An Approach to Web Information Processing

Anatoly Bobkov[1(✉)], Sergey Gafurov[2], Viktor Krasnoproshin[1], and Herman Vissia[2]

[1] Belarusian State University, Minsk, Republic of Belarus
anatoly.bobkov@gmail.com, krasnoproshin@bsu.by
[2] Byelex Multimedia Products BV, Oud Gastel, The Netherlands
sergey_gafurov@by.byelex.com, h.vissia@byelex.com

Abstract. The paper deals with information extraction from the Internet. Special attention is paid to semantic relations.

Keywords: Information extraction · Semantic patterns · Ontology-based approach

1 Introduction

Nowadays, information and data are stored mainly on the Internet. The Internet opens up tremendous opportunities for information extraction that is gaining much popularity [1, 2]. Currently, information extraction from web documents becomes predominant. Information can come from various sources, e.g. media, blogs, personal experiences, books, newspaper and magazine articles, expert opinions, encyclopedias, web pages, etc.

Information extraction comprises methods, algorithms and techniques for finding the desired, relevant information and for storing it in appropriate form for future use.

The field of information extraction is well suited to various types of business, government and social 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 an approach for solving the problem of effective extraction of meaningful, user-oriented information from the web. Semantic patterns approach and an ontology-based approach are proposed as a solution to the problem.

2 Problem Statement and Solution

Numerous models and algorithms are proposed for web information processing and information extraction [3, 4]. But the problem of effective information extraction from texts in a natural language still remains unsolved. Processing of texts in a natural

V.V. Krasnoproshin and S.V. Ablameyko (Eds.): PRIP 2016, CCIS 673, pp. 121–131, 2017.
DOI: 10.1007/978-3-319-54220-1_13

language necessitates the solution of the problem of extracting meaningful information. Semantic relations play a major role [5, 6].

In information extraction and text mining, word collocations show a great potential [7] to be useful in many applications (machine translation, natural language processing, lexicography, word sense disambiguation, 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" [8].

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. The formulas are mainly: "mutual information", "t-test", "z test", "chi-squared test" and "likelihood ratio" [9].

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. Semantic relations (lexical-semantic relations) are meaningful associations between two or more concepts or entities. They can be viewed as links between the concepts or entities that participate in the relation. Associations between concepts can be categorized into different types.

A semantic pattern can be viewed as containing slots that need to be filled. Though most patterns are binary ones having two slots, a pattern may have three or more slots. In general, the proposed semantic patterns include: (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 ("buy", "build", "arrival"); (3) special-purpose rules representing expert knowledge. The patterns cover different types of semantic relations: (1) semantic relations between two concepts/entities, one of which expresses the performance of an operation or process affecting the other ("Much remains to be learned about how nanoparticles affect the environment"); (2) synonymous relationships ("beautiful – attractive - pretty"); (3) antonymy ("wet - dry"); (4) causal relations ("Research identifies new gene that causes osteoporosis"); (5) hyponymous relations ("Jaguar is a powerful vehicle"); (6) locative relations ("Amsterdam is located in the western Netherlands, in the province of North Holland"); (7) part-whole relations ("car transmission - car"); (8) semantic relations in which a concept indicates a time or period of an event designated by another concept ("Second World War, 1939–1945"); (9) associative relations ("baker – bread": "The baker produced bread of excellent quality"); (10) "made-of" relations ("This ring is made of gold"); (11) "made-from" relations ("Cheese made from raw milk imparts different flavors and texture characteristics to the finished cheese"); (12) "used-for" relations ("Database software is used for the management and storage of data and databases"); (13) homonym relations ("bank of the river – bank as a financial institution"), etc. A semantic relation can be expressed in many syntactic forms. Besides words, semantic relations can occur at higher levels of text (between phrases, clauses, sentences and larger text segments), as well as between documents and sets of documents. The variety of semantic relations and their properties play an important role in web

information processing for extracting relevant fragments of information from unstructured text documents.

An ontology-based approach is used for semantic patterns actualization [10].

Ontologies have become common on the World-Wide Web [11]. The broadened interest in ontologies is based on the feature that they provide a machine-processable semantics of information sources that can be communicated among agents as well as between software artifacts and humans. More recently, the notion of ontologies has attracted attention from such fields as intelligent information integration, cooperative information systems, information retrieval, electronic commerce, and knowledge management. For any given knowledge domain, an 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.

3 Implementation of the Proposed Approach

The proposed approach has been successfully realized in BuzzTalk portal [12] for subject domains recognition, opinion mining, mood state detection, event extraction, economic activities detection and named entity recognition.

BuzzTalk is offered to companies as a SaaS (Software as a Service) model and it answers questions like:
What are the burning world and local issues?
Who is involved in burning issues?
What are the consequences?
What is the latest information about my competitors?
What are people writing about my product, organization or CEO?
What are important trends in my industry?
What are the big events inside my industry sector?
Where are my customers located?
When and where are people discussing my brand?
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 BuzzTalk allows you to monitor topic-specific developments within your search. BuzzTalk discovers the latest information about a particular brand, competitors or industry, thus facilitating to make better decisions.

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 finds and links relevant information in natural-language documents while ignoring extraneous, irrelevant information.

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.

3.1 Economic Activities Detection

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 International Standard Industrial Classification of all economic activities (ISIC) reflecting the current structure of the world economy. The classifications provide the internationally accepted standard for categorizing units within an economy. Categories of the classifications have become an accepted way of subdividing the overall economy into useful coherent industries that are widely recognized and used in economic analysis, and as such they have become accepted groupings for data used as indicators of economic activities. The classifications are widely used, both nationally and internationally, in classifying economic activity data in the fields of population, production, employment, gross domestic product and others. They are basic tools for studying economic phenomena, fostering international comparability of data and for promoting the development of sound national statistical systems. The classifications provide a comprehensive framework within which economic data can be collected and reported in a format that is designed for purposes of economic analysis, decision-taking and policy-making.

While extracting and analyzing economic activities, BuzzTalk ensures a continuing flow of information that is indispensable for the monitoring, analysis and evaluation of the performance of an economy over time. Moreover, BuzzTalk facilitates information extraction, presentation and analysis at detailed levels of the economy in an internationally comparable, standardized way.

Examples of economic activities detection:

- *Toyota has maintained its position as the world's biggest car manufacturer.*
  Extracted instances:
  Economic activities = ***Manufacture of motor vehicles*** (NACE code C291)
- *The world's first auto show was held in England in 1895.*
  Extracted instances:
  Economic activities = ***Organisation of conventions and trade shows*** (NACE code N823)
- *Goat cheese has been made for thousands of years, and was probably one of the earliest made dairy products.*
  Extracted instances:
  Economic activities = ***Manufacture of dairy products*** (NACE code C105)

- *This invention relates to a process for the hardening of metals.*
  Extracted instances:
  Economic activities = ***Treatment and coating of metals*** (NACE code C256)
- *India is the largest grower of rice.*
  Extracted instances:
  Economic activities = ***Growing of rice*** (NACE code A0112)
- *OCBC Bank operates its commercial banking business in 15 countries.*
  Extracted instances:
  Economic activities = ***Monetary intermediation*** (NACE code K641)
- *It is even more important to properly plan the preparation of legal documents.*
  Extracted instances:
  Economic activities = ***Legal activities*** (NACE code M691)
- *Doran Polygraph Services specializes in professional certified polygraph testing utilizing the latest equipment and most current software with techniques approved by the American Polygraph Association.*
  Extracted instances:
  Economic activities = ***Security and investigation activities*** (NACE code N80)
- *Florida's aquafarmers grow products for food (fish and shellfish).*
  Extracted instances:
  Economic activities = ***Aquaculture*** (NACE code A032)

3.2 Event Extraction

A specific type of knowledge that can be extracted from texts is an event, which can be represented as a complex combination of relations. Event extraction is beneficial for accurate breaking news analysis, risk analysis, monitoring systems, decision making support systems, etc.

BuzzTalk performs real-time extraction of 35 events, based on lexical-semantic patterns, 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.

Examples:

- *Contract medical research provider, Quintiles, agreed to merge with healthcare information company, IMS Health to make a giant known as Quintiles IMS in an all-stock deal.*
  Extracted instances:
  Event = ***Merger & Acquisition***
- *Mazda Motor Corporation unveiled the all-new Mazda CX-5 crossover SUV.*
  Extracted instances:
  Event = ***Product Release***
- *TCS ranked as top 100 U.S. brand for second consecutive year.*

Extracted instances:
Event = ***Competitive Product/Company***

- *Two Hong Kong men arrested for drug trafficking.*
  Extracted instances:
  Event = ***Illegal Business***
- *A former President of Guatemala, already in jail, has been accused of taking bribes.*
  Extracted instances:
  Event = ***Bribery & Corruption***
- *Yet another green-energy giant faces bankruptcy.*
  Extracted instances:
  Event = ***Bankruptcy***
- *Two Afghans held for attempted rape of woman on Paris train.*
  Extracted instances:
  Event = ***Sex Abuse***

The extracted events play a crucial role in daily decisions taken by people of different professions and occupation.

3.3 Subject Domains Recognition

In BuzzTalk a subject domain is recognized on the basis of a particular set of noun and verb phrases unambiguously describing the domain.

Examples:

- *The Forest Inn Hotel offers hotel accommodation on a weekly basis.*
  Extracted instances:
  Subject domain = ***Travel-Hotel***
- *The goal of the pollution prevention and reduction program is to prevent or minimize polluting discharges.*
  Extracted instances:
  Subject domain = ***Ecology***
- *Mozzarella cheese is a sliceable curd cheese originating in Italy.*
  Extracted instances:
  Subject domain = ***Food***
- *Fresh milk is the common type of milk available in the supermarket.*
  Extracted instances:
  Subject domain = ***Beverage***
- *Distance education includes a range of programs, from elementary and high school to graduate studies.*
  Extracted instances:
  Subject domain = ***Education***
- *The biathlon is a winter sport that combines cross-country skiing and rifle shooting.*
  Extracted instances:
  Subject domain = ***Sport***

- *The aim of nanoelectronics is to process, transmit and store information by taking advantage of properties of matter that are distinctly different from macroscopic properties.*
  Extracted instances:
  Subject domain = ***Innovation***
- *Britain has made a political decision that will have economic effects.*
  Extracted instances:
  Subject domain = ***Politics***
- *Economy from then on meant national economy as a topic for the economic activities of the citizens of a state.*
  Extracted instances:
  Subject domain = ***Economics***
- *The law-making power of the state is the governing power of the state.*
  Extracted instances:
  Subject domain = ***Law***
- *The president called for collective efforts to fight world terrorism.*
  Extracted instances:
  Subject domain = ***Terrorism***
- *Japan was hit by a magnitude 6.5 earthquake followed by an M7.3 quake on Saturday.*
  Extracted instances:
  Subject domain = ***Disaster***

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.

3.4 Named Entity Recognition

Named-entity recognition is a subtask of information extraction that seeks to locate and classify named entities in a text into pre-defined categories such as names of persons, organizations, locations, etc.

BuzzTalk recognizes the following main named entities:

1. Person (first, middle, last names and nicknames, e.g. Steve Jobs, Cristina Fernandez de Kirchner);
2. Title (social, academic titles, etc.);
3. Position (a post of employment/office/job, e.g. president, CEO);
4. Organization (a company, governmental, military or other organizations, e.g. Microsoft, Wells Fargo, The University of Oxford);
5. Location (names of continents, countries, states, provinces, regions, cities, towns, e.g. Africa, The Netherlands, Amsterdam);
6. Technology (technology names or a description of the technology, e.g. 4D printing, advanced driver assistance, affinity chromatography, agricultural robot, airless tire technology);

7. Product (e.g. Sikorsky CH-148 Cyclone, Lockheed Martin F-35 Lightning II, Kalashnikov AKS, Windhoek Lager, Mercedes S550, Apple iPhone 6S Plus, Ultimate Player Edition, Adenosine);
8. Event (a planned public/social/business occasion, e.g. Olympic Summer Games, World Swimming Championship, Paris Air Show, International Book Fair);
9. Industry Term (a term related to a particular industry, e.g. advertising, finance, aviation, automotive, education, film, food, footwear, railway industries);
10. Medical treatment (terms related to the action or manner of treating a patient medically or surgically, e.g. vitamin therapy, vaccination, treatment of cancer, vascular surgery, open heart surgery)

The named entities are hierarchically structured, thus ensuring high precision and recall.

For example:
Organization

- airline company
- automaker
- bank
- football club
- computer manufacturer
- educational institution
- food manufacturer
- apparel manufacturer
- beverage manufacturer

3.5 Opinion Mining

Opinion mining is gaining much popularity within natural language processing [13]. 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) appreciation (e.g. efficient, stable, ideal, worst, highest);
(2) judgment (e.g. decisive, caring, dedicated, intelligent, negligent)

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.

3.6 Mood State Detection

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 [14].

BuzzTalk mood detection uses the classification of the widely-accepted "Profile of Mood States" (POMS), originally developed by McNair, Lorr and Droppleman [15].

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".

Examples:

- *Despite these problems, I feel very happy.*
  Extracted instances:
  Mood state = ***Vigor***
- *I'm feeling angry at the world now.*

Extracted instances:
Mood state = ***Anger***

- *I feel fatigued and exhausted.*
  Extracted instances:
  Mood state = ***Fatigue***
- *I have suicidal thoughts every day.*
  Extracted instances:
  Mood state = ***Depression***

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

4 Conclusion

With the rapid growth of the Internet there is an ever-growing need for reliable multi-functional systems to retrieve relevant and valuable information.

The proposed semantic patterns approach has been successfully realized in BuzzTalk portal for opinion mining, mood state detection, named entity recognition, economic activities detection, subject domain recognition, event extraction. It plays a vital role in information extraction and new knowledge discovery from web documents. The approach ensures high accuracy, flexibility for customization and future diverse applications for information extraction. New semantic relations can be easily created. The relations can be decomposed into simpler relational elements. Semantic relations follow certain general patterns and rules, the same types of semantic relations are used in different languages.

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, p. 246. Springer, Amsterdam (2006)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval: The Concepts and Technology behind Search, p. 944. Addison-Wesley Professional, Harlow (2011)
3. Buettcher, S., Clarke, C., Cormack, G.: Information Retrieval: Implementing and Evaluating Search Engines, p. 632. MIT Press, Cambridge (2010)
4. Machová, K., Bednár, P., Mach, M.: Various approaches to web information processing. Comput. Inf. **26**, 301–327 (2007)
5. Khoo, Ch., Myaeng, S.H.: Identifying semantic relations in text for information retrieval and information extraction. In: Green, R., Bean, C.A., Myaeng, S.H. (eds.) The Semantics of Relationships: An Interdisciplinary Perspective, pp. 161–180. Springer, Amsterdam (2002)

6. Bobkov, A., Gafurov, S., Krasnoproshin, V., Romanchik, V., Vissia, H.: Information extraction based on semantic patterns. In: Proceedings of the 12th International Conference – PRIP 2014, Minsk, pp. 30–35 (2014)
7. Barnbrook, G., Mason, O., Krishnamurthy, R.: Collocation: Applications and Implications, p. 254. Palgrave Macmillan, Basingstoke (2013)
8. Cruse, D.A.: Lexical Semantics, p. 310. Cambridge University Press, Cambridge (1986)
9. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing, p. 620. MIT Press, Cambridge (1999)
10. Bilan, V., Bobkov, A., Gafurov, S., Krasnoproshin, V., van de Laar J., Vissia, H.: An ontology-based approach to opinion mining. In: Proceedings of 10th International Conference PRIP 2009, Minsk, pp. 257–259 (2009)
11. Fensel, D.: Foundations for the Web of Information and Services: A Review of 20 Years of Semantic Web Research, p. 416. Springer, Heidelberg (2011)
12. http://www.buzztalkmonitor.com
13. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis, p. 148. Now Publishers Inc., Hanover (2008)
14. Clark, A.V.: Mood State and Health, p. 213. Nova Publishers, Hauppauge (2005)
15. McNair, D.M., Lorr, M., Droppleman, L.F.: Profile of Mood States. Educational and Industrial Testing Service, San Diego (1971)