

Using Internet News Flows as Marketing Data Component

Sergey Orekhov^[0000-0002-5040-5861], Henadii Malyhon^[0000-0001-5448-2488], Irina Liutenko^[0000-0003-4357-1826], Tetiana Goncharenko^[0000-0001-6630-307X]

National Technical University “Kharkiv Polytechnic Institute”,
Kharkiv, Ukraine
osv@kpi.kharkov.ua

Abstract. The theoretical research of Internet news representation as a marketing data component concerning price strategies in polymer market was provided. The research includes some intermediate tasks. Firstly, the classification of marketing objects in news stream was prepared. Secondly, each object (event) was described with ontology including necessary keywords and using syntactic model based on formal grammar. As a result, the models of the following events were developed: competitors, offer, production level, geography, import, world prices, export and others. Finally, the solution (algorithm) was proposed to implement the study. The efficiency of the approach was proved by successful performance of real projects in the polymer market in Ukraine.

Keywords. Marketing data, Internet news, Ontology, Syntactic model, Keywords, Data mining.

1 Problem and Research Goals

The Internet news flow is a sequence of texts with an indication of the publication date [1]. The complexity of extracting data from text-based news arises from the nature of the information presentation through the spoken language. It creates considerable uncertainty in the automatic analysis and makes it difficult to directly extract and proceed to information processing.

Separate news item constitutes a text based on the grammar used in communication. From the point of view of semantic grammars, text-based news can be represented in the form of the following text hierarchy (Fig. 1).

The read Internet news is presented in the form of a surface text (Fig. 1.a). The task of extracting marketing data is the transition from surface to deep text. We assume that the main meaning (marketing information about market events) is concluded in the deep text. The block of intermediate texts depicted in Fig. 1 corresponds to transitions based on the grammar of a language. It describes the rules of transition from the deep text to the surface text (Fig. 1.b).

Copyright © 2020 for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

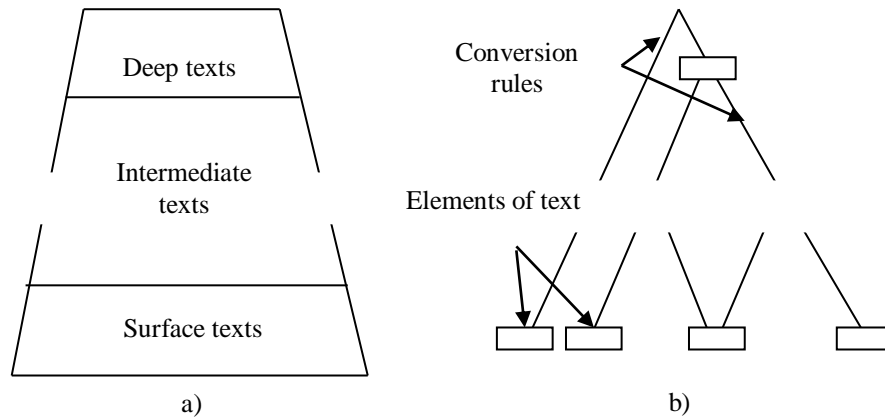


Fig. 1. Model of the text of Internet news

In general, a person's analysis of news passes through the following steps – Fig. 2. At first, a person reads a text and understands that it is news. Then depending on the subject, the corresponding model for data analysis and extraction is being activated. If a person comes across the new data, then his (her) domain model for which the news was received is being updated [1]. It is the data from the updated domain model that is used later in decision making, i.e. news in connection with other domain data.

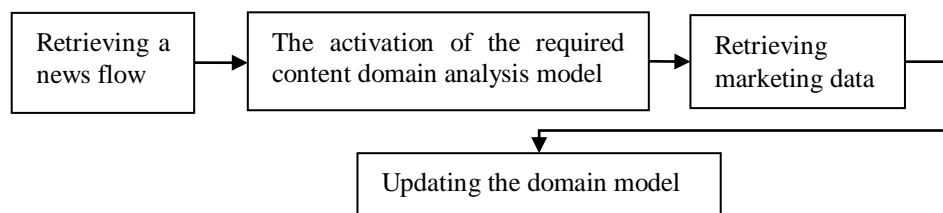


Fig. 2. The stages of news analysis

Existing approaches to extracting data from text-based news are grounded, first of all, on morphological analysis or on the word occurrence frequency. The implementation of this approach involves the formation of a set of lexemes and semantic fields for a particular subject area [1]. Lexem is an abstract set of word-forms for one word. A semantic field is a set of tokens having a common semantic attribute. The quality of

data mining based on morphological analysis is determined by the quality of the lexeme sets and semantic field formation for a particular subject area.

The approach mentioned above is universal. However, it has insufficient accuracy. The reason for the low accuracy of this approach is due to the fact that the data is not stored separately in lexemes, but is also determined by the syntactic structure of the text [2].

2 Problem Statement

The main problems of extracting marketing data from Internet news are:

- **Weak structure.** Internet news is a subjective textual description done by an author about a market event. At the same time there are no generally accepted standards for writing (formats, structures) of news. The news represents a "soup", which includes: a) information about the market event itself, b) "noise". As a result, the information extracted from the news is poorly objective and unreliable. The main risk is related to the price of the error for the decision making based on incorrectly extracted information from the news;
- **Low information gain of news.** To this category, we can relate the news that is not significant. Such news highlights a market event, but its impact on the overall market picture is not significant. As a consequence, when we automatically analyze such Internet news, it may distort the real picture of what is happening. An extreme case of news can be a completely distorted text. As a result, the reliability of the news source is reduced therefore it is necessary to use different sources.

However, when using various sources of news flows, there are problems of a different kind. These problems are associated with the restoration of an overall picture of events, their sequence and importance [3].

The first problem is duplicates. News describing important events will be duplicated in different sources and will characterize one and the same event. The difficulty lies in unambiguous identification of the event.

The second problem is the storylines hidden in news. The news describes only part of the market event, therefore, it is important to analyze not individual news, but some storylines of Internet news.

The simplest solution to the problems described above would be to introduce standardization of the news presentation and description of the event. In the most general case, it could be some artificial language of information transfer - a universal language. However, the complexity associated with the implementation of such a standard, and the absence of monitoring mechanisms, together with little interest in such standardization, make this approach unrealizable. Thus, it is necessary to search for solutions if the language used is unchanged.

To solve the problem of duplicates and storylines, it is necessary to apply machine learning technologies: clustering methods. The application of clustering methods makes it possible to develop an approach for the automatic formation of newsgroups and highlighting of unique events. The main difficulty in this approach is the non-

linear nature of the data in the news, as a consequence, the formation of special criteria for the proximity of two news sites.

To solve the problem of weak news structuring, i.e. tasks of primary information extraction, we analyze the news as a container of marketing information. News is the carrier of information about a market event. According to the rules of journalism the news should have the following structure (Fig. 3).

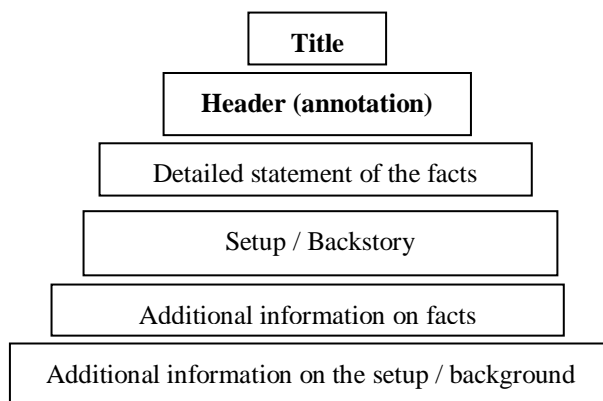


Fig. 3. The structure of Internet news

However, in fact, this news structure is not fully implemented. If it is necessary to get the complete and sufficient information about a market event, then it will be enough to analyze only the top of the hierarchy (Fig. 3): the title and header of the news, which almost always remain.

Header is the beginning of the article. In general, it should be a simple and clear statement about the most important thing. It contains such informational elements: "who", "what", "when", "where", "in what", "how". The same elements can be displayed in the title. The information structured this way will be necessary and sufficient for making a decision.

However, the main problem of such analysis is that news is always a personal statement of the author and characterized by his (her) own style of speech. In addition, the news is not an absolutely abstract medium of information, i.e. in general, as a product, it should be targeted at its consumers-readers, and therefore should attract their attention, as a result the "bare" news information becomes overloaded with different approaches of presentation and psychological tricks to attract attention, making it difficult to extract data event, if not distorting them.

But there is a large enough data layer (detailed information about the event) going after the header. This information would allow identify uniquely the event and its representation in other news. However, there is almost no data structure, which makes the data extraction almost impossible. One of the solutions can be the introduction of additional vocabulary words for syntactic templates of additional explanations about the event and analysis by keywords, carrying out searches for patterns similar to the

search for the header. However, the effectiveness of such an approach is low, due to the lack of structure and unpredictable logics of presentation.

Thus, automatic retrieval of marketing data from Internet news is possible. The proposed approach on the basis of transitions from surface text to depth (using models of analysis of subject areas for news) allows obtaining necessary and sufficient data of high quality about a market event, that is, to identify an external factor that is likely to affect the price dynamics in a given segment market in a given period of time.

Despite the poor structuring of information in the news, it will be enough to use title and header as the main data source, which greatly simplifies the task. To implement the approach, it is necessary to study the rules of the transformation grammar from deep to surface texts for the model analysis for specific subject areas.

Obviously, the primary task is to extract the elements of header and title. The most common approach is the use of linguistic models of text, which together with vocabularies of a specific subject area will be able to give a sufficiently high quality due to the attraction of marketing information.

3 Task Solution

Let us consider an approach to retrieving marketing data from text of Internet news with the use of parsing. We assume that the morphological analysis has already been carried out. Under the extraction of data from the text news is understood the definition of a type of a market event and information characterizing it.

Most of the existing approaches for data extracting from texts do not involve parsing. They provide morphological analysis based on the specific gravity of the word in the text. For such approaches, a large inaccuracy in the extracted data is typical, since one and the same lexeme can enter different sets of syntactic fields, and the same syntactic field, in turn, can enter different sets of fields of different types of events. Besides, the information stored in the syntactic structure of the text is lost. In general, the use of parsing for textual information is complicated by the fact that the potential set of syntactic models is too large and varied.

However, the amount of syntactic models necessary for the analysis of news texts can be significantly reduced by narrowing the subject area. Narrowing is possible by forming a set of syntactic models for analyzing news titles and headers. We assume that they together contain 90% of information on a market event. This is a necessary and sufficient condition for further analysis [3].

The development of a set of syntactic models for the analysis of news, as a preparatory stage, requires the identification of types of events and their characteristic parameters. It is achieved through the development of ontology of a specific subject area (Fig. 4), i.e. complex description of the set of objects and connections between them.

The development of the ontology [4] of a particular subject area is an obligatory stage of analysis and allows the formation of sets of semantic fields, their corresponding lexemes and semantic models.

Ontology (Fig. 4) shows that a market event includes data about the agents involved in it, the date of the event, the market, the product and the type of co-existence. The type of event can correspond to one of the categories such as demand and supply, consumer psychology, consumer profile and competitiveness, consumer value assessment, inflation, world prices, taxes, legislation, antimonopoly actions, import, export, scientific and technological development.

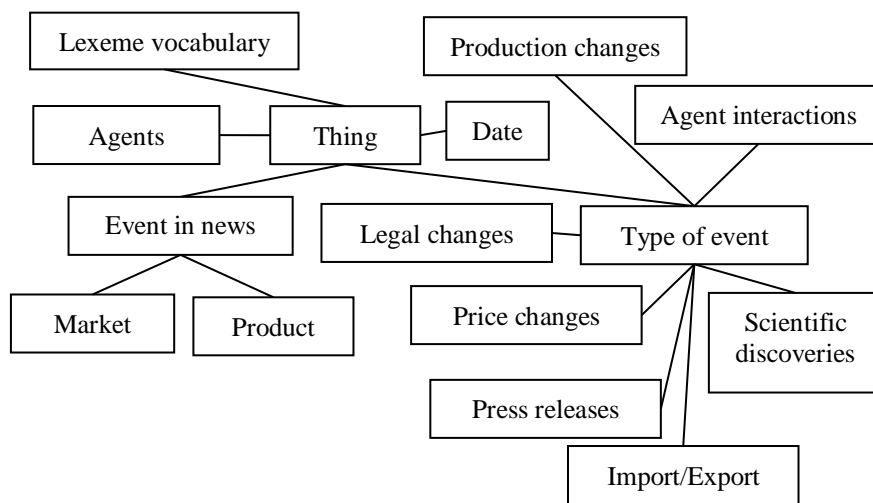


Fig. 4. Ontology of events in the polymer market

On the basis of the ontology, it is proposed to form a set of syntactic models of Internet news corresponding to the specific types of events. Using the methodology of phrase-structure grammar (PSG), it is possible to describe a similar model in the polymer market as an example (Fig. 5) [4].

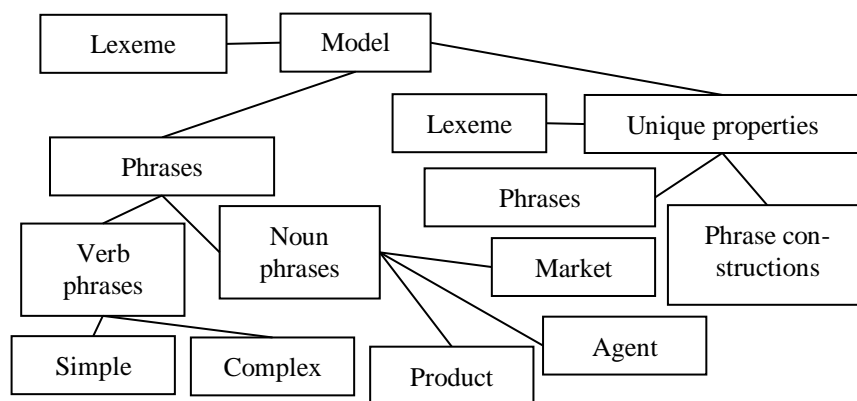


Fig. 5. Polymer Market Event Analysis Model

The model includes the sets of syntactic models of phrases that are divided into verb phrases, noun phrases, as well as many lexemes based on morphological analysis, and additional unique attributes such as events.

Analysis based on syntactic models implies the formation of a set of syntactic phrases for each type of event. The data mining from news on the basis of syntactic models, within the methodology of structure-phrase grammar, implies the determination of the correct construction of phrases by means of the corresponding sets of tokens within the semantic field of the type of event, allowing a priori retrieval of data.

Thus, the general approach to the production of marketing data from the Internet news flow consists of two key stages: the construction of the ontology of the domain area and the formation of syntactic models on its basis for each type of event, highlighting their unique characteristics.

Then, in general, the algorithm of automatic extraction of marketing data from the news will take the form:

- Null hypothesis about the type of event;
- Search for tokens included in the semantic field of the event type;
- An attempt to identify syntactic models;
- Check the text of the news for compliance with unique features;
- Confirm or refute the null hypothesis;
- Extract data when the null hypothesis is fulfilled.

The main advantage of the proposed approach is that the quality of extracting marketing data is significantly increased. However, the syntactical models of the news are entirely dependent on expert opinion [3].

Consider the implementation of this approach in the form of algorithm and a set of syntactic models.

4 Solution

Syntactical and linguistic analysis of Internet news is based on the use of artificial grammars. Using only rules of traditional grammars is ineffective for automatic classification of news due to the inappropriateness of these rules for automatic processing.

An important aspect of traditional grammars, which are of great importance for automatic processing, is a hierarchical interpretation of the structure of the sentence, which is expressed in the form of all sorts of subordinate connections and the allocation of the main members of the sentence.

Among the artificial grammars, grammars are distinguished by functional dependency and grammars of phrase structure. None of the grammars described above is a universal alternative to traditional grammar. The syntactic-linguistic analysis proposed in the paper is based on dependency grammars and phrase structure grammars.

Grammar of dependencies operates with taxonomic units, identifying taxonomic units with functional ones. The nature of all connections is considered to be subordinate. As a vertex of the syntactic tree, a verb-predicate is taken here. Service words for nouns are recognized as managers in relation to subordinate nouns.

The grammar of phrase structure operates with units called "constituent element". As constituent element we consider each of the two maximal volume constructions, which can be identified as a part of the sentence and further within each "constituent element" (CE). The atomic unit in the analysis is not a word, but a morpheme or a lexeme. The links in the grammar are non-directional, because syntactic structure is being established here by successive linear division, and not by clarifying the syntactic hierarchy. The constituent elements are defined in terms of grammatical classes: NP (Noun-phrase) is a nominal component, VP (Verb-phrase) is a verbal component, N (Noun) is a noun, V (Verb) is a verb, Aux (Auxiliary) - a service word, usually a service verb, Adj (Adjective) - an adjective, Prep (Preposition) - a preposition, etc [5-8].

Graphical representation of a sentence in terms of the CE for the polymer market is shown in Fig. 6.

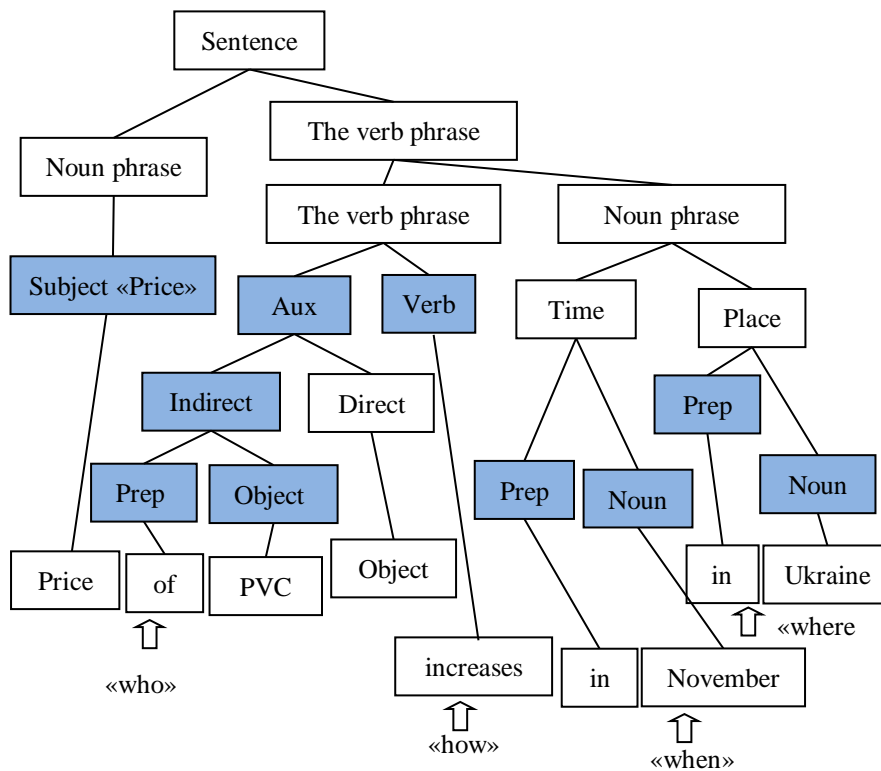


Fig. 6. Graphical representation of a sentence in terms of phrase structure grammar

To extract data from the entire news flow, similar models (grammar rules) must be formed for the header of each category of news.

The main advantage of the phrase structure grammar is that it operates dynamically with the structure of the sentence. It allows analyzing the structure by studying the

sentence consistently in the form of steps to decompose the sentence into its components.

Therefore, for the highest quality of text news analysis, an approach is proposed that includes both grammar dependency elements and phrase structure.

Thus, the proposed approach to the automatic processing and analysis of news objects is based on the formation of sets of units of PSG.

The formation of sets of units of PSG relies on the word-list of the lexemes of the analyzed subject area, and, accordingly, the primary task is to study the multitude of lexemes of the title and the news header.

News as a component of marketing information contains data describing the event and is expressed as "who", "what", "where", "when", "why", "how". Each of the elements is associated with a set of units of grammar. A unique set of units of grammar is a syntactic model of the news text that characterizes a specific type of event.

As an example of the formation of units of the grammar, we have considered the elements of the news and the forming lexemes on the example of the polymer market. The block "when" has two time parameters: the date of news publication and verbal description of the starting point of the event. For example, news published 21.10.2010 "in October price decreases", i.e. 21.10.2010 is the final date and the date of the event beginning is 01.10.2010. The block "where" is sometimes indicated explicitly, for example "in Ukraine, in USA, in Europe", and sometimes hidden in "who", in case of news about manufacturers.

Thus, it is necessary to have a database of producers and location (geography) of their capacities in the markets. The block "who" maybe either the name of the polymer, or the manufacturer, or a specific polymer model of the manufacturer, the rest of the news is usually irrelevant and can be excluded. Another feature of this block can be the enumeration of several types of polymers in conjunction with the names of producers, which can cause problems in automatic processing. The block "how" can be described as verbally "prices have decreased strongly, insignificantly», as in a numerical form "imports decreased by 90 thousand tons, prices grew by 5%". The block "what" describes the action that took place over "who", literary - the verb-action. For specific news special dictionaries are compiled.

So, for example, for the polymer market we will select the following dictionaries:

- "who" block is a dictionary of polymer types (abbreviations and full names), grades of polymers, names of producers;
- "how" block is a dictionary of the verbal description of the nature of the quantitative change;
- "when" block is a dictionary of time measures (week, month, quarter, year) and directly the names of months, seasons.

The peculiarity of the block "why" is connected with the fact that the news often displays a change in markets, the cause of which is not always known, and if there is an explanation of the reasons, then it is usually a hypothesis. Therefore, the "why" block is excluded from the model without loss of informativeness.

However, the separate use of grammar units is insufficient, since the subordination of various elements of the sentence plays an important role in the text analysis. Using

the rules of grammatical dependencies, we can overcome this difficulty. The construction SP in a sentence always consists of two equal elements of the verb phrase VP of the predicate and the phrase SP of the subject (noun). A unique for each category of news is a verbal phrase, a noun phrase for subject and additions can always be determined provided the verb phrase is identified.

Further analysis and allocation of the grammar units is carried out depending on the results of VP and SP identification in accordance with the following mechanism.

Let us denote the hypothesis H0 as a well-chosen syntactic model of news, with the help of which it is possible to extract data. H1 means that the model is not correct and data can not be extracted.

Analysis of the news will look like this: based on verb-action (predicate), a hypothesis is advanced about the classification of news and the corresponding analysis models are used. The use of other members of the sentences is less effective because the objects and subjects of the impact are the same in the news flow. The algorithm for analyzing the incoming Internet news is presented in Fig. 7. Let us consider the example of syntactic models for one of the main categories for the polymer market [9-10].

5 Results

Using mentioned above approach the six syntactic models were developed. In the paper the syntactic model of news of category "competitor's profile" was presented. This category describes competitors and their pricing policy.

To this category of new it is necessary to carry that news in which such market events are displayed as: the conclusion of agreements on cooperation; sale, purchase of enterprises; mergers of enterprises. The syntax model is presented in general form in Fig. 8.

A simple verbal phrase (predicate) of the category "competitor profile", consisting of a single verb, is shown in Fig. 9.a.

The verb varies according to: a) birth: mr, ms.; b) time: the present, the past, the future; c) number: single, plural.

A compound verb phrase can include two verbs, a verb with a preposition, their combinations (Fig. 9).

The case of a compound verb phrase from two verbs includes a verb-bunch and a verb that characterize: the nature of the impact, usually intentions ("are going to build"); time of the event ("start to produce").

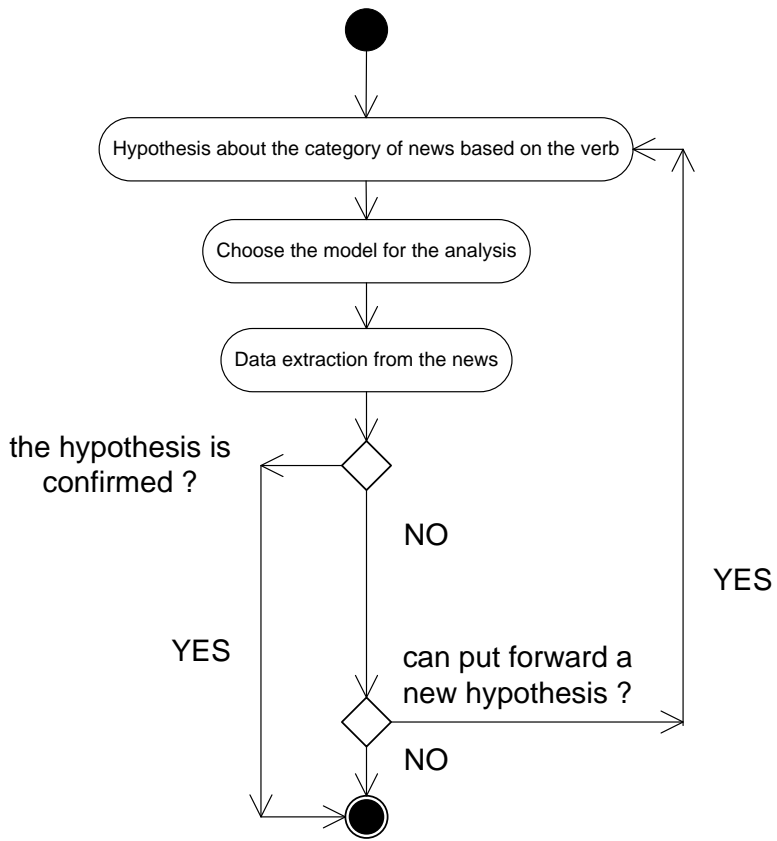


Fig. 7. Algorithm for analyzing incoming Internet news

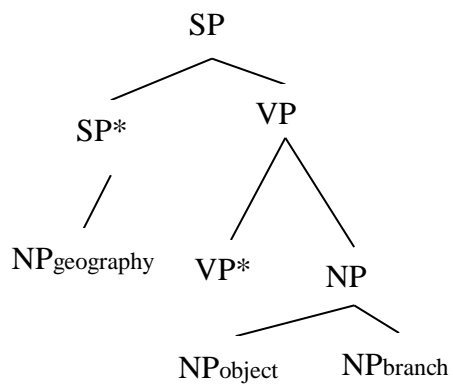


Fig. 8. Syntactic model of the news "profile of the competitor"

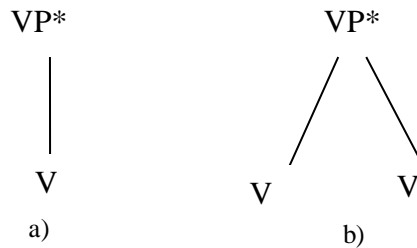


Fig. 9. Simple verb phrase

In the general case, the preposition can be neglected in the verbal phrase, however, it is necessary to take into account the cases for which the pretext is a signal that it is impossible to identify the impact object ("N1 and N2 are feeling in the formation of a new holding", as shown in Fig. 10.c). For a more accurate classification, it is possible to consider a verbal phrase as a phrase from a verb and a nested noun phrase, where the verb derivative is used as a noun, the dictionary of such nouns includes words denoting cooperation, actions of subjects of entrepreneurial activity [5-8].

Thus, the identification of the predicate includes (Fig. 11).

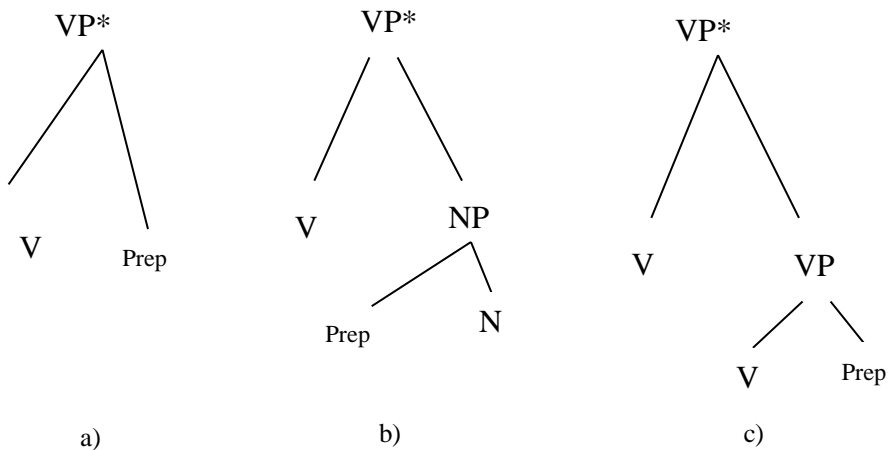


Fig. 10. Complex verb phrase

A simple phrase of the noun subject of the action of the verbal phrase, consisting of one noun, is shown in Fig. 12.

Nouns vary according to: a) childbirth (male, female) and b) number (single, plural). Nouns that are part of the vocabulary of tokens are NP objects that are unique to this category of news and can serve to confirm the hypothesis of a news category. The compound phrase of the substantive object of the action, consisting of an adjective and a noun, is shown in Fig. 13.

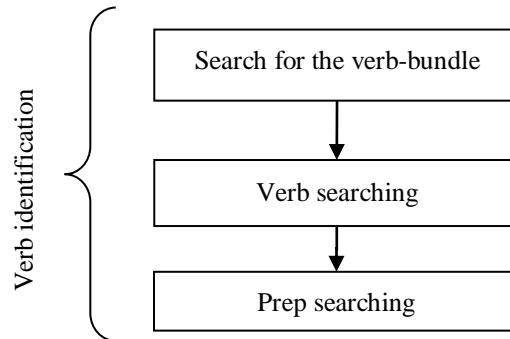


Fig. 11. Verb identification

Nouns used in compound phrases of the noun object of influence are included in the lexeme of the dictionary of simple phrases of the noun object of influence. At the same time, the dictionary of adjectives can be unique. Such a tuple can serve to confirm the hypothesis of this category of news. An example: "will open joint production". Also, the adjective may reflect: a) information on the place of occurrence (geography) of the event ("create a European unit"); b) joint changes in capacity ("will open a new plant in Ukraine").

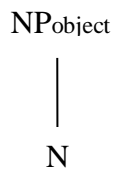


Fig. 12. Simple noun phrase

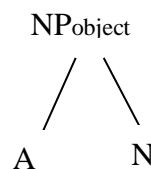


Fig. 13. Complex noun phrase

In this case, it is impossible to confirm the hypothesis for the category of news. The phrase of the noun geography of the event NP geography implements the noun phrase SP for this category of news and stores information about the place of occurrence of the event.

A simple phrase from a single noun is shown in Fig. 14. Nouns that are part of the vocabulary of the lexeme of this phrase usually describe the names of firms. They also express the geography of a market event in a latent form.

The compound phrase of the noun geography of this category of news includes cases of two or more nouns, an adjective with a noun (Fig. 15–16).

The case of two nouns is the enumeration of subjects of interaction in the news, i.e. the transfer of firm names. When identifying such a phrase, in addition to the dictionary, the names of firm names for identification are punctuation, comma, and the conjunction "and" between nouns. Similar constructions can be used to confirm the hypothesis about the category of news.

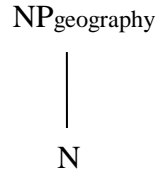


Fig. 14. Simple noun phrase

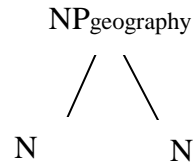


Fig. 15. Compound phrase noun: two nouns

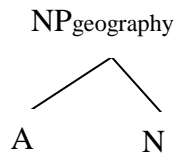


Fig. 16. Compound phrase noun: noun and adjective

It should be noted that the noun in this phrase, which is not a proper name, expresses the subject of the proposal and does not contain information about the place of occurrence of the event, information about which is contained in the adjective phrase: "Ukrainian companies" (Fig. 17). The lexicon of adjectives contains geographic names. The dictionary of lexemes of nouns includes words denoting subjects of entrepreneurial activity in the plural. The presence of such a tuple in front of the verbal phrase of the predicate can provide confirmation for the hypothesis about the category of news.

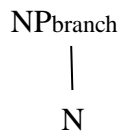


Fig. 17. Simple noun phrase

The phrase of the noun branch (market), to which the NP events relates, can not be used to test the hypothesis. A simple phrase and a single noun is shown in Fig. 17. In the lexicon of the nouns of the given noun phrase: a) product names, incl. abbreviations ("PVC", "ethylene"); b) names of own firms.

There is also a case when the industry is given in an implicit form with the help of a compound phrase of the noun.

The case of the adjective with the noun is given in Fig. 18.a, the dictionary of the noun's lexeme includes words denoting, in general, an object, an article. An adjective characterizes the product and serves as an identifier for the polymer industry. The case of a phrase formed from two or more nouns is shown in Fig. 19.b. Here is the enumeration of the nouns of a simple phrase.

Separately, we should highlight the case of inaccurate categorization when the industry is specified, but the noun of the branch (Fig. 18.c) N1 can not be identified for

any reason. In this case, the noun preceding the branch noun must be subjected to the analysis, i.e. to find a phrase consisting of a preposition and a noun, where the lexeme of prepositions is a pretext, and the lexeme of nouns includes words denoting market processes ("in the market," "in production," in "creation").

Another five models have the same structure [9-10].

Using presented model, the set of news was processed. As the result we have received the series of tags which fully describe the marketing data hidden in the news (Fig. 19).

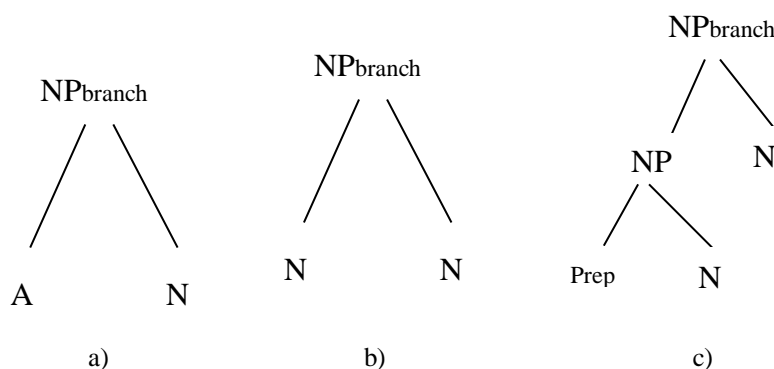


Fig. 18. Compound phrase noun branch

ADNOC to supply Malaysian Lotte Chemical with naphtha in 3-year deal

January 24/2018

MOSCOW (MRC) -- Abu Dhabi National Oil Co said on Tuesday it had signed a three-year agreement with Lotte Chemical Titan, one of the largest polyolefin producers in southeast Asia, to sell the Malaysian firm up to 1 million tonnes of naphtha annually, reported Reuters.

"With this agreement we are implementing a new approach toward our sales of naphtha," said Abdulla Salem al-Dhaheri, ADNOC's director for marketing, sales and trading.

"Previously we have sold the product on shorter-term, one-year contracts. By switching to a three-year contract we are capturing long-term market access and securing offtake."

ADNOC produces over 12 million tonnes of naphtha annually.

As MRC informed before, Abu Dhabi National Oil Company (Adnoc) is planning to increase its crude refining capacity in UAE by at least 60% and triple its petrochemical production as part of its future growth strategy, its chief executive officer said in Abu Dhabi last week. "This proposed expansion will create a single largest integrated refining and chemical site in the world in Ruwais in the UAE," said then Dr Sultan Ahmad Al Jaber speaking at the Atlantic Council Global Energy Forum. "Once complete we will convert almost 20 per cent of our crude into chemicals, diversifying our range of higher value products and providing a natural hedge to oil price movements." He also said hydrocarbons will continue to play a vital role to meet global energy demand despite increase in the diversity of energy mix.

Author: Margaret Volkova

Tags: PP, PE, crude and gaz condensate, petrochemistry, ADNOC, Honam Petrochemical, Lotte Chemical, United Arab Emirates (UAE), South Korea.

Category: General News



Fig. 19. The example of Internet news in polymer market

Next step of the research is building the associative rules, which are used for price forecasting, price strategy development, competitor analysis etc.

6 Conclusion

The performed scientific work allowed us to get some interesting results:

1. In the paper the presentation of the Internet news as a component of marketing information. Based on this presentation, a new approach to the automatic retrieving of marketing data in the Internet news flow has been developed, which makes possible to implement the components of marketing information system in practice with the application of a new data source (Internet news).
2. To implement the proposed approach, a series of syntactic models of Internet news for different categories: competitor's profile, proposal, import, export, world prices and scientific and technical development.
3. For the primary classification of news flows, it is necessary to use the morphological-syntactic models of grammars for phrase structure. Using the developed 6 groups of news models allows identifying up to 85% of news in the stream.

References

1. Liu B. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, USA (2011)
2. Gazdar G. *Phrase Structure Grammar*. In: Jacobson P., Pullum G. (eds) *The Nature of Syntactic Representation*. Synthese Language Library (Texts and Studies in Linguistics and Philosophy), vol 15, 131-186, Springer, Dordrecht (1982)
3. Mittermayer M.-A. *Text Mining Systems for Market Response to News: A Survey*. Institute of Information Systems University of Bern, Working Paper No. 184. Switzerland (2006)
4. Noy N. *Ontology Development 101: A Guide to Creating Your First Ontology*. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880. USA, (2001)
5. Gazdar, G., Klein E., Pullum G., and Sag I. *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge (1985)
6. Polard C., Sag I. *Head-driven Phrase Structure Grammar*. Stanford CA (1994)
7. Georgia M. Green and Robert D. Levine, *Introduction to Studies in Contemporary Phrase Structure Grammar*. Cambridge University Press, (1999)
8. Borsley R. *Modern Phrase Structure Grammar*. Wiley-Blackwell (1996)
9. Cherenkov I, Orekhov S. News data mining based on example of polymer market. *Information Processing Systems*, Vol. 9 (107), 224-227 (2012)
10. Cherenkov I, Orekhov S. Approach for extracting events from news stream. *Eastern-European Journal of Enterprise Technologies*, Vol. 1, No. 4 (61), 62-64 (2013)