Natural Language Processing Based on Semantic Patterns Approach

A. Bobkov
Byelex BV Helium
Oud Gastel, Netherlands
anatoly.bobkov@gmail.com

S. Gafurov
Byelex BV Helium
Oud Gastel, Netherlands
gafurov@gmail.com

V. Krasnoproshin
Faculty of Applied Mathematics
Belarusian State University
Minsk, Belarus
krasnoproshin@bsu.by

H. Vissia
Byelex BV Helium
Oud Gastel, Netherlands
h.vissia@byelex.com

***Abstract*—The paper deals with information extraction from texts in a natural language. Special attention is paid to word collocations on the level of meaning and word sense disambiguation based on semantic patterns.**

Keywords—natural language processing, information extraction, semantic patterns, word collocations, artificial intelligence, ontology-based approach, word sense disambiguation

I. Introduction

Artificial intelligence (AI) is becoming increasingly important in our daily life [1] and continues its rapid development. The future of AI is promising with advancements in machine learning, natural language processing (NLP) and computer vision. AI will enhance various industries and transform the way we live and work. AI has received ever-growing attention in various fields. As a result, it is gradually being applied in industries such as robotics, healthcare, manufacturing, environmental protection, network construction, etc. The technological development of AI goes hand in hand with NLP.

Natural language processing is one of the key elements in artificial intelligence. NLP makes it possible for humans to communicate with machines [2]. This subset of AI enables computers to understand, interpret, and manipulate human language. NLP is one of the fast-growing research domains in AI.

The major part of the most important information can be found in a great variety of texts and documents in a natural language. Information extraction is gaining much popularity within natural language processing [3], [4]. 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, persons, events, organizations.

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

The purpose of this paper is to introduce an approach for effective extraction of meaningful information for solving AI and decision-making problems. Semantic patterns approach is proposed as a solution to the problem.

II. Problem Statement and Solution

Numerous models and algorithms are proposed for information extraction [5]. But the problem of effective information extraction from texts in a natural language still remains unsolved. Processing of texts in a natural language necessitates extraction of meaningful information. Our approach is based on semantic patterns as main constituents for effective information extraction and machine learning. The approach is mainly knowledge-driven thus ensuring extraction of information which is relevant to the topic.

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" [6:40].

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” [7].

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 for solving AI problems and challenges.

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.

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; 2) synonymous relationships; 3) antonymy; 4) causal relations (A is the cause of B); 5) hierarchical subordinate relations; 6) semantic relations between a general concept and individual instances of that concept; 7) semantic relations in which a concept indicates a location of a thing designated by another concept; 8) part-whole relations; 9) semantic relation between two concepts, one of which is affected by or subjected to an operation or process expressed by the other; 10) semantic relations in which a concept indicates a time or period of an event designated by another concept; 11) associative relations, etc. In this way, big data can be processed, information extracted and the meaning of the data determined.

The semantic relations can be considered and represented as an artificial neural network [8]. The main advantage of the proposed approach is that there is no need for training on huge

volumes of data relevant to the topic. Application of the approach is possible for rare/extreme events where data are insufficient to train the model. The approach can provide a clear-cut relationship among interconnected notions and classify complicated relationships.

The proposed approach has a great possibility to know and investigate what is happening, when and where, closely monitor the existing current situation in the world and make predictions.

Of great importance is predictive analytics [9] as an area of big data mining that involves extraction of information and its use to predict events, trends, behavior patterns, etc.

Organizations are turning to predictive analytics to solve difficult problems and uncover new opportunities. Analytical methods can improve crime detection and prevent criminal behavior.

Predictive analytics is used in marketing, financial services, insurance, telecommunications, retail, travel, mobility, healthcare, child protection, pharmaceuticals, capacity planning and other fields.

We consider that extraction and processing of "cause-effect" relations from texts form the basis for predictive analytics. Knowledge of "cause" and "effect" ensures rational decision making and problem solving. It is important in all areas of science and technology.

The semantic patterns approach helps to extract information dealing with "cause-effect" in order to make predictions for decision making. The information could be valuable in many subject areas, including medicine, biology, science, technology, 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.

III. Implementation of the Proposed Semantic Patterns Approach

The proposed approach has been successfully realized in BuzzTalk portal [10] for subject domains recognition, named entity recognition, opinion mining, mood state detection, event extraction and economic activities detection.

BuzzTalk detects, collects and processes big data from over 58.000 of the most active websites around the world. The authors of these documents are well-known writers, journalists and opinion leaders. The total number of monitored websites will continue to grow via Crowd Sourced Learning and manual additions.

BuzzTalk is offered to companies as a SaaS model (Software as a Service) that allows end users to access software applications over the Internet.

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 the user to monitor topic-specific developments within the search.

A. Subject Domains Recognition

A subject domain is recognized on the basis of a particular set of noun, verb phrases and rules 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.

The application recognizes knowledge about the world (more than 80 categories).

In particular:

- "Economics, Finance, Business" (economy type, services, management, market issues, price, economic ecosystems, finance and business issues, etc.);
- "Agricultural Industry" (farming, harvesting, crop production, cultivation technology, etc.);
- "Automotive Industry" (automaker, automobile manufacturing, car systems, self-driving technologies, vehicle safety technologies, autonomous cars, electric car technologies, etc.);
- "Aviation Industry" (aircraft, aircraft manufacturers, air business, aircraft systems, etc.);
- "Arms Industry" (armament, air defense systems, electronic warfare technologies, etc.);
- "Security Industry" (national security, political security, blockchain security, etc.);
- "Sustainable Business" (clean energy, clean transportation, eco-friendly biofuel, etc.);
- "Data Processing Industry" (tokenized data, test mining, natural language processing, etc.);
- "Solar Industry" (crystal silicon solar panel, CSP technologies, solar energy, etc.);
- "Food Industry" (fruits, grapes, vegetable processing; food industry products, etc.);
- "Apparel Industry" (garment industry, beachwear, active wear, evening wear, etc.);
- "Footwear Industry" (footwear, summer footwear, business shoes, combat army boots, etc.);
- "Health" (medical care, treatment, services, staff; disease prevention, etc.);
- "Law" (jurisdiction, legislative norms, agreement, criminal proceedings, crypto crimes, etc.);
- "Politics" (political campaign, conflict, event, system, strategy, regime, sanctions, etc.);
- "Terrorism" (terrorism management, terrorist attack, anti-terrorism preventive measures, etc.);
- "Ecology" (climate change, environment deterioration, biodiversity loss, coastal erosion, etc.);
- "Disaster" (disaster management, natural disaster, man-made disaster, technological disaster, medical disaster, oil spill disaster, space disaster, etc.);
- "Sports" ("Formula One Racing", "Aquatics", "Badminton", "Baseball", "Biathlon", "Boxing", "Cycling", "Equestrian", "Fencing", "Football",

"Golf", "Gymnastics", "Handball", "Hockey", "Judo", "Wrestling", etc.)

B. Named Entity Recognition

BuzzTalk recognizes the following main named entities:

1) "Person" (first, middle, last names and nicknames, e.g. John D. Rockefeller, Bob Dylan);

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, Industrial and Commercial Bank of China Limited, The University of Michigan);

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. Sukhoi Su-57, Lockheed Martin F-35 Lightning II, Kalashnikov AKS, Porsche Cayman GT4 RS, 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, e.g.:

"Organization"

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

The proposed approach helps to understand how entities (persons, organizations, places etc.) relate to each other in a text.

C. Opinion Mining

Opinion mining is gaining much popularity within natural language processing [11]. 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 [12] 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)

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.

D. 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 [13].

BuzzTalk mood detection uses the classification of the widely-accepted "Profile of Mood States" (POMS), originally developed by McNair, Lorr and Droppleman [14].

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

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

E. Event extraction

The developed algorithm performs real-time extraction of 35 events, the recognition of which is vitally important 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.

F. 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 Standard Industrial Classification and North American Industry Classification System. Each of the detected economic activities has a corresponding NACE code.

IV. CONCLUSION

Processing of texts in a natural language necessitates the solution of the problem of extracting meaningful information. Diverse information is of great importance for decision making on products, services, events, persons, organizations. Of great importance is the use of the extracted information for the development of algorithms for predictive analytics with the aim to make predictions about unknown future events. Semantic relations play a major role in solving these problems ensuring interaction with the information in a natural way. Semantic relations ensures tracing of interrelated knowledge. Semantic knowledge modeling can answer diverse questions about persons, their motives and patterns of behavior.

The proposed semantic patterns approach focuses on capturing the meaning of a text. The application analyzes the meanings of the input text and generates meaningful, expressive output. The approach helps to solve AI problems and to improve NLP by automating processes and delivering accurate responses. It helps to solve the problem of word sense disambiguation for effective information extraction.

The knowledge-based approach has been successfully realized in BuzzTalk portal for subject domains recognition, named entity recognition, 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.

Implementation results show that the proposed knowledge-based approach (with statistical methods involved to prevent unwanted results) is correct and justified and the technique is highly effective.

The proposed approach may be improved with reasoning modules to extract more meaningful information from texts in a natural language.

REFERENCES

[1] Russell S., Norvig P. Artificial Intelligence: A Modern Approach. - Pearson, 2020. - 1136 p.

[2] Indurkhya N. (ed.), Fred J. Damerau F. J. (ed.). Handbook of Natural Language Processing. - Chapman & Hall/CRC, 2010. - 702 p.

[3] Moens M. Information Extraction: Algorithms and Prospects in a Retrieval Context. - Springer, 2006. - 246 p.

[4] Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval: The Concepts and Technology behind Search. - Addison-Wesley Professional, 2011. - 944 p.

[5] Buettcher S., Clarke C., Cormack G. Information Retrieval: Implementing and Evaluating Search Engines. - MIT Press, 2010. - 632 p.

[6] Cruse D.A. Lexical Semantics. - Cambridge University Press, 1986. - 310 p.

[7] Manning C. D., Schütze H. Foundations of statistical natural language processing. - Cambridge, MA: MIT Press, 1999. - 620 p.

[8] Laurene V. Fausett. Fundamentals of Neural Networks: Architectures, Algorithms and Applications. – Pearson, 1993. - 480 p.

[9] Siegel E. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. - Wiley, 2013. - 320 p.

[10] [5] “Content Curation and Decision Support - Buzztalk,”, http://www.buzztalkmonitor.com (accessed 05.10.2023).

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

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

[13] Clark A.V. *Mood State and Health.* Nova Publishers, 2005. - 213 p.

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