Issues of Fact-based Information Analysis

Natalia Sharonova, Anastasiia Doroshenko, Olga Cherednichenko

National Technical University "Kharkiv Polytechnic Institute",
2, Kyrpychova str., 61002 Kharkiv, Ukraine
nvsharonova@ukr.net, olha.cherednichenko@gmail.com

Abstract. With the recent growth of Internet, mobile and social networks the spread of fake news and click-baits increases drastically. Today, the fact retrieval system is one of the most effective tools for identifying the information for decision-making. We propose the approach based on factual information systematization. Different interpretations of the same phenomenon, as well as the inconsistency, inaccuracy or mismatch in information coming from different sources, lead to the task of factual information extraction. In this work, we explore how can natural language processing methods help to check contradictions and mismatches in facts automatically. The reference model of the fact-based analytical system is proposed. It consists of such basic components as Document Search component, Fact retrieval component, Fact Analysis component, Visualization component, and Control component.

Keywords: Fact, Natural Language Processing, Information extraction, Comparator identification, Predicate, Reference model

CEUR-WS.org/Vol-2136/10000011.pdf

1 Introduction

The access to the Internet, as well as to social networks has been simplified for the last decade. It has led to information flow growth. Social networking sites give an opportunity for users to share content freely. As a consequence, fake news, hoaxes, and click-baits are spread, circulated, consumed and shared without critical thinking or fact checking. Regardless the form, the reverberations of inaccurate or misleading information could lead to major risks for the society.

Misinformation can be spread both intentionally and accidentally. Among the interested stakeholders of fake information are politicians, marketing managers, sellers, and users with unclear purposes. In media politicians use alternative facts and post-truths in order to manipulate their audience`s opinions creating thus long-term sustainable mindsets. Alternative facts are information with no basis in reality while post-truth technics are defined as beyond the truth or irrelevant information [1].

As for social networking sites, there are several sources of fake information search. Users' profiles contain a lot of misinformation. In many instances it is difficult to match pages of the same user at different social networks because of a clash in personal information. As well as that it is impossible to assess reliable posts presented in the newsfeed.

E-commerce is another sphere where fake information is common. Sellers at market places present the name and the description of products. The same product can be described in different ways by different sellers. Apart from that, sellers, in order to present their products at as many search requests as possible sellers use clashing information.

Thus, the problem of identifying and verifying contradictory or ambiguous information is crucial. The key idea of our work is to develop an approach for checking contradictions and mismatch in facts automatically.

2 Related Works

In recent years, the reliability of information on the Internet has decreased significantly. It is particularly noticeable on social networks, where distorted, inaccurate or false information reaches and affects millions of users within minutes. Therefore the problem of fake information detection has become a popular research sphere. There are several challenges to automatic detection of fake news: determining if the facts in the news article are correct; analyzing the relations between the article headline and article body; estimating the inherent bias of a written text etc. Factual analysis of the text is designed to make possible the intellectual analysis of data extracted from the text flow. The solution of this task should lead to a synergistic effect, to the possibility of using existing information technologies.

The paper [1] presents the analysis of hoax medical news in social media is presented. The stance classification is implemented in hoax analysis particularly with media contents. An interesting framework has been developed to crosscheck claims against fact-checks. In order to check the news in social networks, one important concept emerges in the paper [2]. In each post in the newsfeed, the “fact” should be identified. However, in many situations, it is impossible to identify whether some part of the information is a fact or not. Each fact is composed of something that has happened at some time, somewhere, possibly to someone.

The work [3] proposes an infrastructure to address phenomena of modern online media production, circulation, and manipulation by establishing a distributed architecture for automatic processing and human feedback. A hybrid technology infrastructure that provides user- and machine-generated annotations on top of the whole World Wide Web is proposed. The ultimate goal of the proposed approach is enable internet users to handle fake news and other online media phenomena by providing both automatic assessments of content and by including alternative opinions into the process of media consumption. The paper [4] discusses the role of computational social scientists in the fight against digital misinformation. Clarify the fundamental mechanisms that make us vulnerable to misinformation online, as well as devise effective strategies to counteract misinformation. There is a growing interest in automating the various activities that revolve around fact-checking. The fact-checking automating includes newsgathering, verification and delivery of corrections.

The need for automatic hoax detection systems is a vital task. In the paper [5] they develop an approach which allows classifying posts in a social network with high

accuracy as hoaxes or non-hoaxes on the basis of the users who "liked" them. Two classification techniques are presented. One technique is based on logistic regression, and the other one is based on a novel adaptation of Boolean crowdsourcing algorithms. It was proved that both techniques are robust: they work even when the users' attention is limited to the users who like both hoax and non-hoax posts. These results suggest that mapping the diffusion pattern of information can be a useful component of automatic hoax detection systems.

The goal of the paper [6] is to present a description of UCL Machine Reading's model employed during fake news detection. The presented stance detection model is a single, end-to-end system consisting of lexical and similar features fed through a multi-layer perceptron with one hidden layer. Being relatively simple in nature, the model performs on par with more elaborate, ensemble-based systems of other teams.

In the paper [7] they analyze the link between the article headline and the article body in order to detect whether the presented news is fake or not. Several neural network architectures were explored for stance detection in news articles. The attention-based models, in particular, a variation of the Attentive Reader Model (ARM2) work properly for this task. The given model evaluates each prediction in a two-step process. The first step is to compare the headline and the body and classify as related or unrelated. The second step is to classify related head-body combinations as agrees, disagrees, or discusses.

The research [8] is oriented on fake news detection. "Fake news detection" is defined as the task of categorizing news. The paper discusses a typology of several varieties of veracity assessment methods emerging from two major categories – linguistic cue approaches and network analysis approaches. The paper [9] aims to enable the identification of deliberately deceptive information in text-based online news. Proposed system can alert users to deceptive news in the incoming news stream and prompt users to further fact-check suspicious instances. It is an information system support applied a vector space model to cluster the news.

There are number approaches of information extraction from natural languages texts. We can highlight lack of automated semantic understanding and low consistency of extracted facts. Despite of existing data extraction solutions the task of extracting facts still is not solved.

3 Methodology

Today, the fact retrieval system is one of the most effective tools for identifying the information for decision-making. When you refer to something as a fact you mean that you think it is true or correct. Factual information is information based on facts or relating to the facts. The reliability of automatically extracted facts is the main problem of processing factual information. It is especially important because of increasing density of text information flow in mass media and various social networks, forums and blogs. Different interpretations of the same phenomenon, as well as the inconsistency, inaccuracy or mismatch in information coming from different sources lead to the task of factual information extraction.

We can consider facts as structured objects. This record describes real-world entity with its attributes mentioned in text, usually, who did what to whom, where and when. So, the fact can be extracted from the textual information and can determine the attributes of the object or the relations between objects. The task of Information Extraction is to identify instances, relations, events and their relevant properties in natural language texts. We consider two types of facts. They can be described as triplets. The first kind of fact is a "Subject -> Relation -> Object", where the subject is who acts, the relation defines action with the object. The second kind of fact is a triplet: "Object-Attribute-Value", where the object is the entity about which the fact is fixed. The attribute is predetermined characteristic that identifies the object with the certain values.

The extraction of facts from weakly structured textual information includes the following steps:

1) Entity Extraction – extract words or phrases that are important for describing the meaning of the text (lists of terms of the subject domain, personalities, organizations, geographical names, etc.);

2) Feature Association Extraction is searching the links between the entities extracted;

3) Event and Fact Extraction is extraction of entities, recognition of facts and actions.

To implement the entity extraction, a standard linguistic processor is used. The issue is the extraction of information about the relations between entities. For this purpose we need to define a certain template that reflects the semantic links in the sentence. Based on the fact definition, it is possible to define the minimal semantic unit of factual search, which is a triad: agent-predicate-value. That is, the record of factual information must include a pointer to the fact search agent, the attribute or predicate of this object, and give a specific value of this attribute.

Such a definition makes it possible to extract concepts from weakly structured text sources of information and to represent relations between them in a structured way. The resulting structure is facts, both in the form of fairly simple concepts: keywords, personalities, organizations, geographical names, and in a more complex form, for example, the name of the person with her job and occupation.

Algebra of finite predicates is used as a mathematical tool for describing discrete, determinate and finite objects or processes from real world [10, 11]. We use this math scheme to represent knowledge extracted from natural language texts:

- text information objects;
- the entity of the subject domain,
- grammatical and semantic characteristics of the text units.

An analysis showed that the most natural and convenient tool for modeling natural language relations is the algebra of finite predicates that operates with letter variables [10, 12]. This tool meets all the requirements for linguistic formalisms. In this case, all kinds of morphological processing lead to the solution of algebraic equations with different initial data (fully or partially specified). Having an algorithm for solving these equations, the formalization of various processes of word processing can be greatly simplified. The possibility of equivalent transformations and minimization of

the morphological model are available in this approach. Besides that the commonality of expressive means makes it convenient to analyze different fragments of the model.

Mathematical relation is the basic concept of logical mathematics. A logical network is a processor and it performs various actions on relationships [10]. Relations express the attributes of objects and the connections between them. They are a universal means of describing any objects. The human language, as a means of communication, is only a means for expressing relationships. Speaking to other people, we convey to them the meaning of the sentence, which is an attitude. The exchange of thoughts between people is carried out only through the transfer of relations. Each thought represents some relation. Perceiving objects and events of the external world, we get information about them in the form of relationships.

Any relation can be interpreted meaningfully as knowledge about the fact, expressed by some utterance. The fact is an exhaustive description of the actual state of all places interested to us. Knowledge of the fact only limits the many possible states of places. A statement about a fact can be true or false. It is true if the characterizing relation contains the actual set of place states and otherwise is false.

4 Our approach

The model of fact extraction from natural language text can be presented in the following way. The fact is considered as a triplet: “Subject -> Predicate -> Object”. The predicate defines a relation, and the subject and object defines two entities. In the developed model we introduce a set of grammatical characteristics of the sentence words. To represent the triplet of fact we use approach proposed in [11, 12].

We suggest the model of facts extracting based on the method of comparator identification [10]. It allows matching the data and the template. It is based on the relation between the words and the placement of these words in the text. This method represents the extraction process as a human intelligent activity since a human looking through a text can easily determine whether it corresponds to the template or not and catch attributes of a fact.

We discover that the descriptions of the same commodities in the trading platforms can be presented in a different way. We notice that such description can be presented as a triplet “object-attribute-value”. So, we can consider the second kind of facts. The description of commodity is represented as number of words; usually it is not a sentence, and a table with some characteristics of commodity.

Let E – be the set of structural elements of a web page, W – the set of words. Then $R_{SEARCH} \subseteq E \times W$ – is the binary relation "is used for search". Let $E_q \subseteq E$ – the set of elements of the web page that are selected for estimation and $W_q \subseteq W$ – a set of words that match the topic of the search. Binary relation $R_{SEARCH} = \{(e_{qi}, w_{qj}) | e_{qi} \in E_q, w_{qj} \in W_q\}$ defines a "word-element" pairs. For that pair, the words belong to the set of words which correspond to the topic and the elements which belong to a set of selected elements.

Let $w_{pj} \in W_p$ – a set of words extracted from the web page. Then the predicate which evaluates the binary of "element-word" pair:

$$P_w\left(e_{qi}, w_{pj}\right) = \begin{cases} 1, if\ (e_{qi}, w_{pj}) \in R_{SEARCH}, \\ 0, if\ (e_{qi}, q_{pj}) \notin R_{SEARCH}. \end{cases}$$

The predicate that defines the presence of control words in a particular element:

$$P_e\left(e_{qi}\right) = P_w\left(e_{qi}, w_{p1}\right) \vee P_w\left(e_{qi}, w_{p2}\right) \vee \ldots \vee P_w\left(e_{qi}, w_{pn}\right).$$

The web page estimation combines the estimates for each item and determined by the predicate:

$$P_q = P\left(e_{q1}\right) \vee P\left(e_{q2}\right) \vee \ldots \vee P\left(e_{qs}\right).$$

The page estimation is based on a data source model. The presence of different combinations of words in different combinations of elements of the web page is estimated. Let $R_{SOURCE} \subseteq E \times W$ – be a binary relation "is used for sources selection", this is given as follows:

$$R_{SOURCE} = \left\{\left(e_i, w_j\right) \middle| \, e_i \in E_s, w_j \in w_{qi}, w_{qi} \in W_q\right\},$$

where w_{qi} – set of words according to the element e_i.

The predicate that estimates a pair of "element-words" is defined as:

$$P\left(e_i, w_j\right) = \begin{cases} 1, if\ (e_i, p_p) \in R_{SOURCE}, \\ 0, if\ (e_i, p_p) \notin R_{SOURCE}. \end{cases}$$

The predicate that estimates an item using different word combinations:

$$P(e_i) = (P(e_i, w_{p1}) \wedge P(e_i, w_{p2}) \wedge \ldots \wedge P(e_i, w_{pj})) \vee (P(e_i, w_{pj+1}) \wedge \ldots) \vee \ldots$$

Web page estimation for various combinations of elements is given by predicate:

$$P_S = (P(e_1) \wedge P(e_2) \wedge \ldots \wedge P(e_s)) \vee (P(e_j) \wedge \ldots) \vee \ldots$$

The binary relation "elements and corresponding words that were extracted from the source page for representation the template model, $R_{PAGE} \subseteq E \times W$, $R_{PAGE} = \left\{(e_1, w_1), \ldots, (e_s, w_j)\right\}$. The function of transforming the word combinations into a value template from the set of "standards" $C = \left\{c_1, \ldots, c_m\right\}$ is given as:

$$\forall (e_i, w_j) \in R_{PAGE} : F(e_i) = \begin{cases} c_1, if\ (w_{i1} \wedge w_{i2} \wedge \ldots) \vee (w_{j1} \wedge w_{j2} \wedge \ldots) \vee \ldots \\ \ldots \\ c_m, if\ (w_{im} \wedge w_{im} \wedge \ldots) \vee (w_{jm} \wedge w_{jm} \wedge \ldots) \vee \ldots \end{cases}$$

A set of elements of a web page that contains a certain standard of a set $C = \{c_1, \ldots, c_m\}$ is given as $E_p = \left\{e_j \in E \,|\, c = F(e_j), c \in C\right\}$. Let $R_{PATTERN} \subseteq E \times C$ – the binary relation "elements contain benchmarks", at the same time $R_{PATTERN} = \left\{\left(e_i, c_j\right) \middle| \, e_i \in E_p, c_j \in C\right\}$.

The template predicate looks like:

$$P_{pattern} = \begin{cases} 1, if\ (\exists e_1 \exists e_2 \exists e_3 (E(e_1, e_F) \wedge E(e_2, e_I) \wedge E(e_3, e_O))) \equiv 1, \\ 0, in\ other\ case. \end{cases}$$

where $E(e_1, e_F) = \begin{cases} 1, e \in E_F, \\ 0, e \notin E_F; \end{cases}$, $E(e_2, e_I) = \begin{cases} 1, e \in E_I, \\ 0, e \notin E_I; \end{cases}$ and

$$E(e_3, e_O) = \begin{cases} 1, e \in E_O, \\ 0, e \notin E_O. \end{cases}$$

Let $K = \{k_j\}$ – a set of indicators that are relevant to this signs, then the predicate $M(k, k_j)$ determines, whether the summary template has indicator k_j from this set. Let $I = \{i_\gamma\}$ – the set of indicators according to the given one, then predicate $M(i, i_\gamma)$ determines whether the generic template contains the data for this indicator i_γ.

We can propose the reference model to factual information retrieval and analysis (fig.1). The main concepts are facts that are some knowledge about real-world objects, web-pages which contain text, indicators for representing attributes, and values of those attributes. The appropriate models must formalize the factual data processing.

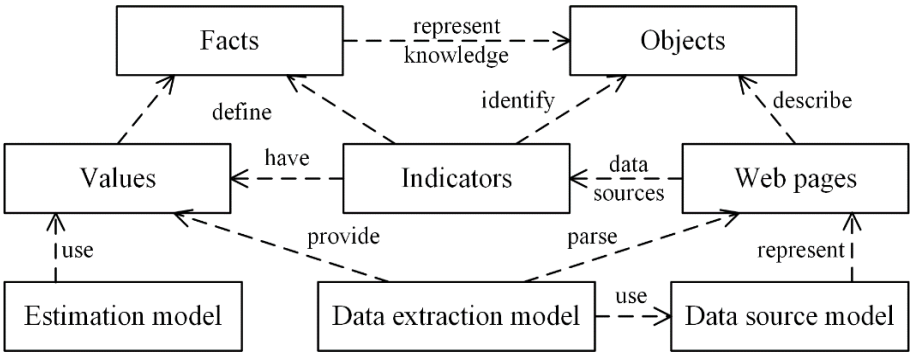

Figure 1. Reference model

In order to realize proposed reference model, the software should be developed. We suggest the basic components which are presented in figure 2. The developed software consists of such basic components as Document Search component, Fact retrieval component, Fact Analysis component, Visualization component, and Control component. Our future work is to make implementing and experimenting with the proposed model.

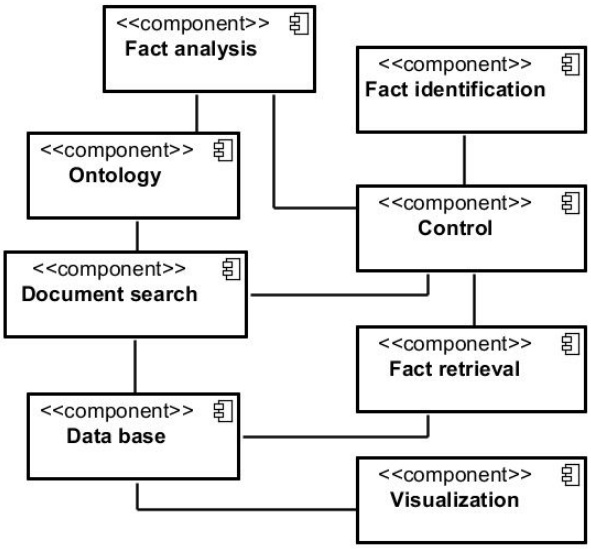

Figure 2. Basic components

5 Conclusions and Future Works

Summarizing, we can say that factual analysis is a rather complex system that has great potential and functionality. The tasks under which the data are built are designed to facilitate the work of analysts, to carry out filtration as well as structuring of huge volumes of information, which in our time are one of the main tasks of a person.

As result, we can underline that the task of identifying instances, relations, events and their relevant properties in natural language texts is still live issue. In general, we consider two kinds of facts. Despite existing data extraction solutions the task of extracting facts still is not solved. We propose to use predicate algebra and method of comparator identification to create a model of searching and extracting factual data. The future work will be devoted to research the similarity of facts and mismatch identification. We hope to develop a mathematical tool based on the relation of tolerance to make a conclusion about similarity or mismatch in the set of extracted facts.

6 References

1. Mauridhi Hery Purnomo et al., Biomedical Engineering Research in the Social Network Analysis Era: Stance Classification for Analysis of Hoax Medical News in Social Media / Procedia Computer Science 116, 2017, pp. 3–9.
2. Álvaro Figueira, Luciana Oliveira, The current state of fake news: challenges and opportunities / Procedia Computer Science 121, 2017, pp. 817–825.

3. Georg Rehm, An Infrastructure for Empowering Internet Users to Handle Fake News and Other Online Media Phenomena / An Infrastructure for Empowering Internet Users, 2017, pp. 216-231.
4. Giovanni Luca Ciampaglia, Fighting fake news: a role for computational social science in the fight against digital misinformation / J Comput Soc Sc (2018) 1:147–153, https://doi.org/10.1007/s42001-017-0005-6
5. E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret and L. de Alfaro, "Some Like it Hoax: Automated Fake News Detection in Social Networks," Cornell University, New York, USA, 2017.
6. B. Riedel, I. Augenstein, G. P. Spithourakis and S. Riedel, "A simple but tough-to-beat baseline for the Fake News Challenge stance detection task," Cornell University, New York, USA, 2017.
7. N. Rakholia and S. Bhargava, "“Is it true?” – Deep Learning for Stance Detection in News," Stanford University, California, USA, 2016.
8. N. J. Conroy, V. L. Rubin and Y. Chen, "Automatic Deception Detection: Methods for Finding Fake News," in Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community , Missouri, USA, 2015.
9. Rubin, V., Conroy, N., and Chen, Y., Towards News Verification: Deception Detection Methods for News Discourse. 2015.
10. Bondarenko M. F., Shabanov-Kushnarenko U. P. Theory of intelligence: a Handbook //SMIT Company, Kharkiv. – 2006.
11. Nina Khairova, Natalia Sharonova. Use of Predicate Categories for Modelling of Operation of the Semantic Analyzer of the Linguistic Processor./Proceedinga of IEEE EAST-West Design & Test Symposium EWDTS'09 (2009).
12. Khairova, N.F., Petrasova, S., Gautam, A.P.S.: The logical-linguistic model of fact extraction from English texts. In: Dregvaite, G., Damasevicius, R. (eds.) ICIST 2016. CCIS, vol. 639, pp. 625–635. Springer, Cham (2016).
13. Khairova N., Lewoniewski W., Węcel K. (2017) Estimating the Quality of Articles in Russian Wikipedia Using the Logical-Linguistic Model of Fact Extraction. In: Abramowicz W. (eds) Business Information Systems. BIS 2017. Lecture Notes in Business Information Processing, vol 288. Springer, Cham
14. Cherednichenko O., Yanholenko O. Information Technology of Web-Monitoring and Measurement of Outcomes in Higher Education Establishment //EuroSymposium on Systems Analysis and Design. – Springer International Publishing, 2015. – P. 103-116. (http://dx.doi.org/10.1007/978-3-319-24366-5_8)