Generalization annotation for its automatic recognition in media texts

Krassovitsky A.M., Ualiyeva I.M., Meirambekkyzy Zh., and Mussabayev R.R.

Institute of Information and Computational Technologies,
Almaty, Kazakhstan
akrassovitskiy@gmail.com, i.ualiyeva@gmail.com, jazirahanim@mail.ru,
rmusab@gmail.com

Abstract. In this paper, we consider a concept of an informational feature generalization, which is analyzed in the media texts. We have researched several linguistic computational approaches for its automatic recognition. The generalization is studied in the field of knowledge extraction as one of author's biases in texts and as a tool of readers manipulation. Automatic recognition of generalization in texts is a new linguistic topic. An annotation scheme on the basis of linguistic rules has been identified for manual generalization detection. Using these rules, an experiment has been conducted to identify generalizations in the texts of the official media.

Keywords: Knowledge Extraction · Generalization · Bias in the Text · Linguistic Rules · Annotation Scheme

1   Introduction

Machine processing of texts involves the procedures for their automatic recognition. By automatic recognition is meant the development of a computer program that solves a certain task of text analysis in an automated way. However, before developing a software, it is necessary to conduct a linguistic analysis of the entity under study. In this case, the problem of selecting linguistic methods for studying language models and text structure is occurred. This research is focusing on recognition of *generalization* in texts.

There are a number of methods that authors use in analytical articles to manipulate the readership [1–9]. One of these techniques is logical transition from the particular to the general, subordination of the particular phenomena to the general principle [10] is called generalization. Generalization is the property of human thinking to make general conclusions about different situations on the basis of his own experience and observations. For example, in the next sentence, the author of the news publication uses G. to express his own negative experience and assigns to all objects: hospitals, science, art, schools: *As a result - there is no science, hospitals do not heal, schools do not teach, universities do not give proper education, theaters are not in demand, television is mediocre, art is secondary and low-standard.* Recognition of G. forms can be used for classification

of authors style. It is possible to detect biased analytical texts on the basis of G. Having analyzed the large corpus of G. articles, one can identify the range of contemporary socially significant problems of society.

We consider generalization as an instrument for readers manipulation, that is why it is important to take into account sentiment analysis of text. Psychologists note [2] that expressive phrases in the text have the strongest impact on the reader. An important aspect of our study is a combination of G. and negative sentiment in one phrase, and marking it as bias in the text.

In our research we analyze ways how to manually recognize generalized statements in a text. The analysis consists of these stages: (1) an annotation scheme by experts; (2) compiling a dictionary of key words expressing G; (3) linguistic rules for manual and automatic recognition of G in texts, which is identified on the basis of the annotation scheme. Open access analytical news publications from Kazakhstan sites in Russian language have been used in the research. The results obtained in the form of rules and dictionary will be included as features for automatic recognition of G in texts. Also, we tried to model how a particular person-annotator perceives G., after that the material is transferred to another annotator for agreement.

2   Related Work

The review revealed that G.'s concept has not been seriously analyzed in the Russian-language scientific literature yet. The question of the algorithmization of the G. de-grees numerical evaluation in the texts has not previously been considered. The works we have studied on G. belong to political scientists and linguists. G. has been more specifically considered by Dankova N.S. [4]. The author investigates means of lexical and grammatical levels of the language, as well as stylistic devices designing G. strategy in the Russian and English press. According to her research, this linguistic phenomenon is also created through the selection of facts, realizing the journalists intention. Dankova N.S. analyses the language means that form G. In her work, she says that this strategy is implemented in combination with sentiment, and states solitary examples of G. strategy use in its positive form and only in English media. As an example, she gives many sentences, one of them is the following: *Read these stories of how the secret courts imprison the elderly in care homes against their will and weep.* Giving this sentence as an example, Dankova N.S. refers it to grammatical means implementing G. strategy, which use nouns and pronouns in plural form, emphasizing the scope covering the described phenomena of reality. Dankova N.S. identifies groups of features expressing G. in texts, such as lexical groups, grammatical language means, stylistic devices, and selection of facts.

Frolova I.F. in her work [3] presents G. as one of the ways of expressing bias in analytical articles of the British and Russian media. Frolova I.F. considers G. as syn-thesis of the individuals opinion into the general opinion. As an example, Frolova I.F. gives the following sentence: *Barroso's options appear very limited. What can Europe do except put Yanukovych in the same box as the dic-*

tator next door, Alexander Lukashenko of Belarus, and blacklist his cronies and oligarch financiers to stop them shopping in London, skiing in Switzerland or holidaying in the Mediterranean? She claims that *in this context ... the British journalist assumes the right to speak out for the whole Europe, demonstrating Europes almost indisputable right to give a negative assessment to Yanukovych, Lukashenko, etc.* Orlova O.G. [5] in her study considers G. as a way to express a stereotype. Stereotypes in relation to Russia are considered as an example.

In Van Dijks work [6], G. is presented as one of the methods of ideological strategy used in political discourse. He states that instead of presenting the facts, the authors use G. to express bias and emphasize negative characteristics of the object under study. He states also that generalization is often made to formulate negative attitudes and prejudices, rather than presenting facts. It follows from here that G. is used to express bias.

In the other work, Van Dijk [7] speaks of a strategic G., aimed at controlling the addressees thinking, carried out by the addressant. A Nigerian political scientist Ajiboye E. in the work [8], claims that G., along with hyperbole and polarization, is a linguistic strategy that expresses an ideological influence on the view and assessment of events. She makes this conclusion by examining user comments regarding political instability in Nigeria and the bias of northern citizens towards southern ones. For ex-ample, G. is used to express bias and mood, when, the author notes that *the north is always cruel.*

Given that G. is a property of thinking, it is found only in analytical opinion piece. The relevance of studying G. by political scientists and linguists is due to the fact that G. is a powerful tool in the process of information warfare for decision-making by people, respectively, for their minds and sentiment.

A number of studies aimed at finding and automatic identification of bias in English texts [11–14]. For example, the work of Recasens M. et al. [11] shows the phenomenon of bias in Wikipedia articles and is aimed at automating the recognition of biased terms. Reviewers follow a predetermined format and do annotational markup of the article sentences with allocation of controversial words. The set of such marked sentences makes up the corpus. Further, the authors classify bias types for their subsequent comparison with respect to a context. The logistic regression carried out on the set of context features of biased words was calculated on the test sets and showed results comparable to expert ones.

Morstatter F. et al. in [14] use machine learning to recognize possible variants of the meaning distortions/corruptions in the text, inherent in biased statements with the use of repeating patterns. Levels where it is possible to define distortions by computer approach are investigated: whether there is a repetition template if yes, that what type of a template it is and what is its sentiment orientation. Such aspects of bias as use of attitude, paraphrase is affected. Both texts of media sources and official texts selected on specific themes and marked by experts are used when comparing corpus texts. Although this paper uses some bias in semantics due to machine translation, good properties of predictions are demonstrated when using simple linguistic attributes. Iyyer M. et al. in [12] apply a

recursive neural network (RNN) framework to identify ideology from text focus on bags of words or wordlists, ignoring syntax.

Rosenthal et al. [13] detect biased words using indicator features such as belief words, question feature using a bag of words approach, lexical features, and social media features.

Recent work in knowledge extraction uses different approaches: lexical approach [15], machine learning approach [13], statistical [16, 17], and deep learning [12], and hybrid approach [18]. We have applied the lexical method based on the dictionary.

The motivation of this work is to study the possibilities of automatic detecting G. in texts, because manually identifying G. is an expensive and impractical process. Also, we consider G. as an additional informative feature as part of various text classification algorithms.

3   Dataset

A corpus from 218 articles has been taken to study properties of generalization, where 30 of them were marked by annotators as generalizing, 30 as analytical with-out generalization, and other marked as news. All publications have been taken from the Kazakhstan news and analytical portals like *tengrinews.kz, ca-monitor.kz.*

Dictionaries A dictionary of generalized words (hereinafter referred to as *KEY-dictionary*) to identify the generalization was developed [19]. A dictionary was compiled manually and consists of 600 words and phrases with regular expressions. It includes the following lexical groups of words that express G:

- non-specific verbs expressing abstractness, nonconcreteness of events, uncertainty of actions: *seem – kazat'sja, tolerate – terpet', think – dumat'*
- abstract nouns: *cold – holod, visions – videnija*
- non-specific pronouns: *some – kakoj-nibud', something – koe-chto*
- universal quantifiers are words that do not allow exceptions: *all – vse, always – vsegda*
- lexical units with a regularity/irregularity value: *as a rule – kak pravilo, usually – obychno*
- quantitative indicators: *every tenth – kazhdyj desjatyj, a lot – mnogo*
- lexical units with general recognition semantics: *it is considered – schitaetsja, it is known – izvestno*
- stylistic devices containing metaphors with a generalized meaning: *beat a record – pobil rekordy, this drug loop – jeta narkoticheskaja petlja.*

For sentiment analysis RuSentiLex dictionary [20] is used. It consists of 16000 words and phrases, manually marked by sentiment. The categories of sentiment annotation are: positive, negative and neutral. Also, an important feature of this RuSentiLex is the additional source in the dictionary: manually annotated for opinion, feeling, fact. The categories opinion, feeling in the dictionary are the words, which express authors attitude.

4 Methods

4.1 Generalization evaluation on the basis of lexical features

Inspiring the lexicon-based method for sentiment analysis in work of Taboada M. et al. [15] we introduce an approach based on lexical features [19]. It was the first approach to evaluate G. Firstly, annotators divide texts on five degrees of G.: *very strongly expressed, strongly expressed, not identified in the text, weakly expressed, very weakly expressed.* A hypothesis, that it is possible to divide the texts into five degrees by dictionary-based lexical approach and get its numerical evaluation was put forward.

Numerical evaluation of G. was calculated as the sum of G. weights of each generalized sentence to the total number of sentences S in the text:

$$gen(S, K, F, O) = \frac{\sum_r weight(s_r)}{|s|}. \tag{1}$$

where S is a set of sentences in the text, K is a set of elements (words and phrases) of the sentence that coincide with the elements of the KEY-dictionary, F is a set of elements in the sentence that express the fact, O is a set of elements in the sentence expressing the opinion of the publications author, s_r sentences that have G., where $1 \leq r \leq q$ and q is a number of sentences in the text.

The weight of the G. sentence $weight(s_r)$ was calculated as the number of elements' matches from the KEY-dictionary with the elements of the sentence. G. sentences, that have only one element from the KEY-dictionary, $weight(s_1) = 1$. In case several elements of the KEY-dictionary are found in the sentence, the numerical estimate of G. is amplified by multiplying by the gain multiplier G. $\alpha : weight(s_r) = \alpha \sum_{i=2}^{l} weight(w_i)$. The initial weight in case of weakening $weight(w_i) = 0.75$. If the sentence contains words expressing an opinion or fact, the numerical evaluation of the G. sentence is reduced by multiplying the elements of the sets F, O in the sentence $weight(s_r) = \prod_{i=1}^{m} weight(w_i)$. When there are elements from the KEY-dictionary in the text and elements of the sets F, O, then G. is calculated as their product. $weight(s_r) = 0$ if there are no elements of the sets K in the sentence under consideration. This integral estimation *gen* takes into account the number of words and phrases in the KEY-dictionary and the presence of facts/opinions found in each sentence. The first strengthen, and the second reduce contribution value of the sentence to the final G. evaluation of the article. The former is strengthened, while the latter reduce the value of the contribution of the sentence to the final evaluation of the G. article.

A special software has been developed in order to numerically estimate the G. and to help annotators to detect it in sentences [21]. However, as experiments show no explicit division according to classes has been found. An advantage of this approach is that it has automatically recognized news texts with an accuracy of 83%.

4.2   Annotation Scheme

By annotation we understand the arrangement of special markers (tags) in the text expressing additional elements of text information. We have distinguished following types of annotations: *lexical* and *formally-grammatical.*

In a lexical annotation, experts identify words and phrases that have a meaning of G. Formal grammatical annotation is the marking by experts of the morphological and syntactic features of generalized words and phrases.

During the lexical annotation we highlight generalizing words and quantifiers. Quantifiers, which includes the KEY-dictionary, are universal words expressing quantity, it can be an absolute maximum (*all*), an absolute minimum (*no one*), for example: *All children have to go to school.* Words *always, every/each/everyone, never* are generalizing words. For example: *Everyone can achieve goals.*

We show an annotation by drawn attention to one analytical publication. In this publication, the author makes generalization: describes a situation about increase of qualified specialists leaving country. The author of the publication has formed a position on this issue, which is revealed through the use of G. Below is annotation scheme.

We use linguistic glosses for translation to show how the text and its translation are related from the point of view of grammar and meaning. We have used Leipzig Glossing Rules and applied Rule 2: Morpheme-by-morpheme correspondence. Annotation of sentences consists of 3 lines: original example in Russian, gloss in the sub-script, its translation into English.

S.1. *The outflow of professionals, or, as it is called, brain drain, eventually leads to a general of life is falling in a country: quality of education, health services and cul-ture has been decreasing. – Ottok professionalov, ili, kak jeto nazyvaetsja, utechka mozgov, v konechnom schete privodit k tomu, chto v strane padaet obshhij uroven' zhizni: snizha-jutsja kachestvo obrazovanija, medicinskih uslug i kul'tura.* Outflow-NOM-SG professional-GEN-PL or as it is call-PRS drain brain in end account leads to that what in country-LOC-SG fall-PRS-SG general life-GEN-SG: lower-PRS-PL quality education-GEN-SG, medicine service-GEN-PL and culture.

The author's subjective attitude is expressed in his/her generalized conclusion: the deterioration in standard of living. Generalization is expressed by subjectivity of opin-ion, the author transfers deterioration to different areas: education, medicine, culture.

In this statement, strategy G. is implemented by the following language means: the semantics of deprivation is presented in the lexemes *outflow, brain drain, falling, decreasing.* Lexically, generalization is given by the phrases: *general level of life is falling, quality of education, health services* and *culture is decreasing.* Formal grammatical indicators include: imperfective verbs that emphasis duration and permanence of actions. Application of homogeneous parts of the sentence adds pressure to the meaning. The authors negative attitude to the question is expressed by the lexical phrase *falling, is decreasing,* and the

expression is intensified by the idiom *brain drain*. Thus, the author expresses a subjective generalized conclusion regarding this problem by means of generalization in negative sentiment.

S.2. *As a result - there is no science, hospitals do not heal, schools do not teach, universi-ties do not give proper education, theaters are not in demand, television is medio-cre/untalented, art is secondary and low-standard. – V rezul'tate nauki net, bol'nicy ne lechat, shkoly ne uchat, vuzy ne dajut dolzhnogo obrazovanija, teatry ne vostrebovany, televidenie bezdarno, ikusstvo vtorichno i nizkoprobno.*
In result science-GEN-SG no hospital-NOM-PL not heal-PRS-PL, school-NOM-PL not teach-PRS-PL, university-NOM-PL not give-PRS-PL proper-GEN-SG education-GEN-SG, theater-NOM-PL not demanded, television-NOM-SG untalented, art secondary and low-standard.

The formal grammatical means of expressing G. and negative sentiment are negation *no, particle not*, prefix *un-* and homogeneous parts of the sentence, plural forms of nouns(*hospitals, schools, universities, theaters*) that express a large number of these objects. When applying gradation(*an arrangement of a number of expressions relating to one subject in a sequential order, increasing the emotional significance of the row members*) G. is vividly expressed and negative semantics of the phrase is amplified. The author's reasoning is based on generalization of certain factors. These statements form an opinion that such spheres as education, culture, science, art are destroyed. Readers who are not familiar with the situation in the country may perceive this generalized conclusion as a true statement.

S.3. *Overall level of unprofessionalism and just amateurism is growing in all spheres of life. – Rastet obshhij uroven' neprofessionalizma i prosto diletantstva vo vseh sferah zhizni.*
Grow-PRS-SG overall level unprofessionalism-GEN-SG and just amateurism-GEN-SG in all-ABL-PL sphere-LOC-PL life-GEN-SG.

Formal grammatical means of G. include an imperfect form of verbs expressing du-ration and permanence of action. Negative sentiment is given by lexemes/lexical units *unprofessionalism* and *amateurism*. Lexical means include a universal generalizer from the dictionary - *in all spheres*. The author transfers these negative properties to all areas, thereby expressing his subjective position.

S.4. *They distort facts, they exaggerate, they moralize. – Oni iskazhajut fakty, oni sgushhajut kraski, oni moralizirujut.*
They distort-PRS-PL fact-NOM-PL, they exaggerate-PRS-PL, they moralize-PRS-PL

Example (S.4.) is taken from another article and conveys the meaning of negative processes' immensity. Formal grammatical G. Indicators include: a generalized object is pronoun in plural form, an imperfect form of verbs with the meaning of continuity, and syntactic enumerations of homogeneous parts of the

sentence. Negative senti-ment is expressed by lexemes with negative semantics: *distort, exaggerate, moralize.* Author's subjectivity with negative meaning is enhanced by the expressive word exaggerate. Thus, the situation is absolutized.

In the process of research, number of generalizing words in the KEY-dictionary in-creases. The list of linguistic rules that expresses G. is increased also.

On the basis of this annotation, linguistic rules (Table 1) were identified by formalizing words and phrases and syntactic links between them. It should be noted that when creating a dictionary, we included this lexical group of words and in the process of research expanded the dictionary with similar words.

Table 1. Linguistic rules to extract generalized sentences in a text.

No.	Markers	Lexeme 1	Lexeme 2	Example
R1	Lexical and grammatical	Quantifiers (in KEY-dictionary)	Singular form of a noun	*Every citizen voted in favor of the reform*
R2	Lexical and grammatical	Quantifier all/everybody	Plural form of a noun	*All citizens voted in favor of the reform*
R3	Grammatical	Plural form of a noun	Plural form of a verb (to be)	*Qualified specialists are gradually leaving the country*
R4	Grammatical	Negative particle not	Finite verb form	*Schools do not teach, universities do not provide proper education, theaters are not in demand*
R5	Lexical-syntactic	A verb with the meaning of the feature intensification	Nouns with negative semantics	*Overall level of unprofessionalism and amateurism is growing*
R6	Phraseological units			*Brain drain*

5   Experiment

In order to estimate linguistic rules, we ask annotators to find sentences in texts to match the rules. Annotators get (1) a set of linguistic rules, (2) instructions explaining the task and (3) a test sample of annotated articles to perform an experiment. The annotator looks for sentences corresponding to linguistic rules in each article. New other publications have been taken for the experiment than

above dataset. The test sample includes five publications: three analytical and two news articles. For each sentence of publications the annotator verifies the rules and writes out counter-examples separately when the rule did not work. Counter-examples of non-generalized sentences, which formally coincide with linguistic rules that means formal-grammatical method couldnt be total approach. Phraseological units are not identified by the rules, therefore it has a sense to compile a separate manual dictionary of phraseological units with G. meaning. It shows that the consistency of these statements to the rules according to grammatical features is not a sufficient condition for generalization recognition. Therefore, both formal grammatical features which are compiled into rules and lexical features that serve for the KEY-dictionary are important in our annotation.

In other examples if there are specific descriptions, events, facts, sources of information, G. in the sentence is not identified. If we expand the G. sentence and add information source, specific details of the event, then the sentence becomes non-generalized. Examples: (1) *all citizens against reform* (2) *according to the survey, 61.4% of 5,000 Kazakhstans citizens do not agree with the reform in education.* Thus, this experiment reveals that linguistic rules allow articles classification for analytical and informational news articles. However, annotators have found counter-examples that suggest the rules cannot develop a set of complete and consistent methods for G. identification in the text.

6   Conclusion and Further Work

A generalization is defined as a biased statement containing lexical and semantic units of generalized statement and uncertainty. At the same time, it has been shown that text sentiment is important an amplifier bias. It was built an annotational scheme for linguistic model construction to identify the G. The linguistic rules for the peer review were developed based on this annotational scheme. There is a problem of creating complete, consistent and non-redundant rules for transferring our cognitive ideas about the author's bias to machine logic. Our goal is to choose a complete model, first at the syntax level, and then by automatically comparing our rule set with the markup of proposals to begin forming a semantic ontology.

References

1. van Dijk TA (2002) Discourse, Ideology and Context. Journal of Asian Economics 35: 11-40. doi:10.1515/flin.2001.35.1-2.11

2. Smit S (2014) Master the power of suggestion! Achieve everything you desire! Series Psychology for everyone. Publishing House AST, Moscow
3. Frolova IV (2015) On the expression of subjectivity in analytical articles of the qualitative British and Russian press. Political Linguistics. Section 2. Political Communication 1(51): 138-145
4. Dankova NS (2016) Strategija generalizacii kak sredstvo reprezentacii sudebnoj vlasti (na materiale rossijskih i anglijskih pechatnyh SMI). Politicheskaja lingvistika 1: 73-81
5. Orlova OG (2013) Sostavlenie slovarja stereotipov o russkih: soderzhanie i forma. Politicheskaja lingvistika 3: 175-182
6. van Dijk TA (1993) Principles of Critical Discourse Analysis. Discourse & Society 4(2): 249-283. doi:10.1177/0957926593004002006
7. van Dijk TA (1994) Critical Discourse Analysis. Discourse & Society 5(4): 435436. https://doi.org/10.1177/0957926594005004001
8. Ajiboye E (2013) Ideological Discourse Analysis of the Functions of Feedback Comments on Online Reports of Socio-political Crises in Nigeria. Covenant Journal of Language Studies (CJLS) 1: 128-147
9. Suzdal'ceva VN (2018) Generalizujushhie vyskazyvanija kak sredstvo manipulirovanija soznaniem massovogo adresata. Medialingvistika 5(1): 6072. http://hdl.handle.net/11701/15091
10. Bol'shoy entsiklopedicheskiy slovar' Homepage, https://www.vedu.ru/bigencdic/13964/. Last accessed 20 Feb 2019
11. Recasens M, Danescu-Niculescu-Mizil C, Jurafsky D (2013) Linguistic models for analyzing and detecting biased language. In: Proc. 51st Annu. Meet. Assoc. Comput. Linguist, Vol 1. Bulgaria, 2013
12. Iyyer M, Enns P, Boyd-Graber JL, Resnik P (2014) Political Ideology Detection Using Recursive Neural Networks. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, USA, June 23-25, 2014
13. Rosenthal S, McKeown K (2012) Detecting opinionated claims in online discussions. In: Proc. IEEE sixth international conference on semantic computing, IEEE, 2012
14. Morstatter F, Wu L, Yavanoglu U, Corman SR, & Liu H (2018) Identifying Framing Bias in Online News. ACM Transactions on Social Computing 1(2): 5. https://doi.org/10.1145/3204948
15. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. Computational linguistics 37(2): 267-307
16. Petrasova S, Khairova, N, Lewoniewski W, Mamyrbayev O, & Mukhsina K (2018) Similar Text Fragments Extraction for Identifying Common Wikipedia Communities. Data 3(4): 66
17. Khairova N, Petrasova S, Lewoniewski W, Mamyrbayev O, Mukhsina K (2018) Automatic Extraction of Synonymous Collocation Pairs from a Text Corpus. In: Proceedings of the Federated Conference on Computer Science and Information Systems (pp. 485488) 15. doi: 10.15439/2018F186
18. Gfu D, Cristea D (2013) Monitoring and predicting journalistic profiles. In: Proc. of the International Conference on Computational Collective Intelligence (pp 276-285). Springer, Berlin, Heidelberg, 2013
19. Krassovitsky A, Ualiyeva I, Meirambekkyzy Z, Mussabayev R (2018) Lexicon-based approach in generalization evaluation in russian language media. Mezhdunarodnyj nauchnyj zhurnal Sovremennye informacionnye tehnologii i IT-obrazovanie 14(3): 567-572

20. Lukashevich NV, Levchik AV (2016) Sozdanie leksikona ocenochnyh slov russkogo jazyka RuSentileks. In: Trudy konferencii OSTIS-2016 (pp 377-382). Belarus, 2016
21. Ualiyeva I, Krassovitsky A, Mussabayev R (2018) Generalization rate evaluation in open publication materials in Russian language. In: Proceedings of the 16th International Conference Information Technologies and Management, ISMA University, Riga, Latvia, April 26-27, 2018