Comparative Analysis of the Informativeness and Encyclopedic Style of the Popular Web Information Sources

Nina Khairova[1](✉), Włodzimierz Lewoniewski[2], Krzysztof Węcel[2],
Mamyrbayev Orken[3], and Mukhsina Kuralai[4]

[1] National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine
khairova@kpi.kharkov.ua
[2] Poznań University of Economics and Business, Poznań, Poland
{wlodzimierz.lewoniewski,krzysztof.wecel}@ue.poznan.pl
[3] Institute of Information and Computational Technologies, Almaty, Kazakhstan
morkenj@mail.ru
[4] Al-Farabi Kazakh National University, Almaty, Kazakhstan
kuka_ai@mail.ru

Abstract. Nowadays, very often decision making relies on information that is found in the various Internet sources. Preferred are texts of the encyclopedic style, which contain mostly factual information. We propose to combine the logic-linguistic model and the universal dependency treebank to extract facts of various quality levels from texts. Based on Random Forest as a classification algorithm, we show the most significant types of facts and types of words that most affect the encyclopedic-style of the text. We evaluate our approach on four corpora based on Wikipedia, social and mass media texts. Our classifier achieves over 90% F-measure.

Keywords: Encyclopedic · Informativeness · Universal dependency
Random Forest · Facts extraction · Wikipedia · Mass media

1 Introduction

Very often the decision making depends on the information that is found in various Internet sources. Enterprises increasingly use various external big data sources in order to extract and integrate information into their own business systems [1]. Meanwhile, the Internet is flooded with meaningless blogs, computer-generated spam, and texts that convey no useful information for business purposes. Firms, organizations, and individual users can publish texts that have different purposes. The quality of information about the same subject in these texts can greatly depend on different aspects. For business purposes, however, organizations and individual users need a condensed presentation of material that identifies the subject accurately, completely and authentically. At the same time, the subject matter should be displayed in a clear and understandable manner.

In other words, the decision making should prefer texts of an encyclopedic style, which is directly related to the informativeness concept i.e. the amount of useful

W. Abramowicz and A. Paschke (Eds.): BIS 2018, LNBIP 320, pp. 333–344, 2018.
https://doi.org/10.1007/978-3-319-93931-5_24

information contained in a document. Obviously, the amount of knowledge that human consciousness can extract from a text has to correlate with the quality and quantity of facts in the text. Based on the definitions of an encyclopedia and encyclopedia articles [2] we can suggest that an encyclopedic text has to focus on factual information concerning the particular field, which is defined. We propose to join the use of the logic-linguistic model [3] and the universal dependency treebank [4] to extract facts of various quality levels from texts.

The model that we described in our previous studies defines the complex and simple facts via correlations between grammatical and semantic features of the words in a sentence. In order to identify these grammatical and semantic features correctly, we employ the Universal Dependencies parser, which can analyze the syntax of verb groups, subordinate clauses, and multi-word expressions in the most sufficient way.

Additionally, we take into account the employing of proper nouns, numerals, foreign words and some others provided by POS-tagging morphological and semantic types of words in the text, which can have an impact on the briefness and concreteness of particular information.

In our study, we focus on using information about the quality and quantity of facts and morphological and semantic types of the words in a text to evaluate the encyclopedic style of the text.

In order to estimate the influence of quality and quantity of factual information and semantic types of words of the text on its encyclopedic style, we decided to use four different corpora. The first one comprises Wikipedia articles which are manually divided into several classes according to their quality. The second Wikipedia corpus comprises only the best Wikipedia articles. The third corpus is The Blog Authorship Corpus[1], which contains posts of 19,320 bloggers gathered from blogger.com. The corpus incorporates a total of 681,288 posts and over 140 million words - or approximately 35 posts and 7250 words per person [5]. The fourth corpus comprises news reports from various topics sections of The New York Times" and "The Guardian", which are extracted in January 2018. We apply the Random Forests algorithm of Weka 3 Data Mining Software in order to estimate the importance of investigated features in obtained classification model.

2 Related Work

Nowadays, the problem of determining informativeness of a text has become one of the most important tasks of NLP. In recent years, many articles devoted to the solution of the problem have appeared.

Usually, informativeness of a text is considered at three levels of the linguistic system: (1) at the sentence level, (2) at the level of the discourse and (3) at the level of the entire text. In traditional linguistics, the definition of the sentence informativeness is based on the division of an utterance into two parts - the topic (or theme) and the comment (or rheme, rema). [6]. At the discourse level, the traditional approach involves the anchoring of information units or events to descriptives and interpretives within a narrative frame [7].

[1] http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm.

Many studies determine ‘informativeness’ of text via ‘informativeness’ of words in the text or - ‘term informativeness’. Herewith, the most part of known approaches to measure term informativeness falls into the statistics-based category. Usually, they estimate informativeness of words by distributional characteristics of words in a corpus. For instance, a recent Rennie and Jaakkola’s study [8] introduced the term informativeness based on the fit of a word’s frequency to a mixture of 2 Unigram distribution.

The considerably fewer studies measure the semantics value of term informativeness. In our opinion, an interesting one is Kireyev’s research [9], which defined informativeness of term via the ratio of a term’s LSA vector length to its document frequency. More recently, Wu and Giles [10] defined a context-aware term informativeness based on the semantic relatedness between the context and the term’s featured contexts (or the top important contexts that cover most of a term’s semantics).

Most of approaches use statistical information in corpora. For instance, Shams [11] explored possibilities to determine informativeness of a text using a set of natural language attributes or features related to Stylometry—the linguistic analysis of writing style. However, he focused mainly on the search for informativeness in the specific biomedical texts.

Allen et al. [12] studied the information content of analysts report texts. Authors suggested that informativeness of a text is determined by its topics, writing style, and features of other signals in the reports that have important implications for investors. At the same time, they emphasized that more information texts are more assertive and concise. Their inferences about informativeness of a text are based on investors’ reaction to the analyst reports for up to five years.

In [13] work, informativeness analysis of language is determined using text-based electronic negotiations, i.e. negotiations conducted by text messages and numerical offers sent through electronic means. Appling Machine Learning methods allowed the authors to build the sets of the most informativeness words and n-grams, which are related to the successful negotiations.

Lex et al. guessed that informativeness of a document could be measured through factual density of a document, i.e. the number of facts contained in a document, normalized by its length [14].

Although the concept of encyclopedicness is closely related to the informativeness, it also includes such notions as brevity and correspondence to a given subject-matter. We suppose that the notion of ‘encyclopedicness of the text’ is more interesting and more useful than ‘informativeness of the text’ because it bases on knowledge concerning the particular subject-matter. Therefore, in our study, we consider the influence both of the various types of facts and semantic types of words of the text on the encyclopedic style of the text.

3 Methodology

3.1 Logical-Linguistic Model

We argue that the encyclopedic style of an article can be represented explicitly by the various linguistic means and semantic characteristics in the text. We have defined four possible semantic types of facts in a sentence, which, in our opinion, might help to determine the encyclopedicness of the text. Figure 1 shows the structural scheme for distinguishing four types of facts from simple English sentences.

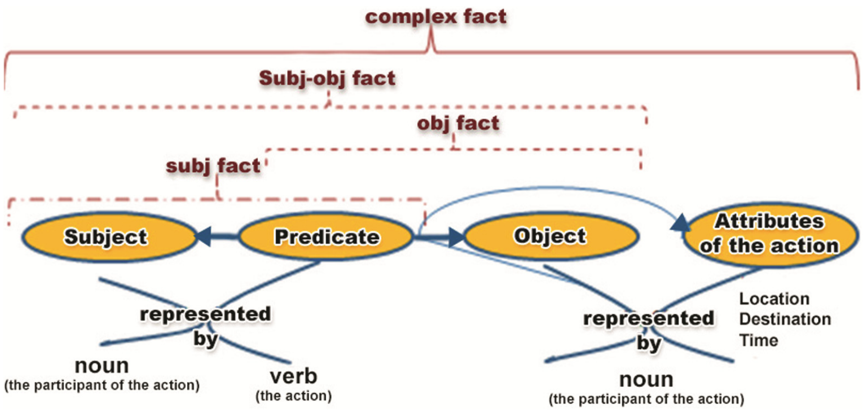

Fig. 1. Structural scheme for distinguishing four types of facts from simple English sentences: subj-fact, obj-fact, subj-obj fact and complex fact.

We called the first type of facts as *subj-fact*. We defined this type of the facts in an English sentence as the smallest grammatical clause that includes a verb and a noun. In this case, the verb represents Predicate of an action and the noun represents Subject[2] of an action. According to our model of fact extraction from English texts [3], the semantic relations that denote the Subject of the fact can be determined as the following logical-linguistic equation:

$$\begin{gathered}\gamma_1(z, y, x, m, p, f, n) = y^{out}((f^{can} \vee f^{may} \vee f^{must} \vee f^{should} \vee f^{could} \vee f^{need} \vee f^{might} \vee \\ f^{would} \vee f^{out})(n^{not} \vee n^{out})\left(p^{I} \vee p^{ed} \vee p^{III}\right)x^{f}m^{out} \vee (x^{l}(m^{is} \vee m^{are} \vee m^{havb}m^{hasb} \vee \\ m^{hadb} \vee m^{was} \vee m^{were} \vee m^{be} \vee m^{out})z^{by}),\end{gathered} \tag{1}$$

where the subject variable z^{prep} defines the syntactic feature of the preposition after the verb in English phrases; the subject variable y ($y^{ap} \vee y^{aps} \vee y^{out} = 1$) defines whether there is an apostrophe in the end of the word; the subject variable x defines the position of the noun with respect to the verb; the subject variable m defines whether there is a form of the verb "to be" in the phrase and the subject variable p defines the basic forms of the verb in English.

Additionally, in this study, we have appended two subject variables *f* and *n* to account for modality and negation. The subject variable *f* defines the possible forms of modal verbs:

$$f^{can} \vee f^{may} \vee f^{must} \vee f^{should} \vee f^{could} \vee f^{need} \vee f^{might} \vee f^{would} \vee f^{out} = 1$$

[2] We use 'Predicate', 'Subject' and 'Object' with the first upper-case letters to emphasize the semantic meaning of words in a sentence.

Using the subject variable *n* we can take into account the negation in a sentence:

$$\mathrm{n}^{\mathrm{not}} \vee \mathrm{n}^{\mathrm{out}} = 1$$

Definition 1. The *subj-fact* in an English sentence is the smallest grammatical clause that includes a verb and a noun (or personal pronoun) that represents the Subject of the fact according to the Eq. (1).

The Object of a verb is the second most important argument of a verb after the subject. We can define grammatical and syntactic characteristics of the Object in English text by the following logical-linguistic equation:

$$\begin{aligned}\gamma_2(\mathrm{z, y, x, m, p, f, n}) = {} & \mathrm{y}^{\mathrm{out}}(\mathrm{n}^{\mathrm{not}} \vee \mathrm{n}^{\mathrm{out}})(\mathrm{f}^{\mathrm{can}} \vee \mathrm{f}^{\mathrm{may}} \vee \mathrm{f}^{\mathrm{must}} \vee \mathrm{f}^{\mathrm{should}} \vee \mathrm{f}^{\mathrm{could}} \vee \mathrm{f}^{\mathrm{need}} \vee \\ & \mathrm{f}^{\mathrm{might}} \vee \mathrm{f}^{\mathrm{would}} \vee \mathrm{f}^{\mathrm{out}})(\mathrm{z}^{\mathrm{out}}\mathrm{x}^{1}\mathrm{m}^{\mathrm{out}}\left(\mathrm{p}^{\mathrm{I}} \vee \mathrm{p}^{\mathrm{ed}} \vee \mathrm{p}^{\mathrm{III}}\right) \vee \mathrm{x}^{\mathrm{f}}\left(\mathrm{z}^{\mathrm{out}} \vee \mathrm{z}^{\mathrm{by}}\right)(\mathrm{m}^{\mathrm{is}} \vee \mathrm{m}^{\mathrm{are}} \vee \\ & \mathrm{m}^{\mathrm{havb}} \vee \mathrm{m}^{\mathrm{hasb}} \vee \mathrm{m}^{\mathrm{hadb}} \vee \mathrm{m}^{\mathrm{was}} \vee \mathrm{m}^{\mathrm{were}} \vee \mathrm{m}^{\mathrm{be}} \vee \mathrm{m}^{\mathrm{out}})(\mathrm{p}^{\mathrm{ed}} \vee \mathrm{p}^{\mathrm{III}}),\end{aligned} \tag{2}$$

Definition 2. The *obj-fact* in an English sentence is the smallest grammatical clause that includes a verb and a noun (or personal pronoun) representing the Object of the fact according to the conjunction of grammatical features in Eq. (2).

The third group of facts includes main clauses, in which one noun has to play the semantic role of the Subject and the other has to be the Object of the fact.

Definition 3. The *subj-obj fact* in an English sentence is the main clause that includes a verb and two nouns (or personal pronouns), one of them represents the Subject of the fact accordance with the Eq. (1) and the other represents the Object of the fact accordance with the Eq. (2).

We also defined the *complex fact* in texts as a grammatical simple English sentence that includes a verb and a few nouns (or personal pronouns). In that case, the verb also represents Predicate, but among nouns, one of these has to play the semantic role of the Subject, other has to be the Object and the others are the attributes of the action.

Definition 4. The *complex fact* in an English sentence is the simple sentence that includes a verb and a few nouns (or personal pronouns), one of them represents the Subject, another represents the Object of the fact in accordance with the Eqs. (1) and (2) respectively and the others represent attributes of the action.

These can be the attributes of time, location, mode of action, affiliation with the Subject or the Object, etc. According to our previous articles, the attributes of an action in English simple sentence can be represented by nouns that were defined by the logical-linguistic equations [12].

Additionally, we distinguish a few semantic kinds of nouns that can be extracted by labels of POS-tagging. Moreover, additionally, we distinguish a few semantic types of nouns that can be extracted by labels of POS-tagging. These are *NNP** (plural and single proper noun), *CD* (numeral), *DT* (determiner, which marked such words as “all”, “any”, “each”, “many” etc.), *FW* (foreign word), *LS* (list item marker), *MD* (modal auxiliary). The approach is based on our hypothesis that occurrence of the proper names, numerals, determiner words, foreign words, items markers, modal auxiliary words in a text can influence its encyclopedicness. For instance, the occurrence of proper nouns, numerals, foreign words and items markers in a sentence can explicitly represent that a statement is formulated more precisely and concisely. On the contrary, we can guess that occurrence of modal auxiliary words in a sentence makes the statement vaguer and more implicit.

3.2 Using Universal Dependencies

In order to correctly pick facts out and properly distinguish their type, we employ the syntactic dependency relation. We exploit Universal Dependencies parser because for this treebanks can the most sufficiently analyze verb groups, subordinate clauses, and multi-word expressions for a lot of languages. The dependency representation of UD evolves out of Stanford Dependencies (SD), which itself follows ideas of grammatical relations-focused description that can be found in many linguistic frameworks. That is, it is centrally organized around notions of subject, object, clausal complement, noun determiner, noun modifier, etc. [4, 15]. These syntactic relations, which connect words of a sentence to each other, often express some semantic content. The verb is taken to be the structural center of clause structure in Dependency grammar and all other words are either directly or indirectly connected to it. In Dependency grammar just as a verb is considered to be the central component of a fact and all participants of the action depend on the Predicate, which expresses the fact and is represented by a verb [16]. For example, Fig. 2 shows the graphical representation of Universal Dependencies for the sentence "*The Marines reported that ten Marines and 139 insurgents died in the offensive*", which is obtained using a special visualization tool for dependency parse - DependenSee[3].

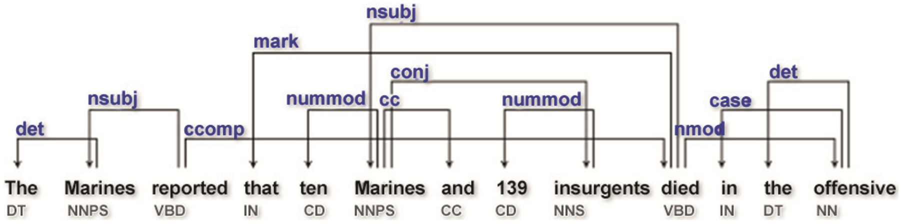

Fig. 2. Graphical representation of Universal Dependencies for the sentence "*The Marines reported that ten Marines and 139 insurgents died in the offensive*". Source: DependenSee

For our analysis we used 7 out of 40 grammatical relations between words in English sentences, which UD v1 contains[4]. In order to pick the *subj-fact* out, we distinguish three types of dependencies: *nsubj*, *nsubjpass* and *csubj*. *Nsubj* label denotes the syntactic subject dependence on a root verb of a sentence, *csubj* label denotes the clausal syntactic subject of a clause, and *nsubjpass* label denotes the syntactic subject of a passive clause. Additionally, in order to pick the *obj_fact* out, we distinguish three types of

[3] http://chaoticity.com/dependensee-a-dependency-parse-visualisation-tool/.
[4] http://universaldependencies.org/en/dep/.

dependencies: *obj*, *iobj*, *dobj* and *ccomp*. *Obj* denotes the entity acted upon or which undergoes a change of state or motion. The labels *iobj*, *dobj* and *ccomp* are used for more specific notation of action object dependencies on the verb.

Considering that the root verb is the structural center of a sentence in Dependency grammar, we distinguish additional types of facts that we can extract from a text. The root grammatical relation points to the root of the sentence [15]. These are such fact types that are formed from the root verb (*root_fact*, *subj_fact_root*, *obj_fact_root*, *subj_obj_fact_root*, *complex_fact_root*) and the other ones, in which the action Predicate is not a root verb (*obj_fact_notroot*, *subj_obj_fact_notroot*, *complex_fact_notroot*).

For completeness of the study, we attribute sentences with copular verbs to a special type of facts, which we called *copular_fact*. We should do this for the following reason. Despite the fact that, as is widely known, the copular verb is not an action verb, such a verb can often be used as an existential verb, meaning "to exist".

4 Source Data and Experimental Results

Our dataset comprises four corpora, two of which include articles from English Wikipedia. We consider texts from Wikipedia for our experiments for a few reasons. First, we assume that since Wikipedia is the biggest public universal encyclopedia, consequently Wikipedia's articles must be well-written and must follow the encyclopedic style guidelines. Furthermore, Wikipedia articles can be divided into a different quality classes [16–18], hence the best Wikipedia's articles have a greater degree of encyclopedicness than most other texts do. These hypotheses allow us to use the dataset of Wikipedia articles in order to evaluate the impact our proposed linguistic features on the encyclopedic style of texts.

The first Wikipedia corpus, which we called "*Wikipedia_6C*", comprises 3000 randomly selected articles from the 6 quality classes of English Wikipedia (from the highest): Featured articles (FA), Good articles (GA), B-class, C-class, Start, Stub. We exclude A-class articles since this quality grade is usually used in conjunction with other ones (more often with FA and GA) as it was done in the previous studies [17, 18]. The second Wikipedia corpus, which is called "*WikipediaFA*", comprises 3000 only the best Wikipedia articles that randomly are selected from the best quality class - FA.

In order to process plain texts of described above corpora, we use Wikipedia database dump from January 2018 and special framework WikiExtractor,[5] which extracts and cleans text from a Wikipedia database dumps.

In addition, in order to compare the encyclopedic style of texts from Wikipedia and texts from other information sources, we have produced two further corpora. The first one is created on the basis of The Blog Authorship Corpus [5]. The corpus collected posts of 19,320 bloggers gathered from blogger.com one day in August 2004. The bloggers' age is from 13 to 47 years[6]. For our purposes, we extract all texts of 3000 randomly selected bloggers (authors) from two age groups: "20s" bloggers (ages 23–27) and "30s" bloggers (ages 33–47). Each age group in our dataset has the same number of bloggers

[5] https://github.com/attardi/wikiextractor.

[6] Groups description available on the page http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm.

(1500 each). Since bloggers have a different number of posts of various size, in our corpus we consider all texts of one blogger as a separate item. Hence we got in total 3000 items for our "*Blogs*" corpus.

The second supplementary corpus, which is called "*News*", is created on the basis of articles and news from popular mass media portals, such as "The New York Times"[7] and "The Guardian"[8]. For a more comprehensive analysis, we extracted an equal number of news from different topics. For our experiment, we selected 10 topics for each news source. So, we extracted 150 recent news from each topic of each news source ("The New York Times" and "The Guardian") in January 2018. In total, we have got 3000 items for our News corpus.

Thus, we have had four corpora with the same number of items for our experiments. Table 1 shows the distributions of the analyzed texts according to the categories. By categories, we mean topics of newspaper posts in the "*News*" corpus, age groups of bloggers in the "*Blogs*" corpus and the special manual mark of assessment quality of Wikipedia articles in "*Wikipedia_6C*" corpus and "*Wikipedia_FA*" corpus.

Table 1. The distributions of the analyzed texts by our four corpora

Corpus name	Categories	Items in each category	Total number of items in corpus
Wikipedia_6C	FA, GA, B-class, C-class, Start, Stub	500	3000
Wikipedia_FA	FA	3000	3000
Blogs	"20s" blogs, "30s" blogs	1500	3000
News	Business, Health, N.Y., Opinion, Politics, Science, Sports, Tech, U.S., World topics of "The New York Times" UK news, World, Sport, Opinion, Culture, Business, Lifestyle, Technology, Environment, Travel topics of "The Guardian"	200	3000

Additionally, based on the Corpus Linguistics approaches [19], in order to compare the frequencies of linguistic features occurrence in the different corpora, we normalized their frequencies per million words. That allows to compare the frequencies of various characteristics in the corpora of different sizes.

Definition 5. The frequency of each feature in a corpus is defined as the number of the feature occurrence in the corpus divided by the number of words in –the corpus multiplied by million.

In order to assess the impact of the various types of facts in a sentence and some types of words in a text on the degree of encyclopedic text, we focus on two experiments. Both of them classify texts from Blogs, News and Wikipedia. The difference lies in the selected Wikipedia corpus. In the first experiment, we used texts from Wikipedia_6C

[7] https://www.nytimes.com/.
[8] https://www.theguardian.com.

corpus, which includes Wikipedia articles of different quality. We called this experiment as BNW6 model. In the other experiment we use texts from ‘Wikipedia_FA’ corpus, which only consists of the best Wikipedia articles. We called second experiment as BNWF model.

The used Random Forests classifier of the Data Mining Software Weka 3[9] allows determining the probability that an article belongs to one of the three corpora. Table 2 shows detailed accuracy by two models respectively.

Table 2. Detailed Accuracy by models

Model	TP rate	FP rate	Precision	Recall	F-Measure	MCC	ROC area	PRC area
BNW	0.887	0.057	0.888	0.887	0.887	0.831	0.974	0.950
BNWF	0.903	0.048	0.904	0.903	0.904	0.856	0.981	0.962

Additionally, the used Data Mining Software Weka 3 allows constructing a confusion matrix that permits to visualize the performance of the model. Such matrices for both classification models are shown in Table 3. Each row in the matrix allows representing the number of instances in a predicted class, while each column represents the instances in an actual class. It makes possible to present which classes were predicted correctly by our model.

Table 3. Confusion Matrices of the obtained models.

BNW6			
Blogs	**News**	**Wikipedia**	
2691	274	34	**Blogs**
199	2575	227	**News**
10	276	2714	**Wikipedia_6C**

BNWF			
Blogs	**News**	**Wikipedia_FA**	
2683	283	33	**Blogs**
206	2655	140	**News**
24	183	2793	**Wikipedia_FA**

Obviously, that the best Wikipedia articles must be well-written and consequently must follow the encyclopedic style guidelines. This is confirmed by higher coefficients of recall and precision of BNWF classification model than BNW6 one.

The Random Forest classifier can show the importance of features in the models. It provides two straightforward methods for feature selection: mean decrease impurity and mean decrease accuracy. Table 4 shows the most significant features, which are based on average impurity decrease and the number of nodes using that feature.

[9] https://www.cs.waikato.ac.nz/ml/weka/index.html.

Table 4. The most significant features of our models based on average impurity decrease (AID) and the number of nodes using the features (NNF)

Feature	BNW6		BNWF	
	AID	NNF	AID	NNF
root_fact	0.53	7526	0.52	6354
subj_fact_root	0.47	7614	0.48	6287
subj_fact_notroot	0.45	6537	0.45	5786
obj_fact_notroot	0.42	5678	0.41	4772
obj_fact_root	0.4	5155	0.4	5354
subj_obj_fact_root	0.38	5270	0.39	4479
complex_fact_root	0.39	3994	0.38	3882
complex_fact_notroot	0.35	5541	0.35	4413
copular_fact	0.34	3924	0.33	3254
CD	0.33	4262	0.32	3668
DT	0.32	4646	0.31	4061
FW	0.29	2083	0.31	2087
MD	0.31	4388	0.3	3702
NNP*	0.29	4893	0.29	4112
LS	0.22	775	0.23	664

5 Conclusions and Future Works

We consider the determination problem of the encyclopedic style and informativeness of text from different sources as a classification task. We have four corpora of texts. Some corpora comprise more encyclopedic texts or articles and others include less encyclopedic ones.

Our study shows that factual information has the greatest impact on encyclopedicness of text. As Fig. 1 shows, we distinguish several types of facts in the sentence. They are complex *fact, subj fact, obj fact, subj-obj fact* and *copular-fact*. Additionally, we highlight the main fact that is represented by a sentence.

Table 4 summarizes that the most significant features that affect the encyclopedic style of the text are (1) the frequency of the main facts (*root_fact*), (2) the frequency of the subj facts, (3) the frequency of the *obj facts* and (4) the frequency of the *subj_obj* facts in a corpus. We definite all these types of facts on the basis of our logical-linguistic model and using Universal Dependencies parser.

The Random Forest classifier, which bases on our features, allows obtaining sufficiently high recall, precision and F-measure. We provide Recall = 0.887 and Precision = 0.888 in the case of the classification of texts by Blogs, News and Wikipedia corpora. In the case of considering only the best of Wikipedia articles in the last corpus, we provide recall = 0.903 and precision = 0.904. Moreover, using the Random Forest classifier allowed us to show the most important features related to informativeness and the encyclopedic style in our classification models.

In future work, we plan to extend obtained approach to compare the encyclopedic style of texts of Wikipedia and of various Web information sources in different languages. In our opinion, it is possible to implement the method in commercial or corporate search engines to provide users with more informative and encyclopedic texts. Such tools must be significant for making important decisions based on the text information from the various Internet sources. On the other hand, firms and organizations will get the opportunity to evaluate the informativeness of the texts that are placed on their Web sites, and make changes to provide more valuable information to potential users. Additionally, more encyclopedic texts can be used to enrich different open knowledge bases (such as Wikipedia, DBpedia) and business information systems in enterprises.

References

1. Cai, L., Zhu, Y.: The challenges of data quality and data quality assessment in the big data era. Data Sci. J. **14**, 1–10 (2015)
2. Béjoint, H.: Modern Lexicography: An Introduction, pp. 30–31. Oxford University Press (2000)
3. Khairova, N., Petrasova, S., Gautam, A.: The logical-linguistic model of fact extraction from English texts. In: International Conference on Information and Software Technologies, Communications in Computer and Information Science, CCIS 2016, pp. 625–635 (2016)
4. Nivre, J., et al.: Universal dependencies v1: a multilingual treebank collection In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, May. European Language Resources Association (ELRA) (2016)
5. Schler, J., Koppel, M., Argamon, S,. Pennebaker, J.: Effects of age and gender on blogging. In: Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, pp. 191–197 (2006)
6. Leafgren, J.: Degrees of explicitness: information structure and the packaging of Bulgarian subjects and objects. John Benjamins, Amsterdam & Philadelphia (2002)
7. Berman, R.A., Ravid, D.: Analyzing narrative informativeness in speech and writing. In: Tyler, A., Kim, Y., Takada, M. (eds.) Language in the Context of Use: Cognitive Approaches to Language and Language Learning. Cognitive Linguistics Research Series. pp. 79–101. Mouton de Gruyter, The Hague (2008)
8. Rennie, J.D.M., Jaakkola, T.: Using term informativeness for named entity detection. In: Proceedings of SIGIR 2005, pp. 353–360 (2005)
9. Kireyev, K.: Semantic-based estimation of term informativeness. In: Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, pp. 530–538 (2009)
10. Wu, Z., Giles, L.C.: Measuring term informativeness in context. In: Proceedings of NAACL 2013, Atlanta, Georgia, pp. 259–269 (2013)
11. Shams, R.: Identification of informativeness in text using natural language stylometry. Electronic Thesis and Dissertation Repository, 2365 (2014)
12. Huang, A.H., Zang, A.Y., Zheng, R.: Evidence on the information content of text in analyst reports. Acc. Rev. **89**(6), 2151–2180 (2014)
13. Sokolova, M., Lapalme, G.: How much do we say? Using informativeness of negotiation text records for early prediction of negotiation outcomes. Group Decis. Negot. **21**(3), 363–379 (2012)

14. Lex, E., Voelske, M., Errecalde, M., Ferretti, E., Cagnina, L., Horn, C., Granitzer, M.: Measuring the quality of web content using factual information. In: Proceedings of the 2nd joint WICOW/AIRWeb Workshop on Web Quality, pp. 7–10. ACM (2012)
15. De Marneffe, M.C., Manning, C.D.: Stanford typed dependencies manual, pp. 338–345. Technical report. Stanford University (2008)
16. Lewoniewski, W.: Enrichment of information in multilingual wikipedia based on quality analysis. In: Abramowicz, W. (ed.) BIS 2017. LNBIP, vol. 303, pp. 216–227. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-69023-0_19
17. Węcel, K., Lewoniewski, W.: Modelling the quality of attributes in wikipedia infoboxes. In: Abramowicz, W. (ed.) BIS 2015. LNBIP, vol. 228, pp. 308–320. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-26762-3_27
18. Lewoniewski, W., Węcel, K., Abramowicz, W.: Quality and importance of wikipedia articles in different languages. In: Dregvaite, G., Damasevicius, R. (eds.) ICIST 2016. CCIS, vol. 639, pp. 613–624. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46254-7_50
19. McEnery, T., Hardie, A.: Corpus Linguistics: Method, Theory and Practice, pp. 48–52. Cambridge University Press, Cambridge (2012)