Opinion Expression Mining by Exploiting Keyphrase Extraction

Gábor Berend
Department of Informatics,
University of Szeged
2. Árpád tér, H-6720, Szeged, Hungary
berendg@inf.u-szeged.hu

Abstract

In this paper, we shall introduce a system for extracting the keyphrases for the reason of authors' opinion from product reviews. The datasets for two fairly different product review domains related to movies and mobile phones were constructed semi-automatically based on the pros and cons entered by the authors. The system illustrates that the classic supervised keyphrase extraction approach – mostly used for scientific genre previously – could be adapted for opinion-related keyphrases. Besides adapting the original framework to this special task through defining novel, task-specific features, an efficient way of representing keyphrase candidates will be demonstrated as well. The paper also provides a comparison of the effectiveness of the standard keyphrase extraction features and that of the system designed for the special task of opinion expression mining.

1 Introduction

The amount of community-generated contents on the Web has been steadily growing and most of the end-user contents (e.g. blogs and customer reviews) are likely to deal with the author's emotions and opinions towards some subject. The automatic analysis of such material is useful for both companies and consumers. Companies can easily get an overview of what people think of their products and services and what their most important strengths and weaknesses are while users can have access to information from the Web before purchasing some product.

In this paper we will introduce a system which assigns pro and con keyphrases (free-text annotation) to product reviews. When dealing with product reviews, our definition of keyphrases is the set of phrases that make the opinion-holder feel negative or positive towards a given product, i.e. they should be the reason why the author likes or dislikes the product in question (e.g. *cheap price*, *convenient user interface*). Here, we adapted the general keyphrase extraction procedure from the scientific publications domain (Witten et al., 1999; Turney, 2003) to the extraction of opinion-reasoning features. However, our task is rather different since we aim at identifying the reasons for opinions, instead of keyphrases that represent the content of the whole document.

The supervised keyphrase extractor to be introduced here was trained on the pros and cons assigned to the reviews by their authors on the `epinions.com` site. These pros and cons are ill-structured free-text annotations and their length, depth and style are extremely heterogeneous. In order to have clean gold-standard corpora, we manually revised the segmentation and the contents of the pros and cons, and obtained sets of tag-like keyphrases.

2 Related work

There have been many studies on opinion mining (Turney, 2002; Pang et al., 2002; Titov and McDonald, 2008; Liu and Seneff, 2009). Our approach relates to previous work on the extraction of reasons for opinions. Most of these papers treat the task of mining reasons from product reviews as one of identifying sentences that express the author's negative or positive feelings (Hu and Liu, 2004a; Popescu and Etzioni, 2005). This paper is clearly distinguishable from them as our goal is to find the reasons for opinions expressed by phrases and we aim the task of phrase extraction instead of sentence recognition.

This work differs in important aspects even from the frequent pattern mining-based approach of (Hu and Liu, 2004b) since they regarded the main task of mining opinion features with respect

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1162–1170,
Chiang Mai, Thailand, November 8 – 13, 2011.

to a group of products, not individually at review-level as we did. Even if an opinion feature phrase is feasible for a given product-type, it is not necessary that all of its occurrence are accompanied with sentiments expressed towards it (e.g. *The phone comes in red and black colors*, where *color* could be an appropriate product feature, but not an opinion-forming phrase).

A similar task to pro and con extraction gathers the key aspects from document sets, which has also gained interest recently (Sullivan, 2008; Branavan et al., 2008; Liu and Seneff, 2009). Existing aspect extraction systems first identify a number of aspects throughout the whole review set, then they automatically assign items from this pre-recognized set of aspects to each unseen review. Hence, they work at the corpus level and restrict themselves to using only a pre-defined number of aspects.

The approach presented here differs from these studies in the sense that it looks for the reason phrases themselves review by review, instead of multi-labeling some aspects. These approaches are intended for applications used by companies who would like to obtain a general overview about a product or would like to monitor the polarity relating to their products in a particular community. In contrast, we introduce here a keyphrase extraction-based approach which works at the document level as it extracts keyphrases from reviews which are handled independently of each other. This approach is more appropriate for the consumers, who would like to be informed before purchasing some product.

The work of Kim and Hovy (2006) lies probably the closest to our one. They addressed the task of extracting con and pro sentences, i.e. the sentences on why the reviewers liked or disliked the product. They also note that such pro and con expressions can differ from positive and negative opinion expressions as factual sentences can also be reason sentences (e.g. *Video drains battery.*). Here the difference is that they extracted sentences, but we targeted phrase extraction.

Most of the keyphrase extraction approaches (Witten et al., 1999; Turney, 2003; Medelyan et al., 2009; Kim et al., 2010) work on the scientific domain and extract phrases from one document that are the most characteristic of its content. In these supervised approaches keyphrase extraction is regarded as a classification task, in which certain n-grams of a specific document function as keyphrase candidates, and the task is to classify them as proper or improper keyphrases. Here, our task formalization of keyphrase extraction is adapted from this line of research for opinion mining and we focus on the extraction of phrases from product reviews that also bear subjectivity and induce sentiments in its author. As community generated pros and cons can provide abundant training samples and our goal is to extract the users' own words, here we also follow this supervised keyphrase extraction procedure.

3 Opinion Phrase Extraction Framework

Here, we employed a supervised machine learning approach for the extraction of reason keyphrases from a given review. Candidate terms were extracted from the text of the review and those present in the extracted set of pros and cons were regarded as positive examples during training and evaluation. Maximum Entropy classifiers were trained and the keyphrase candidates with the highest posteriori probabilities were selected to be keyphrases for a review of a test document in question. In the following subsections we will describe how keyphrase candidates and the feature space representing them were constructed.

3.1 Candidate term generation

One key aspect in keyphrase extraction is the way keyphrase candidates are selected and represented. As usually the number of potentially extracted n-grams and that of genuine keyphrases among them show high imbalancedness, keyphrase candidates are worth to be filtered, instead of using any successive n-grams. For this reason we limited the maximal length of the extracted phrases to at most 4 tokens and also required that the phrases should begin with either a non-stopword adjective, verb or noun and should end to either a non-stopword noun or adjective.

As for the filtration of the candidate set, a new step is introduced here, which omits normalized phrases that had only such occurrences which contained stopwords. This simple step proved effective in excluding many non-proper opinion phrases (i.e. increasing the maximal precision achievable) at the cost of discarding only a small proportion of proper phrases (i.e. slightly decreasing the best recall achievable).

Once we had the keyphrase candidates, they had

to be brought to a normalized form. The normalization of an n-gram consisted of lowercasing and Porter-stemming each of the lemmatized forms of its tokens, then putting these stems into alphabetical order (while omitting the stems of stopword tokens). With this kind of representation it was then possible to handle two orthographically different, but semantically equivalent phrases, such as '*the screen is tiny*' and '*TINY screen*' in the same way.

Previous works on keyphrase extraction also usually carry out this step of normalization, however, here we did it in such a manner that a mapping to each of the original orthographic forms of a normalized form and its corresponding context (i.e. the sentences containing it) was preserved at the same time and that could be successfully utilized at later processing steps.

To provide an alternative way of normalizing phrases, experiments relying on the usage of WordNet (Fellbaum, 1998) were also conducted. In these settings the normalized form of a single token was determined by first searching for all its synsets (in the case of verbs, these were such noun synsets that were in derivative relation with the synsets of the verb word form). Then instead of Porter-stemming the original token, its most frequent word form was stemmed, based on the estimated frequencies of WordNet for all the word forms of the synsets of the original token. In this way two – originally differently stemmed – word forms, such as *decide* and *decision* could be stemmed to the same root forms. Another advantage of this procedure is that it is able to handle semantic similarity to some extent.

The remaining parts of the normalization procedure were left unchanged (i.e. lowercasing and alphabetical ordering of the normalized forms of the individual tokens). Later, in the Results section, the effect of this kind of normalization will be shown.

Candidate terms were handled at the review level instead of occurrence level. This means that each normalized occurrence of a keyphrase candidate was gathered from the document and the feature values for the candidate term aggregate over its occurrences.

3.2 Feature representation

We constructed a rich feature set to represent the review-level keyphrase candidates. The feature space incorporates features calculated on the basis of the normalized phrases themselves, but more importantly, thanks to the mapping between the normalized phrase forms and their original occurrences, new contextual and orthographic features were possible to incorporate.

Features that could be generally used for any kind of keyphrase extracting task (e.g. that makes use of multiword expressions or character suffixes in a special way) and ones designed especially for the novel task of opinion phrase extraction (e.g. that uses SentiWordNet to determine polarity) as well as the standard features of keyphrase extraction are both introduced in the following.

Standard Features Since we assumed that the underlying principles of extracting opinionated phrases are quite similar to that of extracting standard (most of the time scientific) keyphrases, features of the standard setting were applied in this task as well. The most common ones, introduced by KEA (Witten et al., 1999) are the **Tf-idf** value and the **relative position** of the first occurrence of a candidate phrase within a document. We should note that KEA is primarily designed for keyphrase extraction from scientific publications and whereas the position of the first occurrence might be indicative in research papers, product reviews usually do not contain a summarizing "abstract" at the beginning. For these reasons we chose these features as the ones which form our baseline system. **Phrase length** is also a common feature, which was defined here as the number of the non-stopword tokens of an opinion phrase candidate.

Linguistic and orthographic features Since certain POS-codes are more frequent than others among genuine keyphrases, features generated by POS-codes belonging to an occurrence of a normalized phrase were applied. As **POS-code** sequences seem to be more informative, instead of simply indicating which POS-codes were assigned to any orthographic alternation of a normalized keyphrase candidate, it would be desirable to store the POS-code sequences in their full length as well. However, doing so might affect dimensionality in a negative way (especially when having few training data), i.e. the number of all the possible POS-code sequences ranging from lengths of 1 to 4 is too much. To overcome this issue, positional information was added to the POS-code features derived from the tokens of an n-gram. Features

of POS-codes that were assigned to a token being itself a 1-token long keyphrase candidate, at the beginning, at the end, in between an n-gram, got a prefix `S-`, `B-`, `E-` and `I-`, respectively. For instance, the phrase *cheap/JJ phone/NN* induces the features {*B-JJ*, *E-NN*}, whereas the 1-token-long phrase *cheap/JJ* induces the feature {*S-JJ*}. Finally, numeric values for a normalized candidate phrase were assigned based on the distribution of the different POS-related features of all the running-text forms of a normalized phrase.

We introduced features exploiting the syntactic context of a candidate with parse trees. For an n-gram with respect to all the sentences it was contained in a given document, this feature stored the average and the minimal depths of those **NP-rooted trees** that contained the whole n-gram in its yield. These features are intended to express the "noun phraseness" of the phrase.

Features generated from the **character suffixes** of the individual tokens of the occurrences of a normalized keyphrase candidate were also employed. Character suffix features also incorporated positional information, similarly as it was done in the case of POS features. The suffixes themselves came from the last 2 and 3 characters of the tokens constructing an n-gram. For instance, the features induced by (and thus assigned with true value) for the phrase *ch**eap** ph**one*** are {*B-eap*, *B-ap*, *E-one*, *E-ne*}.

Opinionated phrases often bear special orthographic characteristics, e.g. in the case of *so sl**ooo**w* or *CHEAP*. Due to the fact that the original forms of the phrases are stored in our representation, it was possible to construct two features for this phenomenon: the first feature is responsible for **character run**s (i.e. more than 2 of the same consecutive characters), and an other is responsible for **strange capitalization** (i.e. the presence of uppercase characters besides the initial one). The S-,B-,E-,I- prefixes were applied here as well, just like in the case of the **Named Entity** feature, which represented if a token was part of NE (with its type as well).

World knowledge-based features Features relying on the outer resources of Wikipedia and SentiWordNet were also exploited during our experiments. They were useful as world knowledge could be incorporated by their means.

Multiword expressions are lexical items that can be decomposed into single words and display idiosyncratic features (Sag et al., 2002), in other words, they are lexical items that contain space.

To measure the added value of MWEs in the task of opinion phrase extraction, a set of features was designed that indicated whether a certain phrase candidate (1) is an MWE on its own (e.g. *ease of use*), (2) can be composed from more MWEs on the list (e.g. *mobile internet access*), or is just the (3) superstring of at least one MWE from the list (e.g. *send text messages*). In order to be able to make such decisions, a wide list of MWEs was constructed from Wikipedia (dump 2011-01-07): all the links and formatted (i.e. bold or italic) text were gathered that were at least two tokens in length, started with lowercase letters and contained only English characters or some punctuation. Finally, an alignment of the elements of the list and the contexts of the reviews of the dataset was carried out (taking care of linguistic alternations and POS-tag matchings).

A more sophisticated surface-based feature used external information as well on the individual tokens of a phrase. It relied on the **sentiment scores** of SentiWordNet (Esuli et al., 2010), a publicly available database that contains a subset of the synsets of the Princeton Wordnet with positivity, negativity and neutrality scores assigned to each one, depending on the use of its sentiment orientation (which can be regarded as the probability of a phrase belonging to a synset being mentioned in a positive, negative or neutral context). These scores were utilized for the calculation of the sentiment orientations of each token of a keyphrase candidate. Surface-based SentiWordnet-calculated feature values for a keyphrase candidate included the *maximal positivity and negativity and subjectivity* scores of the individual tokens and the *total sum* over all the tokens of one phrase.

Sentence-based features were also defined based on SentiWordNet as it was also used to check for the presence of **indicator terms** within the sentences containing a candidate phrase. Those word forms were gathered from SentiWordNet, for which the sum of the average positivity and negativity sentiments scores among all its synsets were above 0.5 (i.e. the ones that are more likely to have some kind of polarity). Then for a given keyphrase candidate of a given document, a true value was assigned to the SentiWordNet-derived indicator features that had at least one

co-occurrence within the same sentence with the keyphrase candidate in the same document.

SentiWordnet was also used to investigate the entire sentences that contained a phrase candidate. This kind of feature calculated the sum of every sentiment score in each sentence where a given phrase candidate was present. Then the mean and the deviation of the sum of the sentiment scores were calculated for each token of the phrase-containing sentences and assigned to the phrase candidate. The mean of the sentiment scores of the individual sentences yielded a general score on the **sentiment orientation** of the sentences containing a candidate phrase, while higher values for the **deviation** was intended to capture cases when a reviewer writes both factual (i.e. uses few opinionated words) and non-factual (i.e. uses more emotional phrases and opinions) sentences about a product.

Finally, Wikipedia was also used to incorporate semantic features from its category hierarchy. (Wikipedia categories form a taxonomy, indicating which article belongs to which (sub)category). In the case of a candidate phrase all the nominal parts of the normalized titles of **Wikipedia categories** for its related Wikipedia articles were added as separate binary features to the feature space. The normalization of the Wikipedia category names was similar to that of keyphrase candidates. For instance, given the candidate phrase '*service quality*' the feature *wiki_control_qual* is set to true since the Wikipedia article named *Service quality* is in the category *Quality control*.

Document and corpus-level features Among document-level features, the **standard deviation of the relative positions** compared to the document length was a measure to be computed. Higher values of the deviation in the position means that the reviewer keeps repeating some phrase from the beginning to the end of the review, which might indicate that this phrase is of higher importance for them.

As verbs often contribute to the sentiment polarity of the noun phrases they accompany (e.g. '*I adore its fancy screen.*' versus '*I bought this phone one year ago.*'), a set of features was introduced to deal with the **indicative verbs** in the context of candidate phrase occurrences within their document. For this feature to be calculated we took those verbs as indicators that occurred at least 100 times in the whole training dataset. When calculating a feature value for an opinionated-phrase candidate, the algorithm matched all of its occurrences in a document against every indicator verb. For the calculation of the feature value for a given phrase candidate – indicator verb pair, a syntactic distance value was first defined. This syntactic distance was equal to the minimal height of the subtree which contained both the keyphrase candidate and the indicator verb itself to the left among all the sentences associated with a document that contained the keyphrase candidate. The feature value was then determined by simply taking the reciprocal of this semantic distance. This way, the feature value was scaled between 0 and 1. (Note that for indicator verbs that were not present in any of the sentences containing a phrase candidate associated with a document, the semantic distance value was defined to be infinity, the limit value of the reciprocal of which is 0.)

Quite general characteristics of reason-expressing phrases can also be captured at the corpus level. Simply using the number of times an argument phrase aspirant was assigned to a review as a proper phrase on the training dataset was also taken into account as a **corpus-level** feature since the same proper opinion phrases can easily reoccur regarding products of the same type.

4 Experiments

Experiments were carried out on two fairly different types of product reviews, namely mobile phones and movies. We use standard keyphrase extraction evaluation metrics and baselines for evaluating our pros and cons extractor system.

4.1 Datasets

In our experiments, we crawled two quite different domains of product reviews, i.e. mobile phone and movie reviews from the review portal `epinions.com`. For both domains, 2000 reviews were crawled from `epinions.com` and an additional of 50 and 75 reviews for measuring inter-annotator agreement, respectively. This corpus is quite noisy (similarly to other user-generated contents); run-on sentences and improper punctuation were common, as well as grammatically incorrect sentences since reviews were often written by non-native English speakers. [1]

[1] All the data used in our experiments are available at `http://rgai.inf.u-szeged.hu/proCon`

	Mobiles	Movies
Number of reviews	2009	1962
Avg. sentence/review	31.9	29.8
Avg. tokens/sentence	16.1	17.0
Avg. keyphrases/review	4.7	3.2
Avg. keyphrase candidates/review	130.38	135.89

Table 1: Size-related statistics of the corpora

The list of pros and cons was inconsistent too in the sense that some reviewers used full sentences to express their opinions, while usually a few token-long phrases were given by others. The segmentation of their elements was marked in various ways among reviews (e.g. comma, semicolon, ampersand or the *and* token) and even differed sometimes within the very same review. There were many general or uninformative pros and cons (like *none* or *everything* as a pro phrase) as well.

In order to have a consistent gold-standard annotation for training and evaluation, we manually refined the pros and cons of the reviews in the corpora. In the first step, the automatic preprocessing of the segmentation of pros and cons was checked by human annotators. Our automatic segmentation method split the lines containing pros and cons along the most frequent separators. This segmentation was corrected by the annotators in 7.5% of the reviews. Then the human annotators also marked the general pros and cons (11.1% of the pro and con phrases) and the reviews without any identified keyphrases were discarded.

4.2 Evaluation issues

Keyphrase extraction systems are traditionally evaluated on the top-n ranked keyphrase candidates for each document by F-score (Kim et al., 2010), which combines the precision and recall of the correct keyphrases' class. Evaluation is carried out in a strict manner as a top-ranked keyphrase candidate is accepted if it has exactly the same standardized form as one of the keyphrases assigned to the review. The ranking of the phrase candidates was based on a probability estimation of a candidate belonging to the positive keyphrase class. Results reported here were obtained using 5-fold cross validation using Maximum Entropy classifier.

As we treated the mining of pros and cons as a supervised keyphrase extraction task, we conducted measurements with KEA (Witten et al., 1999), which is one of the most cited publicly available automatic keyphrase extraction system. However, we should note that due to the fact that our phrase extraction and representation strategy (and even the determination of true positive instances to some extent) slightly differs from that of KEA, the added values of our features should rather be compared to our second Baseline System (BL_{WN}) which uses WordNet for candidate phrase normalization. The baseline systems use our framework, with the feature set of KEA, which consists of tf-idf feature and the relative first occurrence of a keyphrase candidate. The only difference among the two baseline systems is that BL does not apply the WordNet-based normalization of phrase candidates introduced in Section 3.1.

Since we had the same findings as Branavan et al. (2008) that authors often omit several opinion forming aspects from their pros and cons listings that they later include in their review, we decided to determine the complete lists of pros and cons manually, that is, to compose pro and con phrases on the basis of the reviews. Due to the highly subjective nature of sentiments, the determination of sentiment-affecting pro and con phrases was carried out by three linguists, who were asked to annotate a 25-document subset of the mobile phone dataset. Their averaged agreements for the determination of pro phrases are 0.701 and 0.533 for Dice's coefficient and Jaccard index, and 0.69 and 0.526 for cons, respectively.

4.3 Results

In our experiments all the linguistic processing of the product reviews were carried out using Stanford CoreNLP. It uses the Maximum Entropy POS-tagger of Toutanova and Manning (2000) and syntactic parsing works on the basis of Klein and Manning (2003). The ranking of the candidate keyphrases was based on the posteriori probabilities of the MALLET implementation (McCallum, 2002) of Maximum Entropy classifier (le Cessie and van Houwelingen, 1992).

During the fully automatic evaluation, we followed strict evaluation (see 4.2) that is commonly utilized in scientific keyphrase extraction tasks. Table 2 contains the results of the strict evaluation for both domains. However, since strict evaluation is more likely to suit the evaluation of scientific keyphrase extraction better, i.e. semantically equivalent but different word forms are less common at that domain, we conducted human evaluation on the 25-document subset of the mobile

Feature	Mobiles Top-5	Mobiles Top-10	Mobiles Top-15	Movies Top-5	Movies Top-10	Movies Top-15
KEA	1.72/1.84/1.77	1.42/3.04/1.94	1.39/4.48/2.12	1.21/1.93/1.49	0.98/3.13/1.5	0.89/4.26/1.48
BL	2.6/2.8/2.73	2.6/5.5/3.54	2.6/8.2/3.93	1.6/2.5/1.95	1.5/4.9/2.34	1.6/7.4/2.58
BL_{WN}	2.7/2.9/2.8	2.7/5.8/3.68	2.7/8.7/4.12	1.7/2.8/2.14	1.7/5.4/2.61	1.7/8.2/2.88
IV	3.1/3.4/3.25§	2.9/6.2/3.92	2.8/9.1/4.31	2.4/3.7/2.9†	2.0/6.3/3.04§	1.9/8.8/3.09
KF	2.6/2.8/2.71	2.7/5.9/3.73	2.7/8.7/4.11	1.7/2.7/2.09	1.7/5.4/2.59	1.7/8.2/2.87
$Length$	3.2/3.4/3.26 §	3.1/6.6/4.18†	2.9/9.3/4.4	2.1/3.3/2.6	2.0/6.4/3.08§	2.0/9.1/3.22§
MWE	4.7/5.0/4.88‡	3.8/8.0/5.11‡	3.4/10.8/5.12‡	2.3/3.6/2.81†	2.0/6.3/3.06†	1.9/9.1/3.18§
POS	4.6/4.9/4.71‡	4.2/9.0/5.77‡	3.9/12.6/5.98‡	2.9/4.6/3.57‡	2.8/8.7/4.18‡	2.5/11.7/4.1‡
SWN	6.0/6.4/6.2‡	4.9/10.4/6.65‡	4.3/13.6/6.49‡	3.7/6.0/4.6‡	3.1/9.8/4.73‡	2.8/13.1/4.59‡
$StDev$	3.9/4.2/4.06‡	3.8/8.1/5.15‡	3.5/11.2/5.33‡	2.9/4.6/3.59‡	2.6/8.1/3.9‡	2.5/11.6/4.07‡
$Orth.$	3.2/3.4/3.28§	3.1/6.7/4.27†	2.9/9.5/4.49	3.0/4.7/3.65‡	2.5/7.8/3.76‡	2.3/10.9/3.82‡
$Suffix$	11.5/12.2/11.83‡	8.6/18.2/11.66‡	6.9/22.0/10.54‡	6.8/10.7/8.34‡	5.2/16.4/7.91‡	4.3/20.1/7.08‡
$Syntax$	3.5/3.7/3.61‡	3.0/6.4/4.06	2.8/9.1/4.33	2.3/3.6/2.78†	2.0/6.1/2.97§	1.9/9.1/3.2§
$Wiki$	11.9/12.7/12.25‡	8.1/17.4/11.09‡	6.3/20.1/9.63‡	8.8/13.9/10.78‡	6.3/19.8/9.59‡	4.8/22.5/7.9‡
$COMB$	**14.8/15.7/15.27**‡	10.4/22.0/14.11‡	8.0/25.4/12.17‡	**10.0/15.8/12.22**‡	7.0/21.9/10.63‡	5.3/24.6/8.67‡

Table 2: Performance using different features in the form of Precision/Recall/F-score obtained. IV, KF, SWN and Orth. stands for indicator verbs, corpus-level keyphrase frequency, SentiWordNet and orthography-driven features, respectively. Symbols §, † and ‡ in the upper index of a result indicates that it is significantly better compared to the baseline system which uses the WordNet based candidate phrase normalization (BL_{WN}) at confidence levels of 0.1, 0.05 and 0.01, based on Student's t-test, respectively. As it was only the KF feature which did not yield any significant improvement at all, the combined system (COMB) incorporated all the features but KF.

phone domain. The results of the manual evaluation is shown in 3.

4.4 Discussion

The fact that the highest F-scores for keyphrases are achieved when the number of extracted phrases is around the average number of pro and con phrases per reviews (i.e. between 3 and 5) suggests that our ordering of keyphrase candidates is quite effective (since once we find the number of keyphrases a document has, performance cannot really grow anymore).

Comparing the nature of the task of extracting keyphrases from scientific publications and that of product reviews, we shall take two observations: firstly, keyphrases of scientific documents are more universal, i.e. once we have the knowledge that the expression *distributed computing* was a good keyphrase for one scientific document, we can be more confident about it being a proper keyphrase for other documents within the same domain as well, whereas in the case of opinion phrases such as *pink color* can easily be mentioned in either opinionated and non-opinionated contexts. Secondly, besides scientific keyphrases being more ***universal***, they are more ***deterministic*** in the sense that there are fewer ways to express good keyphrases, e.g. suppose *simulated annealing* is a proper keyphrase for a scientific document, it is unlikely that an automatic system would extract *imitated annealing*, whereas in the case of product review the gold standard keyphrases often differ from their mention in the text (e.g. *tiny keys* and *small keys*).

The above mentioned examples suggest that opinion phrase extraction is more difficult to be performed and evaluated compared to scientific keyphrase extraction. We should note that the best performing system at SemEval-2010 (Kim et al., 2010) that dealt with the much simpler task of scientific keyphrase extraction achieved an F-score of 19.3 when evaluated against author keywords at the top-15 level.

It should be also added here, that among the keyphrases regarded as false positives in our evaluations, there were many near misses due to synonymy, e.g. *tiny keys* and *small keys* or *slow Web* and *slow WAP*. To overcome the synonymy issue to some extent the WordNet-based rewriting of tokens was introduced, which brought improvement in the case of the baseline systems for both domains (so it was employed in the later experiments as well). Another source of false positive classifications was due to the incompleteness of the opinion aspect entered by the user, i.e. not all the important aspects are necessarily listed among the

	Top-5			Top-10			Top-15		
	Prec.	Recall	F-score	Prec.	Recall	F-score	Prec.	Recall	F-score
$\cup$	72.8	20.63	32.14	66.8	33.54	44.66	63.47	46.88	53.92
$\cap$	46.4	27.81	34.77	41.6	44.92	43.2	37.07	56.68	44.82
Author	34.4	22.29	27.05	31.6	35.43	33.4	28.8	45.14	35.17

Table 3: Results of the human evaluation. $\cup$, $\cap$ and Author means when the automatic keyphrases were matched against the union, intersection of the keyphrases of three independent annotators and the keyphrases of the original author, respectively.

pros and cons section, as described earlier. On the other side, many of the author-entered keyphrases were absent in the contents of a review in their same form: only 34,8% and 23,9% of gold standard keyphrases could be found in the texts having the same normalized form for the mobile phone and the movie domains, respectively, setting an upper bound for the recall values when evaluating based on strict matching.

To overcome all the previously mentioned shortcoming during automatic evaluation, human evaluation was performed and it showed that real life application of opinion phrase extraction could be of much higher utility than strict evaluation would suggest. This is due to the fact that human annotators had access to common sense knowledge and during the inspection of keyphrases they could resolve such cases that were impossible during automatic evaluation.

All the features were effective in the sense that expanding the baseline feature set by them separately resulted in better results. Moreover, in the majority of the cases improvements were of high significance (see Table 2). The added value of Wikipedia features (that are likely to work well in other domains as well) should be highlighted as well as the relatively poor effect of keyphrase frequency feature which normally works better in the case of standard scientific keyphrase extraction tasks. A possible reason for keyphrase frequency feature not being that effective in the opinion domain is that in the case of opinionated keyphrases, the presence of such a phrase that was marked as positive in one document is not necessarily marked the same way in other documents, e.g. because one author may write about the feature objectively while the other may write his strong opinions about the very same feature, using similar wording.

5 Conclusions

In this paper, we presented a pros and cons extraction system by pointing out the parallelism between the keyphrases of scientific papers – given by their author – and the pros and cons phrases – given by product reviewers. The WordNet-based phrase normalization and an extended stopword-based filtration of keyphrase candidates introduced here could be of possible use for any kind of phrase extraction tasks. Besides demonstrating their similarity, the main differences of the two tasks were also highlighted, and several ways to adopt to the specialties of opinion phrase extraction have been suggested by introducing a rich feature set, some of which could also be widely used (e.g. Wikipedia-based ones), and others are specifically designed to the special task of opinion phrase extraction (e.g. SentiWordNet-related ones).

Among the most important differences of opinion phrase extraction from scientific keyphrase extraction we should note that for product reviews the pure occurrence of a single phrase is less deterministic to be a keyphrase, i.e. some emotional context is necessary to treat them as genuine ones. Also, the language of reviews is more special since it tends to contain elements that are not present in other genres of documents, such as irony and sarcasm and offers more possibility to express identical things in different ways. In total, our results are competitive with those of other standard keyphrase extraction tasks even when applying strict normalized form matching evaluation. Moreover, human evaluation showed that when semantics are involved into the evaluation, results are significantly better than it is suggested by automatic evaluations.

Acknowledgments

This work was supported by the Project "TÁMOP-4.2.1/B-09/1/KONV-2010-0005 – Creating the Center of Excellence at the University of Szeged", supported by the European Union and co-financed by the European Regional Development Fund and by the project BELAMI financed by the National Innovation Office of the Hungarian government.

References

S.R.K. Branavan, Harr Chen, Jacob Eisenstein, and Regina Barzilay. 2008. Learning document-level semantic properties from free-text annotations. In *Proceedings of ACL-08: HLT*, pages 263–271, Columbus, Ohio. ACL.

Andrea Esuli, Stefano Baccianella, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. ELRA.

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London.

Minqing Hu and Bing Liu. 2004a. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Minqing Hu and Bing Liu. 2004b. Mining opinion features in customer reviews. In *Proceedings of the 19th national conference on Artifical intelligence*, AAAI'04, pages 755–760. AAAI Press.

Soo-Min Kim and Eduard Hovy. 2006. Automatic identification of pro and con reasons in online reviews. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 483–490, Sydney, Australia. ACL.

Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 21–26, Morristown, NJ, USA. ACL.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st ACL*, pages 423–430.

S. le Cessie and J.C. van Houwelingen. 1992. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

Jingjing Liu and Stephanie Seneff. 2009. Review sentiment scoring via a parse-and-paraphrase paradigm. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 161–169, Singapore. ACL.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Olena Medelyan, Eibe Frank, and Ian H. Witten. 2009. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1318–1327, Singapore. ACL.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Morristown, NJ, USA. ACL.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 339–346, Vancouver, British Columbia, Canada. ACL.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of CICLing-2002*, pages 1–15, Mexico City, Mexico.

Todd Sullivan. 2008. Pro, con, and affinity tagging of product reviews. Technical Report 224n, Stanford CS.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio. ACL.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, EMNLP '00, pages 63–70, Stroudsburg, PA, USA. ACL.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 417–424.

Peter Turney. 2003. Coherent keyphrase extraction via web mining. In *Proceedings of IJCAI*, pages 434–439.

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, and Craig. 1999. Kea: Practical automatic keyphrase extraction. In *ACM DL*, pages 254–255.