, but opinion holder and target extraction are very difficult tasks. The results compare reasonably well to those reported by the participants of the NTCIR-7 MOAT workshop, but in general are not the best. This can be expected since all of those systems were supervised, and also monolingual.

Specifically, there were 12 systems entered in the MOAT Chinese opinion holder extraction section. Our system would have ranked 9th in terms of F1 (and 7th with respect to approximate match): the best system's F-measure was 0.46, the worst was 0.02, the macro-average for all systems was 0.19. In contrast, for target extraction, our system would have been 2nd (1st) out of five submissions.

Only two systems extracted opinion holders in the English side of NTCIR-7 and our system would not have outperformed either of them. We attribute this to the difference in the evaluation approaches: at NTCIR the English results were evaluated in a semi-automatic mode where if an automatic fuzzy match did not find any matching string, a human judge decided whether a string was an acceptable match. Obviously the automatic evaluation cannot be as flexible and intelligent as a

human judge, so a lot of potentially good output from our system was tagged as incorrect by the evaluation script.

Unfortunately there were no submissions of opinion holder and target extraction systems for Japanese at NTCIR-7, which makes it impossible to compare our system with any others. But since our results are in line with those for the other languages we assume that our results for Japanese are reasonable. It should be noted that most of the holders in the Japanese collection were tagged as 'AUTHOR', resulting in high recall, which might reflect the usual (impersonal) way of expressing opinions in the Japanese language.

5.5 Error Analysis

There are two types of errors: 1) a holder or a target are not present in a sentence in the gold standard, but the system "finds" them and 2) a holder or a target are present in the gold standard but the system proposes wrong strings as a holder or a target. The most of such errors are the result of the system finding too many candidate strings, many of which consist of functional words: *but that cannot* (a system proposed holder). Such errors could easily be eliminated by a list of stop-words applied to the candidate strings. A lot of mistakes were caused by lack of anaphora resolution, which led to too frequent use of pronouns as opinion holders (which was usually considered to be a mistake[5]). One of the most widespread errors for target extraction was an inability to find correct boundaries of a target phrase (see 6). In preliminary experiments we used the whole target subsentence (the remaining part of the sentence after an extracted holder) as a target. This approach produced much more appropriate and legible target strings (although too long sometimes), but such strings were too long compared to the correct targets.

6. Discussion and Future Work

From manual inspections of data, opinion holders seem to have a simpler structure compared to targets. This makes target extraction much more difficult. The complex structure of opinion targets also means that it is possible for different notions of 'target' to exist. Indeed, it is arguable which of the following variants of the same target is the most appropriate: *Russia and China* or *Non-status quo powers* or *Non-status quo powers, most notably Russia and China?* Should we incorporate all or any (which?) attributes into the target? Or should annotators tag only the shortest noun phrase without any attributes? This decision might explain why results for target extraction are so low. The complex structure of opinion targets makes consistent tagging difficult: for example, the English gold standard turned out to be less consistent, as in some cases annotators tagged only noun phrases as targets, but also rather frequently tagged long substrings as targets, for example :

humanitarian intervention (along with cases of self-defense) has been made an exception from the general condemnation on the use of force when interfering in the domestic affairs of another state.

Such long strings tagged as opinion targets are difficult to extract using only topic words. The Chinese corpus annotators were more consistent, mostly tagging only shortest noun phrases, thus the results (especially for approximate matching) are twice as high compared to English.

It is quite obvious that in principle it would be difficult for a knowledge-poor unsupervised approach to outperform the best supervised (or knowledge-based) systems. But judging from our experiments presented in this paper it is possible to conclude that a system which needs only very basic language-specific adjustments (minimal language description) may perform reasonably well. We noted in the previous section that we were comparing a cross-lingual unsupervised system to monolingual supervised systems. A definitive study would involve comparison to supervised systems on a cross-lingual task.

References

Bender, E. M. (2009) *Linguistically Naïve != Language Independent: Why NLP Needs Linguistic Typology. In Proceedings of EACL.* Athens, Greece (pp 26 – 32).

Bloom, K., Garg, N., & Argamon, S. (2006). *Extracting Appraisal Expressions. In Proc. of the Human Language Technology Conf. of the HLT-NAACL 2007* (pp. 308–315). Rochester, New York, USA.

Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). *Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In Proceedings of HLT/EMNLP* (pp. 355-362). Vancouver, Canada.

Kim, J., Jung, H., Nam, S., Lee, Y., & Lee, J. (2008). *English Opinion Analysis for NTCIR7 at POSTECH. In Proc. of the 7th NTCIR Workshop Meeting, Question Answering and Cross-Lingual Access*. Tokyo, Japan.

Kim, S., & Hovy, E. H. (2006). *Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In Proceedings of the COLING/ACL Workshop on Sentiment and Subjectivity in Text.* Sydney, Australia.

Kim, Y., Kim, S., & Myaeng, S. (2008). *Extracting Topic-related Opinions and their Targets in NTCIR-7. In Proc. of the 7th NTCIR Workshop Meeting*. Tokyo, Japan.

Seki, Y. (2008). *A Multilingual Polarity Classification Method using Multi-label Classification Technique Based on Corpus Analysis. In Proc. of the 7th NTCIR Workshop Meeting.* Tokyo, Japan.

Seki, Y., Evans, D. K., Ku, L., Sun, L., Chen, H., & Kando, N. (2008). *Overview of Multilingual Opinion Analysis Task at NTCIR-7. In Proc. of the 7th NTCIR Workshop Meeting.* Tokyo, Japan.

Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text. *Advances in Intelligent Data Analysis*, *VI*, 121 - 132.