A. Bobkov, S. Gafurov, V. Krasnoproshin, H.Vissia

SEMANTIC-BASED INFORMATION EXTRACTION IN BUZZTALK

Information extraction is gaining much popularity [1; 2]. The field of information extraction is well suited to various types of business and government intelligence applications. Diverse information is of great importance for decision making on products, services, events, persons, organizations.

Creation of systems that can effectively extract meaningful information requires overcoming a number of challenges: identification of documents, knowledge domains, specific opinions, opinion holders, events, activities, as well as representation of the obtained results.

The purpose of this paper is to introduce a system for solving the problem of effective extraction of meaningful information. Semantic patterns approach and an ontology-based approach are proposed as a solution to the problem.

Numerous models and algorithms are proposed for information extraction [3]. The major part of the most useful information is represented as a great variety of texts and documents in a natural language. But the problem of effective information extraction from texts in a natural language still remains unsolved. Processing of texts in a natural language necessitates the solution of the problem of extracting meaningful information. Semantic relations play a major role [4].

In information extraction and text mining, word collocations show a great potential to be useful in many applications (machine translation, natural language processing, lexicography, etc.).

"Collocations" are usually described as "sequences of lexical items which habitually co-occur, but which are nonetheless fully transparent in the sense that each lexical constituent is also a semantic constituent" [5]. Such expressions as "fine weather", "high winds", "cosmetic surgery" are examples of collocations.

The traditional method of performing automatic collocation extraction is to find a formula based on the statistical quantities of words to calculate a score associated to each word pair. Proposed formulas are mainly: "mutual information", "t-test", "z test", "chi-squared test" and "likelihood ratio" [6].

Word collocations from the point of semantic constituents have not yet been widely studied and used for extracting meaningful information, especially when processing texts in a natural language.

The proposed semantic patterns approach is based on word collocations on the semantic level and contextual relations. In general, a semantic pattern includes: 1) *participants* (a person, company, natural/manufactured object, as well as a more abstract entity, such as a plan, policy, etc.) involved in the action or being evaluated; 2) *actions* - a set of verb semantic groups and verbal nouns; 3) *rules for semantic patterns actualization*.

An ontology-based approach is used for semantic patterns actualization [7].

Ontologies have become common on the World-Wide Web ([8]). 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 knowledge domain, the ontology represents the concepts which are held in common by the participants in a particular domain.

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

The proposed approach has been successfully realized in BuzzTalk portal ([9]) for subject domains recognition, opinion mining, mood state detection, event extraction and economic activities detection.

BuzzTalk is offered to companies as a SaaS model (Software as a Service) and it answers questions like:

What is the latest information about my brand, competitors or industry?

What are people writing about my brand, product, organization or CEO?

What are important trends in my industry?

What are the big events inside my industry sector?

Where are my customers/clients located?

When and where are people discussing my brand?

What are the burning world and local issues?

Who is involved in burning issues?

What are the consequences?

BuzzTalk presents a new way of finding content.

The difference between a traditional search engine and a discovery engine such as BuzzTalk, is that search engines list all results for a specific search whereas our engine allows you to monitor topic-specific developments within your search.

BuzzTalk collects all text documents from over 58.000 of the most active websites around the globe, two thirds are news sites and one third are blog sites. The authors of these documents are mainly scientists, journalists and opinion leaders.

BuzzTalk presents a list of articles in chronological order based on publication date. This list grows each day. You can sort and filter this list based on a variety of criteria such as sentiment, mood state, happenings, etc., thus to experience the wealth of real time information without the pain of information overload. For example, you can easily find all publications within your theme that relate to product releases, employment changes, merger & acquisitions and many more.

Below are examples of information extraction in BuzzTalk.

Opinion mining is gaining much popularity within natural language processing [10]. Web reviews, blogs and public articles provide the most essential information for opinion mining. This information is of great importance for decision making on products, services, persons, events, organizations.

The proposed ontology-based approach for semantic patterns actualization was realized in the developed 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).

Opinion words can be expressed by: an adjective (*brilliant, reliable*); 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 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").

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

The developed 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).

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

A valuable addition to opinion mining is detection of individual/public mood states. The relationship between mood states and different human activities has proven a popular area of research [11].

BuzzTalk mood detection uses the classification of the widely-accepted "Profile of Mood States" (POMS), originally developed by McNair, Lorr and Droppleman [12].

In BuzzTalk, mood state detection is based on: 1) mood indicators (e.g. "I feel", "makes me feel", etc.); 2) mood words (e.g. anger, fury, horrified, tired, taken aback, depressed, optimistic); 3) special contextual rules to avoid ambiguation. BuzzTalk automatically recognizes the following mood states: "Anger", "Tension", "Fatigue", "Confusion", "Depression", "Vigor".

Mood state detection alongside with opinion mining can give answers to where we are now and where will be in future.

BuzzTalk detects 233 economic activities from texts in a natural language. The economic activities cover all major activities represented in NACE classification (Statistical Classification of Economic Activities in the European Community), which is similar to the Standard Industrial Classification and North American Industry Classification System.

An example of economic activities detection

General manager Tim Deakin said Orkney Cheddar was produced with locally-sourced milk, following a traditional recipe and process.

Extracted instances:

Economic activity = **Manufacture of Dairy Products**.

BuzzTalk performs real-time extraction of 35 events for decision making in different spheres of business, legal and social activities. The events include: "Environmental Issues", "Natural Disaster", "Health Issues", "Energy Issues",

"Merger & Acquisition", "Company Reorganization", "Competitive Product/ Company", "Money Market", "Product Release", "Bankruptcy", "Bribery & Corruption", "Fraud & Forgery", "Treason", "Hijacking", "Illegal Business", "Sex Abuse", "Conflict", "Conflict Resolution", "Social Life", etc.

For example

A New York woman faced charges for faking cancer to solicit money from unsuspecting donors and a relative.

Extracted instances:

Event = **Fraud & Forgery**

A subject domain is recognized on the basis of a particular set of noun and verb phrases unambiguously describing the domain. For solving the problem of disambiguation special filters, based on the contextual environment (on the level of phrases and the whole text), are introduced.

Subject domains and their concepts are organized hierarchically to state "part-of", "is a kind of" relations.

The proposed semantic patterns approach has been successfully realized in BuzzTalk portal for opinion mining, mood state detection, event extraction and economic activities detection. The approach ensures high accuracy, flexibility for customization and future diverse applications for information extraction.

Semantic word collocations are a major factor in the development of a wide variety of applications including information extraction and information management (retrieval, clustering, categorization, etc.).

Implementation results show that the proposed knowledge-based approach is correct and justified and the technique is highly effective.

REFERENCES

1. *Moens, M.* Information Extraction: Algorithms and Prospects in a Retrieval Context / M. Moens. – Springer, 2006. – 246 p.

2. *Baeza-Yates, R.* Modern Information Retrieval: The Concepts and Technology behind Search / R. Baeza-Yates, B. Ribeiro-Neto. – Addison-Wesley Professional, 2011. – 944 p.

3. *Buettcher, S.* Information Retrieval: Implementing and Evaluating Search Engines / S. Buettcher, C. Clarke, G. Cormack. – MIT Press, 2010. – 632 p.

4. *Khoo, Ch.* Identifying Semantic Relations in Text for Information Retrieval and Information Extraction / Ch. Khoo, H. Myaeng S. – Springer, 2002. – P. 161–180.

5. *Cruse, D. A.* Lexical Semantics / D. A. Cruse. – Cambridge University Press, 1986. – 310 p.

6. *Manning, C. D.* Foundations of statistical natural language processing / C. D. Manning, H. Schütze. – Cambridge, MA : MIT Press, 1999. – 620 p.

7. *Bilan, V.* An Ontology-Based Approach to Opinion Mining / V. Bilan, A. Bobkov, S. Gafurov, V. V. Krasnoproshin, J. van de Laar, H. Vissia // Proceedings of 10-th International Conference PRIP'2009. – Minsk, 2009. – P. 257–259.

8. *Fensel, D.* Foundations for the Web of Information and Services: A Review of 20 Years of Semantic Web Research / D. Fensel. – Springer, 2011. – 416 p.

9. http://www.buzztalkmonitor.com.

10. *Pang, B.* Opinion Mining and Sentiment Analysis / B. Pang, L. Lee. – Now Publishers Inc, 2008. – 148 p.

11. *Clark, A. V.* Mood State and Health / A. V. Clark. – Nova Publishers, 2005. – 213 p.

12. *McNair, D. M.* Profile of Mood States / D. M. McNair, M. Lorr, L. F. Droppleman. – San Diego, Calif. : Educational and Industrial Testing Service, 1971.