EasyChair Preprint

№ 2398

Contextual Predictability of Texts for Texts Processing and Understanding

Olga Krutchenko, Ekaterina Pronoza, Elena Yagunova,
Viktor Timokhov and Alexander Ivanets

January 17, 2020

Contextual Predictability of Texts for Texts Processing and Understanding

Olga Krutchenko[1], Ekaterina Pronoza[1], Elena Yagunova[1], Viktor Timokhov[1] and Alexander Ivanets[1]
[1] St.-Petersburg State University, St.-Petersburg, Russian Federation
{krutchenko.olga, katpronoza, iagounova.elena}@gmail.com,
viktor-timohov@mail.ru, sookol98@yandex.ru

Abstract. This paper is the first part of contextual predictability model investigation for Russian, it is focused on linguistic and psychology interpretation of models, features, metrics and sets of features. The aim of this paper is to identify the dependence of the implementation of contextual predictability procedures on the genre characteristics of the text (or text collection): scientific vs. fictional. We construct a model predicting text elements and designate its features for texts of different genres and domains. We analyze various methods for studying contextual predictability, carry out a computational experiment against scientific and fictional texts, and verify its results by the experiment with informants (cloze-tests) and word embeddings (word2vec CBOW model). As a result, text processing model is built. It is evaluated based on the selected contextual predictability features and experiments with informants.

Keywords: Contextual Predictability, Language Model, Dice, Surprisal, Conditional Probability, Informational Entropy, Cloze test, Fiction texts, Scientific Corpora.

1 Introduction

Information redundancy is an inherent feature of any text, especially from the point of view of information theory. And it is precisely because of this property that a person successfully perceives and understands both oral and written text. Redundancy is an inherent property of any language and is therefore inherent in all texts, without exception, but to varying degrees, depending on the functional style of the text [1].

The concept of contextual predictability is closely connected to the process of predicting words based on their context. The effect of contextual predictability is essentially the opposite of information redundancy, demonstrating that not all the words are equivalent for perception and understanding of a text.

In this paper, an analysis of various computational methods of contextual predictability is carried out, and the most adequate metrics are selected for further verification during constructing a language model. The research involves computational analysis based on the corpora of scientific and fictional texts and experiment with informants and word embeddings.

Our aim is to identify the dependence of the implementation of contextual predictability procedures on the genre and style characteristics of the text.

Contextual predictability involves consideration of many aspects, since this topic is interdisciplinary. One of them is the psychological aspect. There are many different studies about the dependence of contextual predictability and the speed of reading of a person, their eye movement when reading [2], etc.

On the other hand, contextual predictability is directly connected with the fields of linguistics, psychology, perception and analysis of the text. Such research methods as cloze-tests, tests aimed at restoring missing elements of the text, allow to assess the degree of informants knowledge of the language, readability of the text (solving the problem of the comprehensibility of texts) [4], as well as analyze issues which may arise while teaching/studying this language [4, 5, 6, 7, 8].

But the issue of contextual predictability in computational linguistics, when solving problems associated with automatic text processing, is particularly relevant [9]. For example, contextual predictability is highly relevant for the recognition and correction of typos in the text when solving various problems associated with further text processing. Using the principles of contextual predictability, if it is impossible to recognize a word, we can assume that there is a typo in it, and then to restore the correct word with the help of the context.

Contextual predictability can also help in extracting keywords and collocations from text [10]. Since a collocation phrase has signs of a holistic semantic and syntactic unit; contextual predictability indicators values are usually high for collocations. Keywords, on the contrary, are the main source of new and significant information in the text, therefore, their contextual predictability is expected to be small, especially when they occur in the text for the first time.

Contextual predictability is also relevant for the task of predicting the words missing from text, by their context: the higher contextual predictability of text is, the easier it is to predict the missing words (and it is proved in the experiments with both the informants and automatic word prediction model).

Thus, the relevance and practical significance of the research of contextual predictability is very high for a variety of areas related to automatic text processing.

2 Related Work

2.1 General Approaches to Analyzing Contextual Predictability

From the point of view of computational linguistics, the predictability of words in the context has been little studied. However, there has been an increasing amount of research on this topic recently.

The main approaches in the contextual predictability research are the analysis of statistical data based on corpus of texts and the conduct of cloze-tests with informants. To conduct a comprehensive study, it is necessary to use a combination of the two approaches and compare the results at each of the stages. At the initial stage of the analysis of the data, two main questions arise: how to evaluate contextual predictability based on statistical data and on the basis of what corpora to conduct research.

Contextual predictability of the word in the text can be assessed in various ways. First of all, there are statistical measures of association, mainly used to identify collocations. These are measures such as MI, t-score, Dice [10, 14, 15] and others. They may be useful for both separate texts and the corpora [16]. Another possible approach to contextual predictability involves calculating informational entropy and conditional probability. These measures will be considered in 2.3 in detail.

2.2 Contextual Predictability Analysis via Cloze-Test

Cloze-test can be considered the oldest form of analyzing contextual predictability. Cloze-test was proposed by V. Taylor [6] to determine the readability of the text (an indicator of how difficult the text is for reading and perception). Its method is as follows: a prose passage of 100 to 400 words is selected, in which each n-th word is skipped. An informant is asked to recover the missing words. The success of this text is directly dependent on the time it takes for the informant to understand the entire text and restore the connection between the events. This, in turn, is determined by the informant's knowledge of the vocabulary of a given language, the extent to which he/she has developed a language guess and how adequately he/she understands the text of each specific situation [6].

This test can be used to control the process of learning a foreign language, since it allows one to accurately and objectively establish the degree of the formation of reading skills and level of vocabulary knowledge when reading.

Cloze texts have also other possible applications. Using this type of test, one can evaluate language model of a particular language. For example, in [4] it is shown that detailed information about the performance of the language model can be obtained through cloze-tests with informants.

The method of cloze-tests is also used to assess the understanding of speech by ear. Moreover, this approach is important not only for the purpose of control in teaching a foreign language, but also to study the mechanisms of perception of sounding speech, which has its own distinctive features: ellipsis, unclear pronouncing of unstressed syllables, objective interference of a communication channel, etc. This issue is considered in detail in [11] and [12].

2.3 Statistical Models for Contextual Predictability

If we consider studies that propose objective criteria for determining the complexity of an arbitrary language and ranking various languages by complexity, the paper by McWarter [17] can be considered the first work in this direction. In his work, he criticizes the prevailing opinion about the equal complexity of all languages and proves that some modern languages are simpler than the "old" ones. Later, the ideas of McWarter were developed in the works of other researchers, such as Wouter Küsters [18], Esten Dahl [19], Peter Tradgil [20], and others.

The development of contextual predictability models in computer science and related disciplines is more relevant to the our research. Such models often rely on hidden Markov processes. Hidden Markov models allow us to consider the text as a set of

processes of transition from one state to another. In this case, if we analyze the text of a sufficiently large volume, we can use n-gram frequencies to obtain the transition probabilities. For example, after analyzing Liyus Carroll's fairy tale “Alice's Adventures in Wonderland”, we found out that the state “l” (the letter “l”) occurs 100 times in the text. The next state is likely to be the state “i”, since the word “Alice” is a fairly frequent word in the text selected for the initial analysis [13].

In general, statistical models like Hidden Markov models and Conditional Random Fields are often used in natural language processing for such tasks as language modelling, document classification, clustering and information extraction [21].

It should be noted that work related to the study of informational redundancy in text, was also carried out in Russia in the 1960s (see, for example, the studies of N. N. Leontyeva, R. G. Piotrovsky, T. N. Nikitina, M. I. Otkupshchikova, specifically devoted to this topic [22]). This issue is considered in detail by P.G. Piotrovsky ([23] and [24]).

At the first stage of the current research, Hidden Markov models were considered and preliminary results were obtained. However, these results had no strict and formalized linguistic interpretation. Thus we organized our research as follows: firstly, we focus on the statistical metrics of contextual predictability and their interpretation (see this paper); secondly, we compare results of the statistical metrics with other models like Hidden Markov models.

Further in this section we describe several statistical metrics we used in this research.

Informational entropy is calculated as follows:

$$H(x) = -log_2 P(x), \tag{1}$$

where $P(x)$ denotes the probability of occurrence of the word x in text. This is a term from informational theory, and it is a measure of uncertainty of the appearance of a symbol of the primary alphabet.

Conditional probability is the probability of one event, provided that another event has already occurred [25]. Conditional probability for contextual predictability of the word is calculated as follows:

$$P(x|context) = \frac{f(x,context)}{f(context)}, \tag{2}$$

where $f(x, context)$ is the frequency of joint occurrence of the word x after the specified context, and $f(context)$) is the frequency of meeting of the context.

Mutual information (MI) is also a notion from information theory referring to the strength of the connection and allows one to assess the independence of the appearance of two words in the text. In this paper we use pMI which is calculated by the formula:

$$pMI(x_1 x_2) = log_2 \frac{f(x_1 x_2)\times N}{f(x_1)\times f(x_2)}, \tag{3}$$

where $x_2$is the word under study, x_1 is the preceding word, $f(x_1 x_2)$ is the frequency of the occurrence of the two words together, $f(x_1)$ and $f(x_2)$ are the word frequencies ofx_1 and x_2 respectively and N is corpus size (in the number of words) [10, 14]. MI tends to assign greater importance to combinations of rare words, including words with

misprints and foreign words. Therefore, it is necessary to consider a threshold for word frequency values in the corpus [10, 16].

T-score is an association measure which refers to the asymptotic criteria for hypothesis testing. It is calculated by the formula:

$$t - score(x_1, x_2) = \frac{f(x_1x_2) - \frac{f(x_1) \times f(x_2)}{N}}{\sqrt{f(x_1x_2)}}, \quad (4)$$

where x_2 is the word under study, x_1 is the preceding word, $f(x_1x_2)$ is the frequency of occurrence of the two words together, $f(x_1)$ and $f(x_2)$ are the word frequencies ofx_1 and x_2 respectively and N is corpus size (in the number of words) [10, 14].

The Dice coefficient, like MI, refers to the point estimate of a measure of connection. It is calculated by the formula:

$$Dice(x_1, x_2) = \frac{2 * f(x_1x_2)}{f(x_1) + f(x_2)}, \quad (5)$$

where x_2 is the word under study, x_1 is the preceding word, $f(x_1x_2)$ is the frequency of occurrence of the two words together, $f(x_1)$ and $f(x_2)$ are the word frequencies ofx_1 and x_2 respectively. There is also a logarithmic variant of Dice, logDice, which is often used in text processing tasks):

$$logDice(x_1, x_2) = log_2 \frac{2 * f(x_1x_2)}{f(x_1) + f(x_2)}. \quad (6)$$

This measure (both Dice and logDice) does not depend on the size of the corpus (unlike MI and t-score); it takes into account only the frequency of joint occurrence and independent frequencies. However, like MI, this measure gives an overestimation of low-frequency phrases [14, 15], although this overestimate is much less critical for the Dice measure than for the MI measure. To study contextual predictability, the following algorithm can be interesting for estimating n-word combinations using the Dice measure: for all pairs of words in a body (or text), the Dice coefficient is considered, then the elements are arranged into chunks, or linked text segments, according to a particular principle (so-called cosegment procedure [15, 26]).

The term chunk term was introduced as a cognitive term in [3] to designate a fragment (in other words, a piece) of text from several words that are commonly used together in a fixed expression. An example of such phrases: “in my opinion”, “Do you know what I mean?” and others. The selection of these phrases (chunks) was made as part of the study of mastering a foreign language [30].

The union of words in chunks occurs on the basis of a previously discussed feature of the connectivity of the two elements of the text (words).

There are two options when linked text segments extraction is concerned. The first option is as follows: pairs of words are united into one text element based on the value of the coefficients of this pair of words and the closest context. A word is not attached to the previous one, if the value of the Dice coefficient for this pair is lower than the threshold, or if it is lower than the arithmetic average of the same coefficient for the left and right pair. A condition is imposed that related chains cannot consist of more than 7

words [15]. This algorithm was introduced and described in detail by V. Daudaravicius (for example, [26]).

The second option is as follows: for each phrase a group is formed by successively merging it with context phrases. For each group, the Dice coefficient is calculated by taking into account five phrases from the left context and two phrases from the right context (such sizes of the context window is selected as this is approximately how a human perceives context).

In a computational experiment, the Dice coefficient for each bigram has to be calculated.

In this research, a condition based on the arithmetic average of two values of the Dice coefficient to the right and left of the studied words was selected as a feature for combining the two words.

The first word analyzed is always a chunk. To add each subsequent word to the chunk, the following condition must be met:

$$Dice(word2, word3) > \frac{Dict(word1,word2)+Dice(word3,word4)}{2}. \quad (7)$$

A word does not join the previous one if the value of the Dice coefficient for this pair is below the threshold, i.e. than the arithmetic mean of the same coefficient for the left and right pair. An additional limit is imposed on the length of the chunks: the number of elements (words) is not more than 7 [15, 26].

The *surprisal* metric is a measure of the content of information associated with an event in a probabilistic space. The smaller the probability of an event, the greater is the surprisal coefficient associated with the information that this event will occur [27].

This measure, proposed by H. Levvi in 2001 [28], has become standard for the tasks related to the assessment of contextual predictability. It is calculated by the formula:

$$I(x, context) = log_2 \frac{1}{P(x \vee context)} \quad (8)$$

where $P(x \vee context)$ is the conditional probability of the occurrence of the word x in a given context.

The *salience* metric for assessing the compatibility of words is much less common than MI and t-score metrics. However, it can be considered a normalized variant of the Dice metric. The salience coefficient is calculated using the formula:

$$salience(x_1, x_2) = 14 + log_2 \frac{2 \times f(x_1 x_2)}{f(x_1)+f(x_2)}, \quad (9)$$

where x_2 is the word under study, x_1 is the preceding word, $f(x_1 x_2)$ is the frequency of occurrence of the two words together, $f(x_1)$ and $f(x_2)$ are the word frequencies ofx_1 and x_2 respectively [14].

3 Data

As mentioned earlier, redundancy is an essential feature of natural language and natural language text in particular, which is necessary for perception and understanding by a

human. Redundancy is inherent in all texts, without exception, but it is not a constant value and depends on many parameters, one of which is the functional style of the text [1, 12].

The total amount of information contained in the text is called the information richness of the text. Information richness is an absolute indicator of the quality of the text (as opposed to informativeness, which depends on the degree of novelty of the topic for the reader, and therefore is a relative indicator of quality). According to the degree of information richness, the five main functional styles can be arranged as follows in ascending order: colloquial, fictional, publicistic, scientific, official business [1, 29].

According to the defined classification, conversational and artistic styles have the greatest redundancy, while the scientific and official business styles tend to increase the information richness, i.e. to reduce redundancy.

Therefore, for the study of contextual predictability, we selected two functional styles for comparison: the scientific and the fictional ones, which are expected to be the opposites in terms of the redundancy of the texts (and the value of contextual predictability for scientific texts is expected to be much higher compared to the fictional ones due to greater information richness).

We prepared two datasets of scientific text, each of which belongs to one subject area and is homogeneous in genre and theme.

For the corpus of fictional texts, we selected texts that differed in the following parameters: text volume in terms of words number, genre and the "recognition" of the work of art.

The corpus of fictional texts consists of 6 texts. By the number of words, texts range from 9 500 to 363 500 words. The total number of words in the corpus is 782 300.

As for the scientific texts, 2 subcorpora were formed: scientific articles on corpus linguistics (15 093 articles) and cognitive psychology (22 703 articles). The total number of words in the corpus is 37 796.

As the amount of scientific articles is quite small, it makes sense to carry out an analysis directly on the whole corpus, while fictional texts can be considered separately. The results of the analysis of the corpus of scientific texts and individual fictional texts can be comparable due to the common theme of scientific articles, belonging to one subject area, the presence of similar keywords (corpus of scientific articles similar in these characteristics can be perceived as a single text).

The formed corpus of texts serves as the basis for our research and for obtaining preliminary results.

For the experiment with missing words prediction we prepared a third-party fictional corpus of 337 texts and a scientific corpus of 1095 texts. Preprocessing stage included tokenization and lemmatization (using Mystem morphological parser). Continuous Bag-of-Words (CBOW) models were trained on the lemmatized corpora with the following parameters: size=300, window=4, min_count=5 for scientific corpus and size=500, window=2, min_count=5 for fictional corpus.

4 Text Model Construction

All the considered metrics for detecting contextual predictability can be classified as follows: probabilistic estimates (entropy characteristic, conditional probability, surprisal), pointwise (MI, Dice, salience) and asymptotic (t-score) estimates of communication measures. Some of them are very similar to each other, differing only by normalization.

For the practical part of the research, the following measures are of interest:

- Conditional probability and entropy characteristic, since they are the main probability metrics.
- Surprisal, because this metric is a standard for assessing contextual predictability.
- Dice coefficient, which will be used to implement the algorithm of combining collocation into linked text segments.

The selected metrics are quite diverse, and as a result are interesting for comparing their performance. And also all of them have their own characteristics, advantages and disadvantages. In this regard, they are most interesting for further testing on the corpus of texts and individual texts in the course of building a model and analyzing its work. Comparison of various methods allows us to visually identify their differences and work efficiency and analyze the results separately for each of the metrics.

The selected metrics were combined and represented as a graph. This model is a graph (see Fig. 1) where the nodes refer to the words (on the lemma level), and the edges correspond to the connections between the words and their contexts. The edges are annotated with all possible metrics and their values, and the nodes are annotated with morphological information (i.e., lemma, grammemes, etc.).

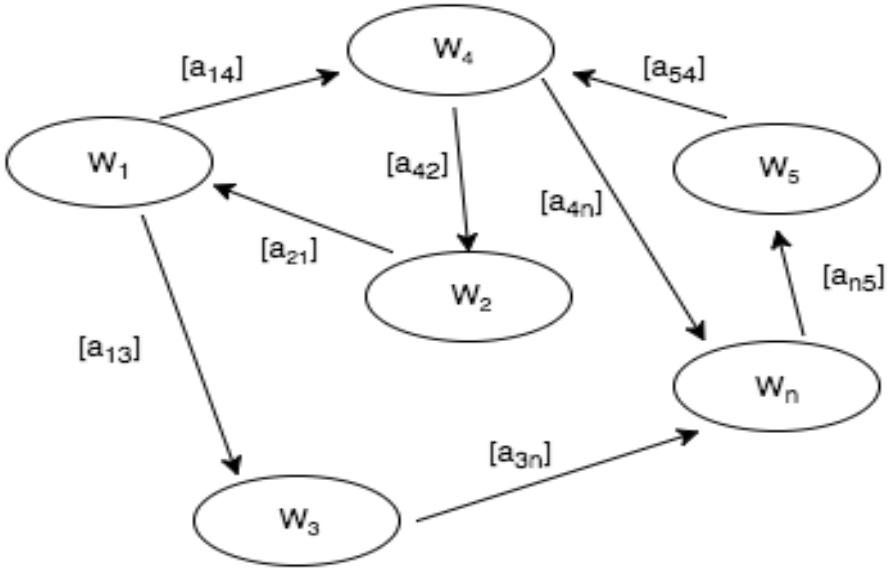

Fig. 1. Graph model for context predictability metrics representation.

To build such a model, the textual data is sequentially processed in several stages, such as:

- preprocessing, which includes tokenization and lemmatization;
- creating a frequency dictionary of tokens;
- extracting bigrams from the text and creating their frequency dictionary;
- calculating necessary attributes and metrics for each of the bigrams;
- generating the graph text model.

As part of the computational experiment, we also extracted linked text segments form the texts using Dice coefficient as described in Section 2.3.

5 Results. Discussion

As a result of the computational experiment, text models (consisting of context predictability metrics values) for each corpus were obtained. For further analysis and comparison of the results, it is necessary to take into account the volume and lexical variety of the texts in question (see Table 1).

Table 1. Volume and lexical variety of the texts.

Text (corpus)	Volume (tokens number)	Percentage of unique word-forms, %	Percentage of unique lemmas, %
Scientific corpus 1 (cognitive psychology)	13434	30.4	23.3
Scientific corpus 2 (computational linguistics)	13434	39.6	25.6
“Catching minnows in Georgia” by V. Astafev	9624	50.2	36.6
“The problem of a werewolf...” by V. Pelevin	10472	38.0	25.3
“Station on the Horizon” by E.M. Remark	49568	28.3	16.0
“Ivanhoe” by W. Scott	148466	19.7	8.3
“Singing in the thorns” by C. McCullough	200852	17.3	7.8
“The Count of Monte Cristo” by A. Dumas	363554	12.1	4.7

To analyze the results obtained for each text, we calculated the arithmetic mean value for each of the studied metrics. Such mean values are shown in Table 2.

Table 2. Volume and lexical variety of the texts.

Text (corpus)	Conditional probability	Entropy	MI	Dice	Surprisal
Scientific corpus 1 (cognitive psychology)	0.45	11.4	23	0.27	2.35
Scientific corpus 2 (computational linguistics)	0.39	11.2	22.6	0.29	2.43
“Catching minnows in Georgia” by V. Astafev	0.53	10.7	21.7	0.32	2.26
“The problem of a werewolf...” by V. Pelevin	0.42	10.3	21.2	0.22	2.68
“Station on the Horizon” by E.M. Remark	0.35	11.8	23.6	0.14	3.64
“Ivanhoe” by W. Scott	0.27	12.4	25.2	0.09	4.38
“Singing in the thorns” by C. McCullough	0.25	12.5	25.3	0.09	4.72
“The Count of Monte Cristo” by A. Dumas	0.2	12.5	25.8	0.06	5.3

At this stage of the research, it can be concluded that the values of the selected metrics depend not only on the contextual predictability of the text, but also on the volume of the given text. This suggests that we need to increase the corpus for further research. In spite of this, on the basis of the results obtained, it can already be concluded that the hypothesis of the expected higher values of contextual predictability features for the body of scientific texts in comparison with the fictional ones is confirmed.

It should be notes that the constructed text models can be used to solve practical natural language processing tasks, for example, those related to the removal of ambiguity and the correction of typos (see Table 3).

To recognize typos, one needs to identify such pairs of words where the two words differ in one letter, and lemma for one of the words is not known (in practice, it means that lemma cannot be found by a morphological parser). In this case, one can to compare the entropy values for these words. If the entropy value of one of the words is higher than the total entropy of the text, then there can be a typo in this word and it is necessary to check the values of other features with the same context for these words (if there is a context).

Table 3. Example of typos correction.

Context	Word	Entropy	Dice	Surprisal	Lemma
сегодня /today/	вечером /evening/	12	0.079	4.2	вечер /evening/
сегодня /today/	вчером /evning/	**18**	0.026	6.2	unknown

Another case of applying the results of the research in practice is the disambiguation task. Since morphological analysis is carried out automatically, it is imperfect, and errors are possible. To verify its results, it is necessary to compare the values of some contextual predictability features for the same words with different contexts. First of all, it is necessary to pay attention to the values of the surprisal metric, which is significantly higher than the average value in the text for rare bigrams (and this is the case when a disambiguation error takes place) – see Table 4.

Table 4. Disambiguation example.

Context	Word	Entropy	Dice	Surprisal	Lemma
моя /my/	вина /guilt/	16	0.026	**6.0**	вино /wine/
глоток /sip of/	вина /wine/	13	0.045	**1.6**	вино /wine/

As a result of the experiments with chunks extraction, a list of linked text segments was obtained, with their lengths ranging from 1 to 7 elements.

Comparing the results obtained for fictional and scientific functional styles, the following tendencies are found out:

- The average length of chunks in fiction texts is 5 elements, while in scientific ones it is 3 elements.
- In scientific style, each sentence is almost completely divided into chunks, while in fictional style only 1-2 integrated blocks are more common in long sentences.
- Chunks in scientific style texts are cliche phrases, introductory constructions and turns. In fictional style chunks are constituted by steady combinations and collocations.

To evaluate the performance of the constructed contextual predictability model, we conducted cloze-tests with informants.

We selected 4 fragments of texts (2 fragments of fictional style, 2 – of scientific style), belonging to various works. Each of the fragments ranges from 100 to 120 words in volume. In each fragment, 10 words are missing, which are proposed to be restored by the informants.

The choice of the omitted words was made on the basis of the surprisal (which is a standard for such procedure) and entropy metrics. In each text fragment, words with high (8–11), medium (4–8), and low (0–3) meanings of the surprisal metric were selected to be missing. It is assumed that a higher surprisal value means that the given word is worse recovered from the context. The results of the experiment with surprisal metric are also compared with those with the entropy metric.

The cloze-test in our research consists of two parts - the main and the additional. The first one contains the content part (fragments of texts with missing words), the second one includes questions for informants that need to be answered after passing the test (regarding their age category, sex and whether they recognized the books from which

the fragments were taken). The informants were offered instructions for passing the test and a test form, they were no time restrictions.

In our experiment, 10 informants participated.

Some of the excluded words, together with the informants' results, are presented in Table 5and 6 for the text fragments selected for the cloze-test, and an example of such text fragment (with omitted words) is provided as follows:

"She walked on the deck and 1) _____ new, unfamiliar Australia. In the transparent, colorless 2) _____ was slowly spreading, rose above the pearl rose 3) _____, and already in the east, on the edge of the ocean there rose 4) _____, the newborn scarlet light turned into a white day…".

In Table 5, results of the experiment on fictional texts are presented, and Table 6 contains the results obtained for the scientific corpora.

Table 5. Cloze-test results for fictional texts (a fragment).

Original word	Answer 1	Answer 2	…	Amount of correct answers	Percent of correct answers, %	Amount of correct part of speech tags	Percent of correct part of speech tags, %
увидела /saw/	увидела /saw/	увидела /saw/		10	100	10	100
сиянье /glow/	облако /cloud/	солнце /sun/		1	10	10	100
солнце /sun/	солнце /sun/	светило /luminary/		9	90	10	100
бортом /board/	бортом /board/	ней /her/		4	40	10	100
собой /self/	собой /self/	бриллианты /brilliants/		5	50	6	60
водой /water/	землей /earth/	небом /sky/		4	40	10	100

Table 6. Cloze-test results for scientific texts (a fragment).

Original word	Answer 1	Answer 2	…	Amount of correct answers	Percent of correct answers, %	Amount of correct part of speech tags	Percent of correct part of speech tags, %
иначе /else/	иначе /else/	иначе /else/		8	80	8	80
многообразие /variety/	обширность /vastness/	разнообразие /diversity/		2	20	10	100
несмотря /despite/	несмотря /despite/	взгляды /views/		7	70	8	80

правило /rule/	правило /rule/	чувство /sense/	3	30	8	80
рассматривают /consider/	считают /think/	определяют /define/	1	10	10	100
сталкиваются /encounter/	работают /work/	исследуют /investigate/	5	50	10	100
опыт /experience/	опыт /experience/	опыт /experience/	9	90	10	100

According to the results of the informants' answers, the main assumption of the experiment was confirmed: the words with low surprisal value are restored by the informants correctly or using synonyms in 85-100% of cases.

We also selected a group of words (10% of the total number of the missing words), which are unequivocally restored by informants, but have a high surprisal value (from 8 to 11). Initially it was assumed that this group of words would be less recoverable. However, entropy values of these words slightly exceed the average entropy of the text (they make 13-14, with the average entropy of the text equal to 11).

This result can be explained by the fact that for the calculation of the surprisal measure one previous word was used as the context, and the informant, filling in the blanks, was guided by the context of greater length. The fact that the context is used more widely is confirmed by the fact that the words of this group are included in the selected chunks formed on the basis of the Dice metric with the broader context (five words from the left side and two words from the right side).

For example, in the sentence "She went on deck and saw a new, unfamiliar Australia" the word "saw" was omitted. According to the results of the experiment, 100% of the informants correctly restored the word form in this case. The value of the surprisal metric for the missing word is 10.7. However, despite the high value of the surprisal metrics, when allocating chunks based on the Dice metric, the "went and saw" chain was selected. This example in particular and the whole group of these words in general confirm the need to use a broader context in the study of contextual predictability, which is closer to human perception of information.

It should also be noted that the words of this group (easily restored and with high surprisal value) in the fragments of fictional texts are more than twice as many as compared with the scientific ones.

In scientific texts, words that are cliches are almost unmistakably (in 95% of cases) restored (e.g., "in other words", "despite", "first, ..., second, ..." and others. But despite the low surprisal value (and therefore supposedly the best recoverability from the context), the terms and scientific vocabulary in fragments of scientific texts in 75% of cases are not restored by informants.

In general, the results confirm the effectiveness of the contextual predictability text model built during the computational experiment. However, some characteristics of human perception were not taken into account in the experiment (e.g., analysis of a wider context, certain knowledge and experience of a particular informant).

Our second experiment was similar to cloze test but the language model had to fill in the missing words instead of the informants. We trained two CBOW models (using word2vec tool) on fictional and scientific third-party corpora and automatically predicted missing words from their context. Results of this experiment for the two text collections are shown in Table 7.

Table 7. Missing word prediction task results.

Text collection	Percent of correct* answers, %	Percent of correct part of speech tags, %
Fictional	30	20
Scientific	0	10

Since cloze-test is a hard task for a language model, its answer is considered correct when the true missing word or its synonym is present in top-10 most probable words returned by the model. Results of the missing words prediction task are much worse than that of the informants, which is not surprising. At the same time, the three correct answers given by the language model refer to the situations when a word is actually a part of the collocation (e.g., "your" in "your excellency", "say" in "better say"), while other words, not predicted by the language model correctly, are less dependent on their context.

6 Conclusion

In this paper, we consider various metrics for calculating contextual predictability and construct language model using these metrics. Our data incudes scientific and fictional texts, and we experiment with informants (cloze tests) and word embeddings language models (missing words prediction task).

Results of the experiments prove that the implementation of contextual predictability procedures depends on the genre and style characteristics of the text. Words with higher context predictability values of various metrics are easier restored by both human informants and language models than those with lower context predictability values.

Acknowledgements

The authors acknowledge the RSF for the research grant 18-18-00114.

References

1. Yagunova E.V.: Fundamentals of theoretical, computational and experimental linguistics, or Reflections on the place of the linguist in computational linguistics. Automatic processing of texts in natural language and computational linguistics: studies. allowance / Bolshakova E.I., Klyshinsky E.S., Lande D.V., Noskov A.A., Peskova O.V., Yagunova E.V. - M .: MIEM, 2011. [in Russian]

2. Biemann Ch., Remus St. and Hofmann M. J.: Predicting word 'predictability' in cloze completion, electroencephalographic and eye movement data. Natural Language Processing and Cognitive Science / Bernadette Sharp, Wiesław Lubaszewski and Rodolfo Delmonte (eds). Libreria Editrice Cafoscarina, Venezia. P.83-95.
3. Miller, G. A.: The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review. 63 (2): 81–97. 1956.
4. Owens M., O'Boyle P., Mcmahon J., Ming J., Smith Fj.: A comparison of human and statistical language model performance using missing-word tests. Language and speech, 1997, vol. 40, №4. – P. 377-389.
5. Richard D. Robinson: The Cloze Procedure: a New Tool for Adult Education. Adult Education Qarterly. 1973 – P. 23, 97-98.
6. Taylor W. L.: Cloze procedure: a new tool for measuring readability. Journalism Quarterly, 1953. – P. 415-433.
7. Oller J. W., Jr., Grover Kh Yii, Greenberg L.A., Hurtado R.: The learning effect from textual coherence measured with cloze. Cloze and coherence. Eds J. W. Oller, Jr., J. Jonz (Eds). – Cranbury, NJ, 1994. P. 247-268.
8. Nusbaum H. C. et al. : Why cloze procedure? Cloze and coherence. Eds J.W. Oller, Jr., J. Jonz (Eds) – Cranbury, NJ, 1994. – P. 1-20.
9. Yagunova E.V.: Study of the contextual predictability of text units using corpus resources. Proceedings of the International Conference "Corpus Linguistics - 2008". - SPb.: SPSU, 2008b. p. 396-403 [in Russian]
10. Yagunova E.V., Pivovarova L.M.: The nature of collocations in the Russian language. The experience of automatic extraction and classification on the material of news texts. Proc. STI, Ser.2, №6. M., 2010. [in Russian]
11. Yagunova E.V.: Variability of perception strategies of sounding text (experimental research based on Russian-language texts of various functional styles). SPSU - Perm, 2008. [in Russian]
12. Yagunova E.V.: Investigation of the redundancy of Russian sounding text. Redundancy in the grammatical structure of the language. Ed. ed. M. D. Voeikov. SPb .: Science, 2010. - 462 p. [in Russian]
13. Markov Models for Text Analysis. Purdue University, Department of Statistics. 2009. http://www.stat.purdue.edu/~mdw/CSOI/MarkovLab.html (15.04.2016).
14. Khokhlova M.V.: The study of lexical-semantic compatibility in Russian with the help of statistical methods (based on corpus text). /St. Petersburg, 2010. [in Russian]
15. Yagunova E.V., Pivovarova L.M:. Study of the structure of news text as a sequence of connected segments // Computational linguistics and intellectual technologies: Based on the materials of the annual International Conference "Dialogue" (Bekasovo, May 25-29, 2011). Issue 10 (17) .- M .: Izd-vo RSUH, 2011. [in Russian]
16. Yagunova E.V.: Study of the contextual predictability of text units using corpus resources. Proceedings of the International Conference "Corpus Linguistics - 2008". - SPb. : SPSU, 2008. - p. 396-403 [in Russian]
17. J. McWhorter: The world's simplest grammars are creole grammars. Linguistic typology. 2001. 5(2–3).
18. W. Kusters.: Linguistic complexity: the influence of social change on verbal inflection. Utrecht, 2003.
19. Ö. Dahl: The growth and maintenance of linguistic complexity. Amsterdam, 2004.
20. P. Trudgill: Sociolinguistic typology: social determinants of linguistic complexity. Oxford, 2011.

21. Y. Sun, H. Deng, J. Han: Probabilistic Models for Text Mining. Mining Text Data. 2012. - P 259-295.
22. Berdichevsky A.: Language complexity (Language complexity). Questions of linguistics. 2012. №5. [in Russian]
23. Piotrovsky R.G.: Linguistic Automaton (in research and continuous learning). SPb., 1999. [in Russian]
24. Piotrovsky R.G.: Informational measurements of language. L., 1968. [in Russian]
25. D. MacKay: Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003.
26. V. Daudaravicius: Automatic Identification of Lexical Units. Computational Linguistics and Intelligent text processing CICling, 2009.
27. Decision Trees: Entropy, Information Gain, Gain Ratio. Marina Santini: http://www.slideshare.net/marinasantini1/lecture-4-decision-trees-2-entropy-information-gain-gain-ratio-55241087?related=1 (18.03.2016).
28. Myslín, Mark, & Roger Levy: Codeswitching and predictability of meaning in discourse. Language 91(4), 2015.
29. Babaylova A.E.: Text as a product, means and object of communication in teaching a non-native language. Ed. Saratov University, 1987. [in Russian]
30. Miller J.A.: The magic number is seven plus or minus two. On some limits of our ability to process information / Ed. Yu.B. Gippenreiter, V.Ya. Ro-manov. - Moscow: CheRo, 1998. - p. 564-582. [in Russian]