Discovery of Subjective Evaluations of Product Features in Hotel Reviews

Viktor Pekar and Shiyan Ou
HLSS, University of Wolverhampton, United Kingdom
{V.Pekar;Shiyan.Ou}@wlv.ac.uk

1. Introduction

Tourism is a dynamic and growing industry, with the Internet offering a multitude of new ways of conducting tourism business and promoting tourism destinations (World Tourism Organization, 2007). As the Forrester Research[1] reported in 2005, in Europe 40% of users of tourism services book their travels on the Web. However, current information technologies are hardly capable of making full use of the potential of the Web for tourism business. Traditional search engines do not provide users efficient means to access the information they require, retrieving vast numbers of web pages in response to queries expressed in keywords. Instead, users often want specific and brief answers to complex queries like "*Tell me the telephone numbers of Hilton Birmingham Hotel*" or "*What are the major places of interest in the Black Country*". The purpose of the EU-funded QALL-ME[2] project, is to address this need by providing an information facility to that accepts natural language questions from users interested in tourism services and retrieves succinct answers to them from the Web.

The main focus of QALL-ME are factual questions, expected answers to which are often contained in the text of web pages and can be obtained using answer extraction and reformulation techniques. User information requests, however, often have to do not only with factual information, but also involve subjective evaluation of a tourism product, such as "*What do people think of the Hilton Birmingham Hotel?*" and "*Does this hotel have a good restaurant?*". Such kinds of questions are not amenable to question answering techniques, since exact answers to them are not directly available in the text of a specific web page. To answer such questions, Opinion Mining techniques must be employed.

Opinion mining seeks to determine the sentiment, attitude or opinion of an author expressed in texts with respect to a certain topic. On the web, there is an increasing number of review web sites (such as *epinions*[3] and *tripadvisor*[4]), where users post their comments on a product (e.g. hotel and restaurant) and provide their positive or negative evaluation. These websites are important resources providing advice to new users and helping them with their travel plans. On the other hand, customer comments on a product found on such websites can be used for the purposes of marketing research and customer relationship management by tourism businesses. Automatic analysis of sentiment expressed in such customer reviews has a lot of potential for applications in the tourism domain.

In this study, the overall problem we address is the analysis of customer reviews with respect to specific features of a tourism product. Our eventual goal is to generate a feature-based report on a product based on this analysis that would describe the importance of each feature

[1] http://www.forrester.com/
[2] http://qallme.itc.it/
[3] http://www.epinions.com/
[4] http://www.tripadvisor.com

from the customer's point of view, pinpoint the product's strengths and weaknesses and enable detailed comparison of different products with regard to customer preferences. Within the QALL-ME project, feature-based reports on a hotel are meant to supplement factual information that the question answering techniques retrieve from the Web.

The specific problem we address is how to associate descriptions of different product features with sentiment expressions found in a review. We present a method for identification of extraction patterns that relate the two types of expressions. Embedding the method within a system for feature-based report generation, we evaluate it against human-assigned grades to various aspects of hotels, based on customer hotel reviews. Our evaluation demonstrates a high correlation between the grades assigned to the hotel features by our system and the grades assigned by human reviewers.

2. Related work

Previous work has attempted to perform opinion mining at three different levels – the document level, the sentence level and the feature level, which correspond to increasing granularity of automatic opinion analysis. At the document level, whole documents are classified into either "positive" or "negative" according to the overall sentiment expressed in the text. To predict the polarity of the opinion expressed in documents, sentiment words such as "*excellent*", "*poor*", "*enjoy*", and "*dislike*", are used as input into statistical or machine learning classification algorithms (e.g. Turney, 2002; Tong, 2001), or manually labelled documents are used to train a classifier (Pang et al., 2002; Das & Chen, 2001; Gamon, 2004). The document-level sentiment classification is based on the assumption that each document focuses on a single object and contains unique opinion from a single opinion holder. However, the assumption does not always hold and not all sentences in a product review express subjective opinions. Instead, many sentences present factual information. To deal with this fact, the sentence- or clause-level sentiment classification is performed, which consists of two subtasks – distinguishing subjective from objective sentences and determining the polarity (*positive* vs. *negative*) of each subjective sentence. The representative studies on subjectivity sentence classification include Bruce and Wiebe (2000), Hatzivassiloglou and Wiebe (2000), Pang and Lee (2004), and Riloff et al. (2005).

A product review usually contains comments on different aspects or features of a product, e.g. *picture quality* and *battery life* for a camera, or opinions of different subjects on a topic, e.g. *persons* or *organizations*. The document-level and sentence-level sentiment classification can determine the overall sentiment in a document or sentence but is unable to indicate which specific features of an object are *evaluated positively and which negatively*. The third variety of opinion mining techniques is intended to reveal the opinions expressed towards individual features. This problem involves two subtasks – extracting different features of a product and associating each feature with its corresponding opinions. To address the first subproblem, Yi et al. (2003) extracted nouns and noun phrases as candidate feature terms based on patterns of part-of-speech tags and selected feature terms using likelihood-ratio test. Hu and Liu (2004) used sequential pattern mining, a kind of special association mining which considers word order in a sentence, to extract frequent features. For infrequent features that were talked about only by a small number of reviewers, known sentiment words were used to identify the nearby infrequent features which were modified by them on the assumption that the same sentiment word can modify both frequent and infrequent features. To improve Hu and Liu's association mining rule method for frequent features, Popescu and Etzioni (2005) removed

those frequent noun phrases that may not be product features by computing the PMI scores between each phrase and part discriminators associated with the product class (e.g., "*of scanner*", "*scanner has*", "*scanner comes with*", etc, for the scanner class).

To associate features and their corresponding opinions, some researchers considered that a product feature and its opinion words/phrases usually co-occur within a certain distance in the text (Hu & Liu, 2004; Kim & Hovy, 2004). Hu and Liu (2004) focused more on adjacent adjectives that modify feature nouns or noun phrases, than other opinion words/phrases. Kim and Hovy (2004) explored the following four sizes of regions which may contain both of product features and their opinions: (1) full sentences; (2) words between the opinion holder (i.e., the reference to the person expressing the opinion) and the topic; (3) region 2 +/- two words; and (4) from the first word behind the holder to the end of sentences. The forth region was found to outperform others.

However, the simple statistics-based approaches (e.g. co-occurrence) are not sufficient in some situations, for example, if more than one feature or topic is mentioned in a sentence. Yi et al. (2003) applied complicated linguistic analysis to identify associations between entities (i.e. features, topics) and opinions at finer granularity within sentences. They focused on analyzing the grammatical structure of sentences and representing it using a formal T-expression (e.g. <*camera, take, excellent picture*> for the sentence "*this camera takes excellent pictures*") or B-expression (e.g. <*poor, performance*> for the phrase "*poor performance in a dark room*") and derived associations from the expression. Popescu and Etzioni (2005) took advantage of manually constructed syntactic dependency rules (e.g. *if* $\exists$*(subject, predicate, object = feature)*$\rightarrow$ *potential opinion = predicate*) to find potential opinion phrases (e.g. "*I hate this scanner*") associated with the known product features. They then used an unsupervised collective technique – relaxation labelling – to determine the polarity of the lexical head word of each potential opinion phrase in the context of an associated feature and sentence, which was expressed in a (*word, feature, sentence*) tuple, e.g., (*sluggish, driver, "I am not happy with this sluggish driver"*). The phrases whose head words have been assigned positive or negative labels were then retained as opinion phrases.

The work reported in this paper is most closely related to the studies by Kim and Hovy (2005) and Popescu and Etzioni (2005) and addresses the problem of derivation of extraction patterns involving features and sentiment expressions. However rather than using manually constructed extraction patterns or very general extraction cues such as sentence boundaries, we propose a method that derives extraction patterns from dependency tree parses.

3. Generating feature-based product report

We identify three subtasks in the overall problem of analysing customer reviews with respect to specific features of a product as follows:

1. Construction of a semantic lexicon containing terms that refer to features of a product that are important to the customer. The compilation of such a resource requires approaches fundamentally different from those used in the mainstream of lexical acquisition: the semantic relations between the terms are not adequately covered by the synonymy, hyponymy, or part-whole relations. The terms need to be identified and arranged into semantic classes from the point of view of their importance to the customer, rather than an objective measure of semantic similarity. For example, statements containing the terms "windows", "bottled

water", "access card", and "welcome note" all contribute to the subjective evaluation of a hotel room, but have no apparent semantic relation outside this context.

2. Construction of a lexicon with expressions containing either negative or positive sentiment. The lexicon needs to be customised to a particular kind of product, because there are many important sentiment-bearing expressions that normally would not be considered as such or would be expressing different kind or intensity of sentiment when used outside the product review or even attributed to a different feature of the same product. For example, "small" expresses a highly negative sentiment when it is attributed to a hotel room and a positive one when attributed to a price.

3. Given comprehensive lexical resources encoding different ways to refer to product features and to express sentiment towards them, the next problem to recognise the semantic relation between the two kinds of expressions in the text of the review. Again, this task has a certain resemblance to a well-known text mining problem, that of information extraction (IE), where one needs to identify and label relations between named entities (such as WORKS_AT (PERSON,ORGANISATION)). The important difference however is that while in IE the related entities have very concrete meaning and hence the relation between them can be captured with a relatively limited number of lexico-syntactic constructions, in the opinion extraction context, the terms have much more varied semantics and as a result, there is a much broader variety of linguistic constructions that express the relation between the two. Hence a direct application of IE techniques to the problem at hand does not appear feasible.

In the present paper, our specific goal is to address the third subproblem and to develop ways to establish the semantic relationship between the ***topic*** of the sentiment (a term for a particular product feature) and the ***sentiment*** itself (an expression expressing a subjective evaluation of the topic).

4. Relating opinion topics and sentiment expressions

In the area of information extraction, semantic relations between entities are extracted using linguistic cues such as the dependency relations between words as well as lexical expressions with relational semantics such as verbs and prepositions. For example, an extraction pattern can have the following form:

PERSON →subj→ *is employed by* ←obj← ORGANISATION

where the verb *employ* and the subject and object positions near it are used as cues to recognise the WORKS_AT relation between two entities.

In order to cater for a much broader variety of ways in which feature terms and sentiment expressions can be related in the text, we attempt to capture such relations using looser extraction patterns, such that would allow for indirect syntactic relations between words of potential interest. More specifically, we use patterns which represent a path in a dependency tree consisting of one or more dependency links, at both ends of which there are placeholders for a topic and a sentiment expression.

Figures 1 and 2 illustrate such patterns and their mapping to concrete sentences expressing subjective evaluation. In Figure 1, in the first sentence, the sentiment-bearing word *luscious*

expresses sentiment towards both *bed* and *room*, while in the second, *elegantly* expresses sentiment towards *lobby*.

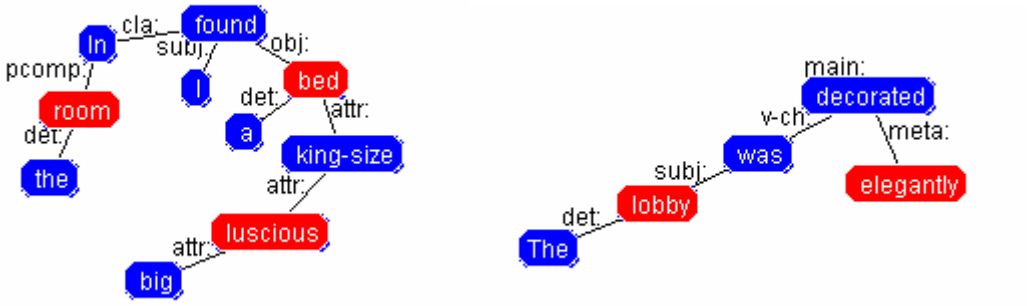

Figure 1. Dependency trees for "*In the room I found a big, luscious king-size bed.*" and "*The lobby was elegantly decorated*". Each edge in the tree has a label specifying the dependency relation between words in the sentence.

Figure 2 shows the tree paths linking topics and sentiment words in these sentences. We assume that there is a conceptual relation between a topic and a sentiment if they appear along a path within a certain distance from each other and study the effect of different distance thresholds on the accuracy of the extracted relations.

T:N→ PREP→ V ← N←A← S:A	found a luscious king-size bed in the room
T:N→A←S:A	luscious king-size bed
T:N← AUXV→ V ← S:ADV	lobby was decorated elegantly

Figure 2. Examples of opinion extraction patterns and fragments of the sentences corresponding to the patterns. The part-of-speech tags designate parse tree nodes, "T" the placeholder for a topic and "S" for a sentiment word.

Once a relation between words is established, we determine whether the relation is negated or postulated by checking if the main verb in the clause at hand is modified by a negative particle or adverb. In case the relation is negated, the polarity of the sentiment is switched. After topic-sentiment word pairs have been identified in the text, for each feature we calculate an average of the intensity scores per mention of the feature in the review, both for positive and negative sentiments. The final evaluation score for the feature is the difference between the averages of the positive and negative scores. Finally, in order to deal with cases when an actual product feature and its sentiment word appear in different sentences and thus to improve the recall of our system, we resolve pronoun references in the document using the MARS system (Mitkov 1998) and substitute the pronouns for the heads of their antecedent noun phrases.

5. Evaluation

5.1 Data

Corpus. The evaluation data used in this study were 268 reviews of hotels automatically downloaded from the epinions.com web site[5]. From the downloaded pages, the plain text of the reviews and the star ratings assigned to the hotels by the reviewers were extracted.

[5] The URLs of the reviews will be made publicly available for interested parties via the Internet.

Frequent domain terms in the corpus were identified using TermExtractor (Navigli and Velardi, 2004). The reviews were processed with the help of the FDG dependency parser (Jarvinen and Tapanainen, 1997).

Lexicons. A lexicon with expressions describing different hotel features was constructed after manual inspection of frequency lists for single nouns and multiword nouns extracted by TermExtractor. Based on that, the following six features of a hotel were identified:

- LOCATION: the area where the hotel is situated, how accessible the points of interest in the area are from this hotel, how far the airport is;
- FOOD: food and drinks served in the hotel;
- ROOM: the room interior and its facilities;
- SERVICES: the services offered by the hotel, how helpful the staff is;
- FACILITIES: the facilities in the hotel outside those found in the room, e.g. a swimming pool, lounge, lobby, casino;
- PRICE: what is the value for money for this hotel.

We experiment with three different general resources encoding sentiment words: General Inquirer's Harvard-4 dictionary (GI, Stone et al. 1966), SentiWordNet (SWN, Esuli & Sebastiani 2006), and a subsection of the Roget thesaurus manually annotated for both the polarity and the intensity of sentiment (Heng 2004). It should be noted that the GI lexicon encodes only the polarity of each sentiment word, but not its intensity; therefore all positive words in the lexicon were assumed to have the intensity of "+1" and negative "-1". SWN encodes intensity scores for different senses of a word; in order to apply the resource to our task, a single score for each word was calculated by averaging the scores of its WordNet synsets.

5.2 Evaluation method

The quality of the extracted topic-sentiment pairs was evaluated in two different experiments: against the overall hotel ratings assigned by the authors of the reviews themselves and against grades assigned manually to the six hotel features by human judges.

In the first experiment, we evaluated the method against the overall grade for the hotel assigned by the authors of the reviews. After an overall score for a review was calculated by averaging scores for its individual features, we measured its correlation with the grades given by the authors. Because in our corpus the authors' grades are highly skewed towards five-star grades (out of the 268 reviews, 240 reviews have five stars, 26 have four stars and 2 have three stars), for this experiment we randomly selected 26 five-star reviews and created all their possible pairs with the 26 four-star reviews, obtaining a total of 676 pairs. The experimental task was to decide which review in each pair expressed a more positive sentiment. The accuracy of the method was measured as the proportion of correct decisions to the total number of review pairs, with ties, i.e. case when the algorithm assign the same score for both reviews, contributing to the accuracy score with half the weight of a correct decision:

$$A = \frac{\#correct + 0.5 \cdot \#ties}{\#total}$$

The data for the second experiment consisted of 19 reviews for two different hotels selected randomly from the corpus. Two judges were asked to read the reviews and assign a grade for each of the six hotel features, based on the text of the review, on the 5-point Likert scale, ranging from "poor" (1) to "excellent" (5). The correlation between the annotators' grades

was significant at the 0.01 level according to Pearson's *r* coefficient (*r*=0.453). During the experiment, we measured the correlation between the system-assigned sentiment scores for each feature and the average of those assigned by the judges.

The proposed topic-sentiment extraction method was evaluated against two baselines. The first one consisted in creating topic-sentiment pairs based on their co-occurrence within the same sentence, and the second – within the same clause (clauses were extracted from the parsed text as subtrees in the sentence trees whose root corresponded to a main verb; each clause could contain only one main verb).

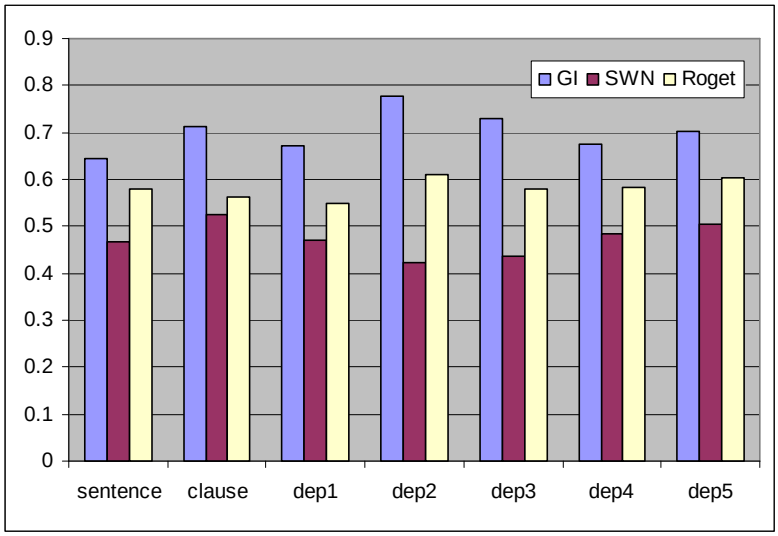

Figure 3. A comparison of different thresholds on the distance between a topic and a sentiment word in a parse tree (dep 1 to dep 5), against the sentence- and clause-based delineation of the context of their co-occurrence.

5.3 Results

Figure 3 shows the results of the first experiment, comparing the accuracy scores achieved when using different thresholds on the distance in the parse tree and when using the sentence or clause as the context of the words' co-occurrence. These results suggest the best way to relate a topic and a sentiment is to allow for two or three edges between them in a dependency tree: these settings produced one of the top accuracy scores for the GI and Roget lexicons. Using the clause to delimit the co-occurrence context seems to produce good results as well, also on the SWN data, which again indicates that related words appear relatively close to each other in the sentence, but they are not necessarily directly related by a dependency relation.

Tables 1 and 2 present results of the second experiment. For each sentiment lexicon (SWN, GI, and Roget), Table 1 describes the ten configurations of the method with the highest correlation, measured in terms of Pearson's *r*, with the average of the judges' grades. Table 2 similarly presents the top ten configurations with prior resolution of pronominal anaphora. Table 3 compares overall scores for a review derived from scores for individual features with and without prior anaphora resolution. In all the three tables, statistically significant positive correlation values are shown in bold.

Comparing the results in the two tables (Tables 1 and 2), we see that anaphora resolution helps to substantially increase the correlation of the system's score with those of the humans, by up to 24 points. In many cases, statistically significant correlation could be achieved only after performing anaphora resolution (LOCATION, FOOD, and ROOM). An increase in the

correlation values is observed also when calculating an average score for a review from the scores of individual hotel features (see Table 3).

LOCATION		FOOD		ROOM	
SWN dep2	0.447	GI dep2	0.335	SWN s	0.247
GI s	0.437	Roget c	0.319	SWN c	0.182
GI dep4	0.334	Roget s	0.278	Roget dep2	-0.03
GI dep3	0.278	GI c	0.263	Roget dep1	-0.03
GI dep5	0.272	SWN s	0.262	SWN dep3	-0.09
GI dep1	0.259	GI dep3	0.210	SWN dep5	-0.09
GI c	0.254	GI dep1	0.195	SWN dep2	-0.12
GI dep2	0.246	SWN dep4	0.064	GI dep2	-0.12
SWN c	0.223	SWN dep5	0.027	GI c	-0.13
SWN dep1	0.064	SWN c	-0.03	Roget s	-0.15
SERVICES		FACILITIES		PRICE	
Roget dep2	**0.536**	GI dep1	**0.459**	SWN c	**0.525**
SWN dep1	**0.483**	SWN c	0.400	SWN dep2	0.444
GI dep2	0.407	Roget dep4	0.277	SWN s	0.412
GI c	0.378	SWN dep4	0.269	Roget dep2	0.384
Roget dep4	0.363	Roget s	0.260	SWN dep1	0.360
Roget dep5	0.329	GI dep3	0.249	Roget dep1	0.318
GI dep5	0.311	Roget dep5	0.240	SWN dep3	0.311
Roget dep3	0.309	Roget dep2	0.228	Roget dep5	0.269
Roget s	0.307	GI s	0.203	Roget dep4	0.269
GI dep4	0.285	SWN dep5	0.197	SWN dep4	0.213

Table 1. Correlation according to Pearson's *r* between different thresholds on the distance between words and the average of human judges' grades without prior anaphora resolution (c – sentence-based delineation of the co-occurrence context, s – clause-based delineation, dep1…5 – distance thresholds in the dependency tree).

LOCATION		FOOD		ROOM	
GI dep5	**0.514**	GI dep1	**0.579**	Roget c	0.442
GI dep4	**0.491**	GI c	0.210	GI dep2	0.249
GI dep3	**0.460**	GI dep2	0.135	GI c	0.164
SWN dep2	**0.458**	SWN dep5	0.133	Roget dep4	0.115
GI s	0.444	SWN dep4	0.118	Roget s	0.106
GI dep2	0.376	Roget c	0.098	Roget dep2	0.071
GI dep1	0.296	SWN s	0.080	SWN dep2	0.062
SWN dep5	0.155	Roget dep1	0.058	Roget dep5	0.060
SWN c	0.095	Roget s	0.031	Roget dep1	0.056
SWN dep4	0.032	GI dep4	0.009	SWN s	0.050
SERVICES		FACILITIES		PRICE	
Roget s	**0.487**	GI dep1	**0.494**	SWN c	**0.515**
Roget dep5	0.352	GI dep2	**0.466**	SWN dep2	0.333
Roget dep4	0.305	Roget s	0.261	SWN dep1	0.326
Roget dep3	0.262	Roget dep3	0.239	Roget dep2	0.319
GI dep2	0.256	Roget dep5	0.233	GI dep2	0.303
Roget dep2	0.255	SWN dep5	0.226	SWN dep3	0.297
GI dep4	0.249	Roget dep2	0.223	SWN dep4	0.284
GI dep5	0.248	SWN dep1	0.202	SWN s	0.273
GI s	0.224	SWN s	0.202	Roget dep1	0.263
SWN dep1	0.185	Roget dep4	0.191	Roget dep4	0.232

Table 2. Correlation according to Pearson's *r* between different thresholds on the distance between words and the average of human judges' grades with prior pronominal anaphora resolution.

The results further indicate that shorter distances between topics and sentiment words typically deliver the best performance overall, which is also consistent with the findings of the first experiment. At the same time, different threshold settings seem to work better for some feature names than others. For example, longer distances between words and their co-occurrence inside a sentence seem to be the best way to recognise sentiment statements about services of a hotel.

One other interesting finding is that different lexicons perform differently with respect to different features: while, as in the first experiment, the GI lexicon seems to be better suited for the domain at hand, SentiWordNet yields the best results on the PRICE feature, while Roget on the SERVICES feature. This appears to illustrate the importance of domain-customisation of the lexicons used by the system.

AR-		AR+	
Roget dep2	**0.234**	GI dep2	**0.293**
GI dep2	**0.234**	GI dep1	**0.238**
Roget s 1	0.188	Roget s 1	**0.233**
GI dep3	0.161	Roget dep2	0.146
Roget dep4	0.152	GI dep4	0.142
Roget dep5	0.151	Roget dep5	0.129
GI dep4	0.149	Roget dep3	0.122
Roget c 1	0.140	GI s 1	0.116
GI c 1	0.136	GI c 1	0.115
GI dep1	0.133	GI dep5	0.115

Table 3. A comparison of overall scores for a review derived from score for individual features with (AR+) and without (AR-) prior anaphora resolution.

Table 4 illustrates some of the patterns discovered by the algorithm when using a 3 edges limit on the distance in the parse tree.

Patterns	Frequency	Examples
T:N ← S:A	282	*wonderful* **room**, *lovely* **service**, *attentive* **staff**, *comfortable* **décor**, *wonderful* **views**, *elegant* **lobby**, *luxurious* **bathrooms**
T:N → V ← S:A	269	**staff** are *friendly*, **price** was *outrageous*, **lunch** is *expensive*, **rooms** looked *dark*, **menu** looked *decent*, find **staff** *rude*
T:N ← N ← S:A	95	*elegant* lobby **bar**, *outstanding* lake **views**, *amazing* breakfast **buffet**, *wonderful* salad **bar**, *spectacular* ocean **view**
T:N → V ← N ← S:A	50	**service** did *excellent* job, **rooms** were *adequate* size, **baths** are *wonderful* dream, **food** is *attractive* feature, **bed** had *fantastic* mattress, **place** has *wonderful* atmosphere, **restaurants** provided *excellent* variety
T:N → N → V ← S:A	45	**room** proportions are *generous*, **room** service is *exquisite*, **pool** staff is *attentive*, **price** structure is *consistent*, **bath** towels are *plentiful*, **bath** products were *excellent*

Table 4. Examples of frequent patterns extracted using a 3 edges limit on the distance in the parse tree. Topics are shown in bold, sentiment words in italics.

6. Conclusion

In this paper we have addressed the task of automatic analysis of sentiment expressed in a product review towards different features of the product. The specific problem we examined was automatic recognition of semantic relations between expressions describing product features and sentiment-bearing expressions. This problem is key to generating numerical feature-based reports on customer satisfaction with a product based on its customer reviews. Nonetheless, to our knowledge the work presented here is the first that explicitly addresses the relation extraction problem in the context of opinion mining. Despite certain affinities with information extraction, extracting opinions about product features has its unique challenges, such as the semantic diversity of expressions referring to product features and of linguistic constructions explicating their relations with sentiment expressions. We have proposed a novel way to recognise such relations which makes use of extraction patterns derived from dependency parses of sentences.

Our main findings can be summarised as follows. Topic-sentiment relations are best captured by using short distances between words inside a parse tree, such that one word is up to 3 edges away from the other in the tree. While this setting is the most optimal in the general case, it seems that sentiment towards a particular feature may be more accurately captured with the help of altogether different extraction patterns. Furthermore, our experimental results indicate that the performance of the opinion extraction system greatly depends on a particular lexicon used, with different lexicons performing differently relative to different product features.

In general, by generating evaluation scores with statistically significant correlations with human judgements, the proposed approach has demonstrated quite promising results. Yet, there is a substantial room for further improvement, because the lexicons used in this study were all general-purpose. In the future, we intend to work on automatic customisation of domain lexicons with respect to each product feature. Another line of extension of this work is concerned with learning extraction patterns from a training corpus.

7. Acknowledgement

This work has been partially supported by the EU funded project QALL-ME (FP6 IST-033860).

References

World Tourism Organization. (2007). http://www.world-tourism.org/

Bruce, R., & Wiebe, J. (2000). Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering*, 6(2).

Das, S., & Chen, M. (2001). Yahoo! for amazon: Extracting market sentiment from stock message boards. In *Proceedings of the 8th Asia Pacific Finance Association Annual Conference* (APFA-2001).

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation* (LREC-2005), pp.417-422.

Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th International Conference on Computational*

Linguistics (COLING-2004), pp. 841-847.
Hatzivassiloglou, V., & Wiebe, J.M. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of 18th International Conference on Computational Linguistics (COLING-2000).* San Fransisco, CA: Morgan Kaufmann Publishers Inc.
Hu, M., & Lin, B. (2004). Mining opinion features in customer reviews. In *Proceedings of the 9th National Conference on Artificial Intelligence* (AAAI-2004).
Heng, A. (2004). An exploratory study into the use of faceted classification for emotional words. *Master Thesis*. Nanyang Technological University, Singapore.
Tapanainen, P., Jarvinen, T. (1997). A non-projective dependency parser. In: Proceedings of the 5th Conference of Applied Natural Language Processing, Washington D.C., USA. pp. 64–71.
Kim, S.-M., & Hovy, E. (2004). Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics* (COLING-2004), pp.1367–1373.
Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. In: Proceedings of the 18th International Conference on Computational Linguistics (COLING'98/ACL'98), Montreal, Canada. pp. 867-875.
Navigli R. & Velardi, P. (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Websites. Computational Linguistics, 30(2), MIT Press, pp. 151-179.
Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing* (EMNLP-2002), pp. 79–86.
Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, (ACL-2004), pp. 271–278.
Popescu, A. M., & Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the 2005 Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing* (HLT/EMNLP-2005).
Riloff, E., Wiebe , J. , & Phillips, W. (2005). Exploiting subjectivity classification to improve information extraction. In Proceeding of the 20th National Conference on Artificial Intelligence (AAAI-2005).
Stone, P., Dunphy, D., Smith, M., Ogilvie, D., & associates. (1966). The General Inquirer: A computer approach to content analysis. Cambridge, MA: The MIT Press.
Tong, R.M. (2001). An operational system for detecting and tracking opinions in on-line discussions. In *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification* (pp. 1-6). New York, NY: ACM.
Turney, D. (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-2002, pp.417-424). Morristown, NJ: ACL.
Wilson, T., Wiebe, J., & Hwa, R. (2004). Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of the 19th National Conference on Artificial Intelligence* (AAAI-2004).
Yi, J., Nasukawa, T., Bunescu R., & Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings* of the 3rd *IEEE International Conference on Data Mining* (ICDM-2003), pp. 423-434.