An Ontology-Based Approach to Opinion Mining

Bilan V.[1)], Bobkov A.[1)], Gafurov S.[1)], Krasnoproshin V.[1)], van de Laar J.[2)] Vissia H.[2)]

1) Belarus State University, 4 Nezavisimost av., Minsk, 220050, Belarus, e-mails: vomavr@yahoo.com, anatoly.bobkov@gmail.com, gafurov@bsu.by, krasnoproshin@bsu.by

2) Byelex Multimedia Products BV Argon 1 – 4751 XC Oud Gastel, The Netherlands, e-mail: j.vandelaar@byelex.com, h.vissia@byelex.com

Abstract: *The paper deals with the identification and extraction of opinion words from arbitrary texts. Special attention is paid to linguistic resources and opinion word sense disambiguation based on opinion collocations within the proposed ontology-based approach.*

Keywords**:** opinion mining, opinion words, opinion collocations, ontology, word sense disambiguation.

1. INTRODUCTION

Natural language processing is one of the key elements in information technologies. The major part of the most interesting and important information is represented as a great variety of texts and documents in a natural language. Opinion mining is gaining much popularity within natural language processing. Web reviews, blogs and public articles provide the most essential information for opinion mining. The field of opinion mining and sentiment analysis is well suited to various types of business and government intelligence applications. This information is of great importance for decision making on products, services, persons, events, organizations.

Creation of systems that can effectively process subjective information requires overcoming a number of new challenges: identification of opinion-oriented documents, knowledge domains, specific opinions, opinion holders, representation of the obtained results.

The purpose of this study is to address the problem of opinion word sense disambiguation. An ontology-based approach is proposed as a solution to the problem.

2. PROBLEM STATEMENT AND SOLUTION

Opinion words are the main constituents of opinion mining and sentiment analysis. Numerous models and algorithms are proposed to identify and extract opinion words, positive or negative assessment of the object being evaluated [1, 2, 3]. But the problem of effective identification and extraction of opinion words and phrases from an arbitrary text, irrespective of the knowledge domain, still remains unsolved. Processing of an arbitrary text necessitates the solution of the problem of opinion word sense disambiguation. Thus, our particular interest is in opinion word sense disambiguation.

The automatic disambiguation of word senses has been a concern since the earliest days of natural language processing [4]. There are several approaches to word sense disambiguation: 1) dictionary- and knowledge-based methods that rely primarily on dictionaries, thesauri, and lexical knowledge bases, without using any corpus evidence [5]; 2) supervised methods that make use of sense-annotated corpora for training [6]; 3) semi-supervised methods that use a secondary source of knowledge such as a small annotated corpus [7]; 4) unsupervised methods working directly from raw unannotated corpora [8].

In information extraction and text mining, word sense disambiguation shows a great potential to be useful in many applications (machine translation, information extraction, lexicography, etc.), but word sense disambiguation has not yet been widely studied and used in opinion mining, especially when processing an arbitrary text.

Opinion word sense disambiguation is of great importance in effective opinion mining. Opinion word sense disambiguation is essential for opinion mining, since the same words in different contexts can have different meaning and different sentiment orientations (e.g. *high price*, *high performance*).

As a possibility to overcome ambiguity of opinion words in an arbitrary text, we propose an ontology-based approach. Ontologies have become common on the World-Wide Web. Ontologies on the Web range from large taxonomies categorizing Web sites (such as on Yahoo!) to categorizations of products for sale and their features (such as on Amazon.com). For any given domain, the ontology represents the concepts which are held in common by the participants in that domain.

Since ontologies explicitly represent knowledge domain semantics (specifications of the terms in the domain and relations among them), they can effectively be used in solving opinion mining problems, opinion word sense disambiguation in particular.

The proposed ontology-based approach was realized in the developed knowledge base, called “Opinion Miner Knowledge Base”, which contains opinion words expressing:

1) personal emotional state (e.g. happy, delighted, proud, sad, angry, horrified);

2) appreciation (e.g. flexible, efficient, stable, reduced, ideal, backward, poor, highest);

3) judgement (e.g. active, decisive, caring, dedicated, intelligent, negligent, evil)

While “judgement” evaluates human behaviors, “appreciation” typically deals with natural objects, manufactured objects, as well as more abstract entities, such as plans and policies. Humans may also be evaluated by means of “appreciation”, rather than “judgement”, when viewed more as entities than as participants, e.g. *lovely medical staff*.

Opinion words can be expressed by: an adjective (*brilliant, happy*); a verb (*like, love, hate, blame*); a noun (*garbage, triumph, catastrophe*); a phrase (*easy to use, simple to use*). Adjectives derive almost all disambiguating information from the nouns they modify, and nouns are best disambiguated by directly adjacent adjectives or nouns.

Information about the force of evaluation (low, high, the highest) and orientation (positive/negative) is also included in the knowledge base. For example, *safe* (low force, positive orientation), *safer* (high force, positive orientation), *the safest* (the highest force, positive orientation), *unsafe* (low force, negative orientation)

In the Opinion Miner Knowledge Base opinion words go together with their accompanying words, thus forming "opinion collocations" (e.g. *deep depression*, *deep devotion*, *warm greetings*, *discuss calmly*, *beautifully furnished*). By an "opinion collocation" we understand a combination of an opinion word and accompanying words, which commonly occur together in an opinion-oriented text. The use of opinion collocations is a way to solve the problem of opinion word sense disambiguation (e.g. *well-balanced political leader* and *well-balanced wheel*) and to exclude words that do not relate to opinions (cf. *attractive idea* and *attractive energy*).

We assume that the number of opinion collocations, which can be listed in a knowledge base, is fixed.

The use of opinion collocations within the ontology-based approach opens a possibility to assign names of knowledge domains to them, because opinion collocations are generally domain specific. For example, *helpful medical staff* ("health care"), *helpful hotel reception staff* ("travel-hotel"), *stable economy* ("economics"), *well-balanced politician* ("politics"). More than one knowledge domain may be assigned to an opinion collocation, e.g. *fast service* ("economics-company", "travel-hotel").

Knowledge domains and their concepts are organized hierarchically to state "part-of", "is a kind of" relations. For example, the knowledge domain "Medical Care" includes:

Medical Care
- medical service
- medical staff
- disease prevention
- disease treatment
- medications

Associative relationships, which relate concepts across the tree structure, are also taken into consideration: 1) nominative relationships describing the names of concepts; 2) locative relationships describing the location of one concept with respect to another; 3) associative relationships that represent, for example, the functions, processes a concept has or is involved in; 4) cause-effect relationships.

Based on the proposed ontology approach, an object of the particular class of interest may have its own specific sets of sub-classes, opinion collocations and evaluation. In the automobile domain, for a car model they can be: engine, transmission, suspension, size, color, design, condition under which an evaluation applies (e.g. driving on slippery roads), a supporting factor for the evaluation.

For example: "The C180K is the *cheapest* in the C-Class range. Everything about the C180K's interior reeks of *superior design* and *craftsmanship*. The *engine revs* as *smoothly* as it sounds. The *steering* was *light*. Whether surmounting cobblestones, concrete or brick, the C180K was a *planted*, *communicative* and *comfortable city car*. The C180K delivers *excellent fuel economy*."

Opinion Miner Knowledge Base also provides additional information about quality characteristics and relationships for different objects on which an opinion is expressed (e.g. *software product* evaluation includes: usability, reliability, efficiency, reusability, maintainability, portability, testability; *travel-hotel* evaluation includes: value, rooms, location, cleanliness, check in/front desk, service).

For example: "The *location* of the Golden Well hotel is *excellent*. The *hotel* is *beautifully furnished* without being overdone. *Check-in* was *fast* and *easy*. The *room* was *fabulous*, and the *breakfasts amazing*. The *bed* was *comfortable* and the *bathroom* was a *pleasure*. *Friendly* and *attentive staff*."

Domain-specific information helps to solve the problem of opinion word sense disambiguation and ensures customized search, i.e. detection of sentences relevant to a given knowledge domain or topic.

3. OPINION-MINING SYSTEM

Opinion Miner Knowledge Base is the core of the developed intelligent opinion-mining system, called "Opinion Miner". The system consists of several modules (Fig.1).

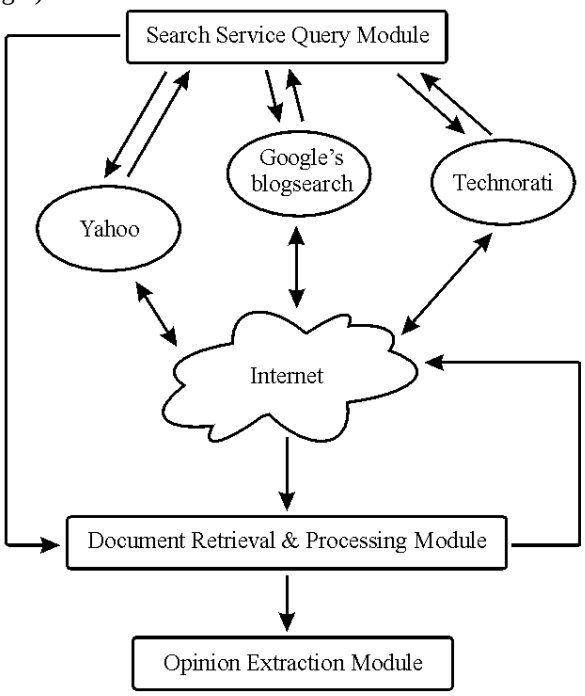

Fig.1 - General scheme of "Opinion Miner"

"Search Service Query Module" detects an object on which an opinion is expressed *(product, organization, person, event, or topic*). A query is sent to popular specialized services that search by blogs and articles. Query parameters are predefined objects on which an opinion is of interest ("base tags" and their synonyms). The result of query processing is a list of references in RSS format.

"Document Retrieval & Processing Module" processes the obtained list by referring to Internet for the full text of an article or a message. The obtained documents are analyzed, relevant documents are collected and passed to the next module.

"Opinion Extraction Module" examines the potential opinion phrases on the basis of information contained in

Opinion Miner Knowledge Base. Identification of the actual opinions is carried out by processing opinion collocations.

Processing of the extracted opinion collocations is carried out in their contextual environment. Opinion Miner checks for the presence of modifiers that can change the force of evaluation and orientation, indicated in the knowledge base.

Let's consider the following example: *reliable company*. This opinion collocation has the following information in the knowledge base: "low force" of evaluation and "positive orientation". E.g. "Sunpak is a *reliable company*". The evaluation force is changed to "higher force" in *more reliable company*. E.g. "MSI a *more reliable company*".

The algorithm changes the force of evaluation to "the highest force" when processing the opinion collocations *very reliable company*, *the most reliable company*, *extremely reliable company*. E.g. "Electa Limited Company is a *very reliable company*".

The orientation is changed to the opposite ("negative orientation") in the following examples: *unreliable company*, *not reliable company* ("low force"), *the most unreliable company* ("the highest force"). E.g. "With regards to security, Websense may be *the most unreliable company*".

The opinion collocation *not reliable enough company* has positive orientation, but the "low force" is weakened. E.g. "Obviously because InPlant is *not a reliable enough company*".

"The highest force" of evaluation is weakened in *not a very reliable company* ("positive orientation"). E.g. "MedZilla is *not a very reliable company* in terms of loyalty to the sales force".

The results of opinion collocations processing are grouped and evaluated to recognize the quality of the opinion-related text. The results are also visualized.

4. CONCLUSION

The proposed ontology-based approach with the use of opinion collocations ensures high accuracy, flexibility for customization and future diverse applications for question-answering systems dealing with opinions and reviews. Opinion collocations are a major factor in the development of both human and machine readable dictionaries for a wide variety of applications including information extraction and information management (retrieval, clustering, categorization).

The further development of the approach envisages the solution of opinion-related problems concerning the analysis of comparative sentences, multiple objects being evaluated (relations expressing similarities or differences of more than one object), the order in which opinion-related sentences are presented, multi-document opinion-oriented summarization.

The proposed ontology-based approach will be enhanced with reasoning modules and effective algorithms for opinion holder detection. Multi-language applications will also be developed.

In future work, natural language processing techniques can be developed aiming at a specific knowledge domain of the document, Opinion Miner could also be trained with respect to the particular knowledge domain to improve the accuracy.

Experimental results, using the proposed ontology-based approach for opinion mining, show that the approach is correct and justified and the technique is highly effective.

6. REFERENCES

[1] Pang B., Lee L. Opinion Mining and Sentiment Analysis. Now Publishers Inc, 2008, p. 148

[2] Devitt A., Ahmad K. Sentiment. Analysis in Financial News: A Cohesion-based Approach. *Proceedings of the Association for Computational Linguistics (ACL' 2007)*, pp. 984–991.

[3] Eguchi K., Lavrenko V. Sentiment retrieval using generative models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP' 2006)*, – pp. 345–354.

[4] Eneko A., Edmonds P. (eds.). Word Sense Disambiguation: Algorithms and Applications. Dordrecht: Springer, 2006, 364 p.

[5] Navigli, R., Velardi, P. Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation, *PAMI (27), No. 7*, July 2005, pp. 1075-1086.

[6] Cabezas C., Resnik P., Stevens J. Supervised Sense Tagging using Support Vector Machines. *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*. Toulouse, France, 2001, pp. 59-62.

[7] Cuong L. A Study of Classifier Combination and Semi-Supervised Learning for Word Sense Disambiguation: PhD Thesis. Japan Advanced Institute of Science and Technology, 2007, 114 p.

[8] McCarthy, Diana, Rob Koeling, Julie Weeds & John Carroll. Finding predominant senses in untagged text. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Barcelona, Spain, 2004, pp. 280–287.