Generating Domain-Specific Ontology from Common-Sense Semantic Network for Target-Specific Sentiment Analysis

Ashish Sureka
IIIT Delhi
New Delhi, India
ashish@iiitd.ac.in

Vikram Goyal
IIIT Delhi
New Delhi, India
vikram@iiitd.ac.in

Denzil Correa
IIIT Delhi
New Delhi, India
denzilc@iiitd.ac.in

Anirban Mondal
IIIT Delhi
New Delhi, India
anirban@iiitd.ac.in

Abstract

Target or feature specific sentiment classification of a product review consists of extracting opinion or sentiment expressing phrases, extracting the targets (features in a product domain), computing the semantic orientation of the sentiment expressing phrase and assigning the sentiment expression to the product feature it targets. Each of the tasks is fundamental to the problem of target-specific sentiment analysis. In this paper, we present an algorithm to automatically build a domain-specific ontology (a graph consisting of product features and semantic relations between them) which can be used as a lexical resource for performing target-specific sentiment analysis in real-time. We use ConceptNet (a large semantic network of commonsense knowledge) for extracting domain-specific ontology. We evaluate our approach on publicly available pre-annotated dataset from *phone* and *camera* domain. The advantages of our approach are that it uses a resource which is created by volunteers on the Internet and not by trained or specialized knowledge engineers. Another advantage is the product feature lexicon that is created is in the form of semantically rich domain ontology rather than a flat list of phrases. We investigate the usefulness of commonsense knowledge for generating domain-specific ontology for feature extraction task in sentiment analysis application and conclude that the approach is feasible.

1 Introduction

Opinion mining and sentiment analysis is a subtopic of natural language processing and text mining that deals with the automated discovery and extraction of knowledge about people's sentiments, evaluation and opinions from textual data such as personal blogs, review websites and customer feedback forms. Opinion mining and sentiment analysis is an area that has received significant interest in recent times because of its practical usage and application in today's environment. One example usage of sentiment analysis systems is user opinion summarization and sentiment extraction for a particular model or brand of digital camera and its various features based on reviews posted on an e-commerce website. Another example is summarizing the general opinion of people on policy decisions made by a political leader based on user generated content on the Web. Such analysis is quite useful to product manufacturers who want to get insight into the voice of the customer, for buyers who want to make purchase decisions based on the experience of others who have used the same product and also companies or governments who want to get feedback on what people think about their company or a policy. There has been a surge in research activity on opinion mining and several commercial products are now available in the market. A detailed survey on opinion mining and sentiment analysis is provided by Pang et al. and a chapter on sentiment analysis is written by Liu in Handbook of Natural Language Processing (Liu et al., 2010) (Pang et al., 2008). Feature-based (also referred as aspect-based or target-specific) opinion analysis consists of a fine-grained analysis with respect to the attributes, features or aspects of an object (a product, an organization or a person) commented on by reviewers. In the context of a product domain, the problem of feature-based sentiment classification can be decomposed into the following four main tasks:

1. Product feature extraction (for example battery life, image quality and resolution in a camera domain and seating comfort, maximum speed, wheels and steering in a car domain).

2. Opinion and sentiment expressing phrase extraction (for example extremely comfortable,

not smooth, quite heavy, good and bad).

3. Polarity classification or semantic orientation determination of sentiment expressing phrases (for example the word denotes a positive sentiment and the word bad denotes a negative sentiment or evaluation).

4. Intensity or strength determination of sentiment expressing phrases (the word excellent is a strong positive word whereas the word good is a weak positive word).

The focus of the work presented in this paper is on automatic product feature extraction from customer reviews.

1.1 Related Work

Automatic product feature extraction (a sub-problem of target-specific sentiment analysis) is an important problem as the numbers of product domains are huge and it is hard to manually create and maintain a comprehensive product feature lexicon for all domains (manual approach is not scalable). Product feature extraction is fundamental to the problem of target-specific sentiment analysis and forms a building block of a larger system. Product feature extraction from customer reviews is a challenging problem and has thus received a great deal of interest in recent times. In this section, we provide an overview of the related work by presenting the main idea behind traditional approaches. To the best of our knowledge, the paper by Chih-Ping Wei et al. is the most recent paper specifically on the problem of product feature extraction (Wei et al., 2009). Chih-Ping Wei et al. provide a comprehensive literature survey and classify existing product feature extraction techniques into supervised and unsupervised techniques (Wei et al., 2009). The work done by Wong et al. (Wong et al., 2005) (Wong et al., 2008) falls into the category of supervised learning approach (requiring pre-annotated training dataset) as they employ Hidden Markov Models and Conditional Random Fields as the underlying learning method for extracting product features. The technique proposed by Hu et al. (Hu et al., 2004) (Hu et al., 2004b) falls into the category of unsupervised learning as their techniques not require pre-annotated training dataset. The main features of the approach by Hu et al. is the application of association rule mining algorithm to discover product features (nouns and noun-phrases in a review represented as item-sets). The approach by Wei et al. is an improvement over the approach by Hu et al. (Wei at al., 2009) which introduces an additional step of semantic-based refinement that leverages subjective adjectives from General Inquirer. Another class of solution for solving the problem of product feature extraction is based on information extraction techniques (tagging specific pieces of information from a free-form text based in hand-crafted rules or models induced by applying machine learning techniques). The approaches proposed by Popescu et al. (Popescu et al., 2005) and Kobayashi et al. (Kobayashi et al., 2004) falls into the category of information-extraction based techniques. Ferreira et al. (Ferreira et al., 2008) presents a comparative study of feature extraction algorithms in customer reviews and provides an overview of related work on product feature extraction. The paper by Ferreira et al. and Chih-Ping Wei et al. provides a comprehensive overview of product feature extraction algorithms based on various techniques and approaches like: supervised learning methods like Hidden Markov Models and Conditional Random Fields, application of information extraction, application of association rule mining, creation of pre-build databases, application of Point-wise Mutual Information and usage of General Enquirer. Ferreira et al. systematically compares two feature extraction algorithms: an approach by Nasukawa et al. (Nasukawa et al., 2003) which consists of identifying candidate features by applying a set of POS patterns and pruning the candidate set based on the Log Likelihood Ratio test and the other approach by Hu et al. (Hu et al., 2004b) which consists of performing noun-phrase extraction and then applying association rule mining for identifying frequent features.

1.2 Paper Contributions

In this section, we present the main idea behind our solution and list the advantages of our solution over traditional approaches. The key difference between the proposed solution in this paper and previous approaches is that we employ a semantic network of common-sense knowledge-base (ConceptNet) for automatically creating a domain specific ontology of product features and attributes. Previous approaches create a flat list of product features whereas we create ontology where product features are concepts or nodes in a seman-

tic network connected to other nodes using multiple types of semantic relationship. Thus, the product ontology and lexicon created from our approach is semantically richer than the lexicon created from previous approaches. Creating ontology from ConceptNet gives the ability to perform reasoning and inferences (i.e. if direct knowledge is not available one can exploit the network of semantically related knowledge to make inferences). The types of semantic relations in ConceptNet are more than in WordNet. For example, we not only make use of semantic relationships like *IsA* and *HasA* but also other relationships like *CreatedBy*, *MadeOf*, *PartOf*, *DesireOf* and *DefinedAs*. Another difference is that the nodes in ConceptNet can be higher level concepts (compound concepts) such as "battery life", "read book" and "eat food" which are useful to us for solving the problem at hand. ConceptNet knowledge is also informal as unlike WordNet it is not handcrafted by experts and knowledge engineers but contributed by thousands of ordinary people as volunteers on the Internet. The scope of ConceptNet is general world knowledge and not limited to a specific domain. The concepts in ConceptNet are informal in nature which is useful for the task of extracting product features from user comments. ConceptNet has redundant concepts and multiple ways of expressing the same concept which is useful for the problem at hand (A camera is used for: photography, record images, making fotos, take photographs, take pictures). To the best of our knowledge, the application of ConceptNet for solving sub-problems in sentiment analysis and opinion mining is an unexplored area and this work is a step in the direction of our research motivation on investigating the usefulness of ConceptNet for opinion mining applications. The contributions of this paper are as follows:

1. A novel use of common-sense knowledge-base (ConceptNet) for automatically constructing product domain ontology for target-specific sentiment analysis. The product domain ontology is represented as a directed and labeled graph which is semantically richer than a lexicon consisting of a flat list of product features for a particular domain.

2. An algorithm to extract product features from customer reviews using text chunking and then pruning based on the product domain ontology created using common-sense knowledge-base. The extracted product feature is connected to the product domain by finding all paths from the product feature node to the product domain node in a directed and labeled graph representing the product domain ontology.

2 Solution Approach

We leverage ConceptNet (which is machine-interpretable semantic network representing common-sense knowledge) for creating a domain specific ontology. The common-sense knowledge present in ConceptNet is collected from volunteers on the Internet since the year 2000 and represents facts that ordinary people knows about the world (Havasi et al., 2007). The data present in ConceptNet is contributed by ordinary people unlike lexical resources such as WordNet and FrameNet which are mainly created by trained and specialized knowledge engineers. As ConceptNet is a semantic network, it consists of nodes connected by edges. The nodes represent the concepts and the edges represent predicates. Predicates express semantic relationships between two concepts. Some relationships between concepts in the ConceptNet semantic network are: IsA, MadeOf, UsedFor, CapableOf, DesireOf, CreatedBy, InstanceOf, PartOf, PropertyOf and EffectOf (Havasi et al., 2007). The relation types are grouped into various thematic such as: Things (IsA, PropertyOf, PartOf, MadeOf, DefinedAs), Agents (CapableOf), Events (PrerequisiteEventOf, FirstSubeventOf, SubeventOf, LastSubeventOf), Spatial (LocationOf), Causal (EffectOf, DesirousEffectOf), Functional (UsedFor, CapableOfReceivingAction) and Affective (MotivationOf, DesireOf) (Liu et al., 2004),(Liu et al., 2004B). In ConceptNet, an assertion is uniquely defined by five properties: language, relation, concept1, concept2 and frequency. The Language property defines the language an assertion is expressed in (such as English). The Relation property defines the relation or the name of the predicate that connects the two concepts in the assertion (such as IsA, PartOf). Concept 1 and Concept 2 defines the first and the second argument of the relation (words and phrases). The Frequency property expresses how often the given concepts would be related by the given relation, ranging from never to always. Also for each assertion, there is a field which defines the

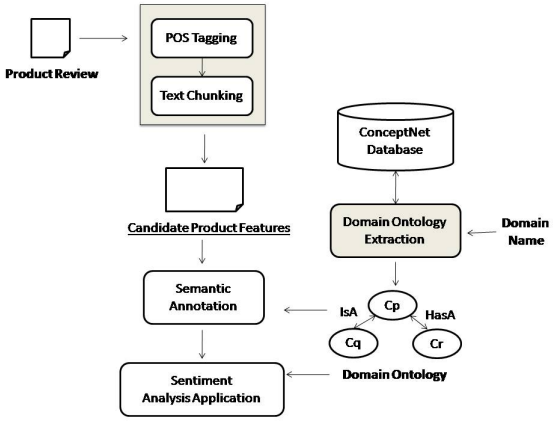

Figure 1: Architecture diagram of the product ontology and product feature extraction system

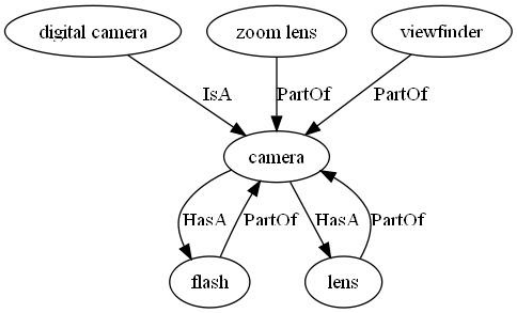

Figure 3: Snapshot of Camera product-domain ontology derived from ConceptNet

assertion type. The value of the assertion type is +1 if the assertion makes a positive statement (such as Diamonds are pretty) and -1 if it makes a negative statement (such as a person doesn't want anxiety). The process fo creating a domain specific ontology is a multi-step process which exploits certain types of relations and connection between words.

Figure 1 presents the proposed multi-step process for automatically creating a domain specific ontology from ConceptNet and then using it for extracting product-features from customer reviews. The domain ontology is created automatically from the product domain name (for example *camera* or *telephone*). Given a domain name (for example, *camera* or *telephone* which exists as a concept in ConceptNet), we first extract all assertions (forward as well as reverse) where the domain name is one of the concept in the assertion. We add only those assertions in the domain ontology which satisfies any one of the rule. One rule consists of forward assertions (i.e., the concept searched occurs as concept1 in the assertion) where the relationship is of type *IsA*, *HasA*, *DefinedAs*, *MadeOf* and *CreatedBy* and the assertion type is positive. The other rule consists of reverse assertions (i.e., the concept searched occurs as concept2 in the assertion) where the relationship is of type *IsA*, *part of* and *DefinedAs* and the assertion type is positive. We create a directed labeled graph from the extracted assertions. The vertices in the graph represents the concepts (concept1 and concept2) present in the assertion, the direction of the arc in the graph corresponds to the direction (forward or reverse) of the assertion (between concept1 and concept2) and the label of the graph corresponds to the semantic relation (such as *IsA*, *HasA*, *MadeOf*, *PartOf*). In the first step of ontology creation, we expand only the domain name (the seed) to one level. We apply the same procedure two more times: one for the first level nodes connected to the domain name and the second time to the second level nodes connected to the domain name. Figure 2 and 3 illustrates a snapshot (and not the complete ontology) of product ontology automatically extracted for the *telephone* and the *camera* domain. We chose to present an ontology (for the *telephone* domain) which is big enough to fit one complete page inorder to showcase the richness of the ontology that can be derived. Automatic creation of product domain ontology is one aspect of the proposed solution. The other aspect is to extract candidate product features from the product review and assign it to the nodes in the ontology for further analysis. As shown in Figure 1, product domain ontology is constructed offline and product feature extraction is performed for each customer review. To identify candidate product features, we apply a technique called as text *chunking* wherein non-overlapping multi-sequence tokens (for example noun-phrases) are extracted from a sentence. *Chunking* is a commonly used pre-processing step for natural language processing tasks such as named-entity extraction, information extraction and syntactic parsing of text. We perform sentence segmentation on each product review (a review is basically a sequence of sentences) and then recognize the chunks that consists of contiguous sequence of nouns within each sentence. We used Natural Language Toolkit (NLTK) [1] in our research and development for performing Natural Language Processing (NLP) tasks such as sentence segmentation,

[1] http://www.nltk.org/

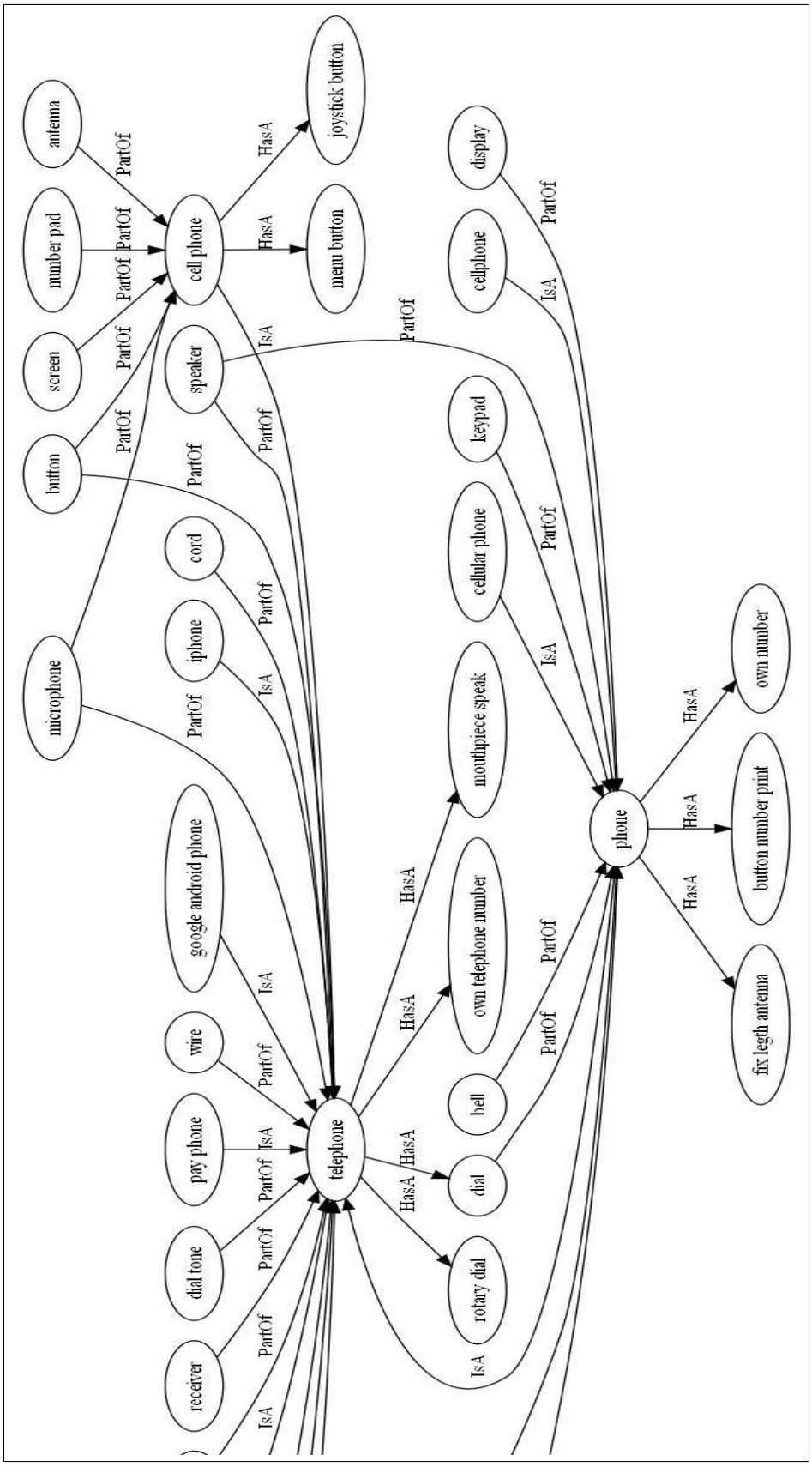

Figure 2: Snapshot of Telephone product-domain ontology derived from ConceptNet

part-of-speech tagging and text chunking (Bird et al., 2004). The chunk grammar (regular expression or linguistic rules that specify how sentences should be chunked) consists of one or more contiguous sequences of noun forms such as Singular Noun, Plural Noun, Proper Singular Noun and Proper Plural Noun. The output of this step is a set of chunks which is pruned in later steps to identify the chunks representing product features and eliminating chunks that are not product features. The text chunks extracted consists of a large number of chunks some of which may not be product features and thus needs to be pruned. We compute a text similarity analysis between the candidate product features (text chunks extracted) and the nodes in the domain ontology. We consider two strings (one from the set of candidate product features and the other from the set of nodes in the product domain ontology) as a match if their Levenshtein distance (also called as edit distance) is less than or equal to two or if there is a common token between the two strings. Since the concepts in ConceptNet are natural language fragments and the candidate phrases extracted from the customer review can have typos, short-forms and informal notations, we apply a fuzzy string matching function. The matched string from the candidate product feature is assigned to the appropriate node in the product domain ontology and the un-matched strings present in the set of candidate product features are omitted. Only the strings in the candidate product features that matched one of the nodes in the product domain ontology is tagged as the final product feature that is extracted from the customer review.

3 Empirical Evaluation

The evaluation dataset for our experiments consists of publicly available (at University of Illinois at Chicago website) annotated customer reviews of consumer electronics products [2]. The annotated reviews were from amazon.com and has been used in the paper by Hu et a. (Hu et al., 2004) and other papers (Ferreira et al., 2008),(Wei et al., 2009). We used the annotated reviews for two products: Canon G3 digital camera and Nokia 6610 cellular phone. We parsed the input data files and computed the number of manually tagged unique product features as well as the total number of product features tagged. The Nokia Phone dataset has 112 manually tagged product feature whereas the total number of product features tagged is 340 (including duplicates as several product features were mentioned in multiple reviews). The Canon Camera dataset has 105 manually tagged product feature whereas the total number of product features tagged is 286. The first step of text chunking (extracting non-overlapping multi-sequence tokens consisting of various noun forms) for the Nokia cellular-phone dataset resulted in a set of 868 chunks which constitutes the candidate feature set. We observed that from the 112 manually annotated product features in the experimental dataset for Nokia phone, 73 phrases were present in the candidate feature set and 39 phrases were not present in the candidate feature set. This amounts to a recall of $73/112 = 65.17\%$. We investigated the part-of-speech tagging results and noticed that phrases like bluetooth, infrared, key lock, pc cable, screensaver, software, vibrate setting, voice dialing and wallpaper which are manually tagged as product features in the experimental dataset were not recognized as candidate features because the constituents words were assigned a non-noun lexical category. For example, bluetooth was assigned a lexical category of Determiner, infrared as Past Participle, key lock as adjective and Noun, pc cable as noun and adjective, screensaver as Adverb, vibrate setting as noun and Gerund, vibrate as infinitive, dialing as Gerund and voice-activated dialing as Adjective and Gerund. In the experimental dataset, Hu et al. (Hu et al., 2004) have provided annotations for product features on which the user has stated an opinion or evaluated the respective product feature. Hu et al. have not manually annotated product features on which no sentiment has been expressed. This issue has also been noted by Ferreira et al. (Ferreira et al., 2008) and hence we notice many product features that are present in the candidate feature set but are not present in the set of manually annotated features in the experimental dataset. There were several product features extracted by the text chunking process but are not part of the manually annotated dataset such as: aol instant messenger software, appearance, audio quality, background wallpaper, battery power, bluetooth functionality, buttons, camera attachment, cdma tri-band, color screens, data kit, display, extra features, family plan, fm radio option, games, signal strength, speaker qual-

[2] http://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

	Telephone	Camera
Nodes in Ontology	433	26
Edges in Ontology	466	27
Targets (Annotated Dataset)	112	105
Targets Extracted from Ontology	43	20
Candidate Targets (Noun Phrases)	868	1032
Targets Extracted from Candidates	219	96

Table 1: Recall results for product feature extraction for telephone and camera domain.

ity, stereo headphones, speakerphone feature, usb, voice recognition and wap browser. We observed that from the 105 manually annotated product features in the experimental dataset for Nokia phone, 70 phrases were present in the candidate feature set and 35 phrases were not present in the candidate feature set. This amounts to a recall of $70/105 = 66.67\%$. Similar to the cellular-phone dataset, there were several product features extracted by the text chunking process in the camera dataset which are not part of the manually annotated product features such as: accessories, add-on flash unit, adjustability, aperture priority, auto exposure settings, autofocus delay, backup battery, battery duration, battery power, buttons, close-up photos, compact flash, control panel, default settings, digital pictures, durability, exposure settings, flash photography, focus range, knobs, lcd panel, lcd viewfinder, lens protector, megapixels, photographs, pictures, remote control, resolution, sharpness, shutter button, user-interface and zoom lens unit.

Table 1 presents experimental result on the publicly available *camera* and *telephone* customer review dataset. The number of nodes (representing features and attributes of the object) and edges (representing semantic relations between the nodes) in the product ontology for the *telephone* object or domain were 433 and 466 respectively. The number of nodes and edges in the product domain ontology for *camera* were 26 and 27 respectively. We noticed that the number of assertions in ConceptNet related to *camera* domain were much less than the *telephone* domain. The overall quality of the product domain ontology is dependent on the amount of data present in ConceptNet for that particular domain. The quality of the ontology affects the performance of our product feature extraction system as only those candidate phrases that are similar (fuzzy similarity) to the nodes in the ontology are extracted. The number of product features that were manually tagged in the evaluation dataset for the *camera* and *telephone* were 105 and 112 respectively. Our system was able to recall 20 and 43 product features respectively. This amounts to a recall of 19.04% and 38.39% respectively. Table 1 presents the number of candidate product features extracted from the corpus of *camera* and *telephone* customer reviews. The number of product features extracted from the candidate phrases (i.e., the candidate phrases that matched at-least one node in the domain ontology) for the *camera* and *telephone* domain were 96 and 219 respectively. Notice that the number 96 and 219 are not unique product features (the same product feature such as *lens* or *battery* can be mentioned multiple times in the customer review corpus) as it is also important for the analyst to capture the frequency of the mention of each product feature. The number of times a product feature is talked about in customer reviews is useful information for the business analyst. Also, a single noun phrase can map to multiple nodes in the domain ontology. This is due to the fuzzy string matching technique employed in our system (for example *zoom lens* and *camera lens* and *lens* candidate phrases extracted from the customer reviews will all map to the node *lens* in the domain ontology). Since, we extract product domain ontology from ConceptNet semantic network and model the extracted ontology as a directed labeled graph, we can apply graph operations on the extracted ontology or graph. This is a useful feature for the business analyst as he can print all paths from an extracted product feature to the root node (the name of the domain). The limitations of our approach are that currently the recall is not high as the extracted product domain ontology does not contain many important concepts (attributes and features of the product) present in the customer review. Any candidate phrase (which is a true positive) which does not match with any one of the

nodes in the product ontology is omitted.

4 Conclusions

This paper investigates the usefulness of common-sense knowledge for extracting product features from customer reviews and constructing a domain ontology for target-specific sentiment analysis. Evaluation on test data consisting of publicly available pre-annotated customer reviews shows that leveraging common-sense knowledge that is shared by the vast majority of people for the task of product feature extraction and domain ontology construction is feasible. Certain types of relationships between concepts and the connection between words and concepts in a semantic network like ConceptNet can be exploited to build a product domain ontology consisting of product features and semantic relations. The accuracy and coverage of the words is a function of the number of concepts, assertions, relations and quality of data in the common-sense knowledge-base.

References

Ana-Maria Popescu, and Oren Etzioni. 2005. *Extracting product features and opinions from reviews Human Language Technology and Empirical Methods in Natural Language Processing*, 339–346

Bing Liu. 2010. *Sentiment Analysis and Subjectivity,* , Second Edition Handbook of Natural Language Processing, (editors: N. Indurkhya and F. J. Damerau),

Bo Pang, and Lillian Lee, 2008. *Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval*, 2(1-2):1135

Catherine Havasi, Robert Speer and Henry Lieberman, 2007. *ConceptNet 3: A Flexible Multilingual Semantic Network for Common Sense Knowledge,* Recent Advances in Natural Languges Processing

Chih-Ping Wei, Yen-Ming Chen, Chin-Sheng Yang, and Christopher C. Yang. 2009. *Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews*, Information Systems and E-Business Management

Guang Qiu, Bing Liu, Jiajun Bu and Chun Chen, 2009. *Expanding Domain Sentiment Lexicon through Double Propagation, 21st International Joint Conference on Artificial Intelligence (IJCAI)*, Pasadena

Hugo Liu and Push Singh, 2004. *Commonsense Reasoning in and over Natural Language, International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES)*, Wellington

Hugo Liu and Push Singh, 2004. *ConceptNet: A Practical Commonsense Reasoning Toolkit, BT Technology Journal*, Volume 22 Kluwer Academic Publishers

J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack 2003. *Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques, 3rd IEEE International Conference on Data Mining*, 427–434

Liliana Ferreira, Niklas Jakob, and Iryna Gurevych, 2008. *A Comparative Study of Feature Extraction Algorithms in Customer Reviews, IEEE International Conference on Semantic Computing*, 144–151

Minqing Hu and Bing Liu, 2004. *Mining and summarizing customer reviews, International Conference on Knowledge Discovery and Data Mining (KDD)*, Seattle, Washington.

Minqing Hu and Bing Liu, 2004. *Mining opinion features in customer reviews, International Proceedings of American association for artificial intelligence (AAAI) conference*, 755–760

Nozomi Kobayashi, Ryu Iida, Kentaro Inui, and Yuji Matsumoto, 2004. *Opinion extraction using a learning-based anaphora resolution technique, International joint conference on natural language processing, Jeju Island*, 173–178

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani, 2009. *Multi-facet Rating of Product Reviews, 31st European Conference on IR Research on Advances in Information Retrieval*, 461–472

Steven Bird, and Edward Loper, 2004. *NLTK: the natural language toolkit, Proceedings of the ACL demonstration session*, Barcelona, 214–217

Tak-Lam Wong, and Wai Lam, 2008. *Learning to extract and summarize hot item features from multiple auction Web sites, Knowl Inf Syst*, 14(2):143–160

Tak-Lam Wong, and Wai Lam, 2005. *Hot item mining and summarization from multiple auction Web sites, IEEE International conference on data mining*, Houston