Machine learning vs human learning

J. Henno*, H. Jaakkola** and J. Mäkelä***
* Tallinn University of Technology, Estonia
** Tampere University, Pori
*** University of Lapland, Rovaniemi, Finland
jaak@cc.ttu.ee

***Abstract* – Machine Learning (ML) is a technology to make messages created by humans (text, images, speech etc.) more understandable for computers so that they could better answer humans' queries and needs when recalling this information. Here is considered the ML sub-area – Natural Language Processing (NLP) and presented examples of its methods using text corpuses created from MiproCE presentations.**

I. INTRODUCTION

Communication with computers involves one major difficulty: computers do not (yet) understand our human language and only a limited number of humans (called programmers) understand (some of) computer's languages.

It is unrealistic to expect, that all humans learn to understand computer's language, i.e. become programmers; besides, for understanding computer code our brains use quite different mechanisms than when understanding human language [1][2] – for our brains program code is not at all a language.

We reason, define our problems, goals and wishes always first in our everyday human language – English, Estonian, Croatian etc., but we express ourselves also in other formats – in music, dance, images. In order to improve human-computer communication we have to teach computers better to understand our multimodal formats of communication.

II. NEW APPROACHES

A. *From linguists to programmes*

Understanding, i.e. parsing human language in computers is a difficult problem. Traditionally this was a research area for linguists, who built for parsers big handcrafted grammars. These parsers often generated large numbers of possible parses for a given input sentence. But in tests performed in 1990 where to best of handcrafted parsers were presented short (max length -13 words) sentences from the Associated Press news these 'hand-made' by linguists parsers understood only 30% or less of presented sentences [3].

Therefore currently NLP does not use traditional linguistic concepts – grammar, parsing, parts of speech, it is totally based on statistical properties of texts. In NLP linguists have been displaced with programmers; characteristic is the quote from Frederick Jelinek, long-time head of the Center for Language and Speech Processing at Johns Hopkins University, where he developed speech recognition systems : "*Whenever I fire a linguist our system performance improves*" [4]. The currently most famous living linguist Noam Chomsky commented the situation ironically: "*There is a notion of success ... which I think is novel in the history of science. It interprets success as approximating unanalyzed data.*"

III. NLP - FROM STATISTICS TO DATA SCIENCE

Polonius: What do you read, my lord?
Hamlet: Words, words, words.

A. *NLP – Data Entry, Text corpus*

The computer-based natural language research uses large corpora of human-produced text, often billions of words. In the following are used texts from the Mipro CE subconference from years 2017-2020, both in English and in Croatian languages. Most NLP tools do not depend on concrete language, thus it were interesting to find similarities and differences what NLP methods reveal in the English and Croatian-language text corpuses and what 'new' can be distilled from these texts.

B. *NLP - Statistics*

Text analyzing begins with classical statistical methods, e.g. count of total number of words and count vocabularies – unique words. It turned out, that while the corpuses (total number of words) have been steadily growing, vocabularies (unique words) used in both languages – English and Croatian - remained during 2017..2020 nearly the same.

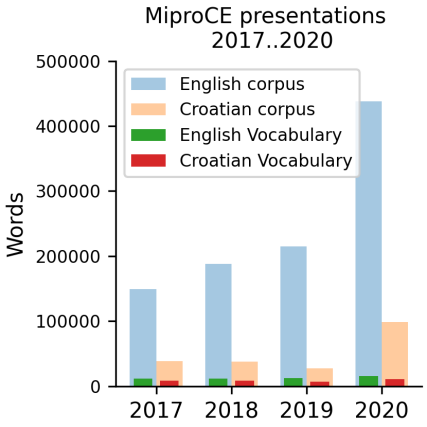


Figure 1. Number of words and size of vocabularies (unique words). It seems, that pandemic (in spring 2020) has essentially increased interest in making conference presentations

Another common statistics in text analysis is distribution of frequencies of words – this is considered one of the first important tasks in content analysis [5], basis of the Google search, basis for market trends analysis [6]. For frequency analysis are first from corpora removed stopwords – very frequent words ('*a*', '*the*', '*it*',,). The well-known package nltk (Natural Language ToolKit) lists 127 English stopwords [7] and for the Croatian language 177 stopwords [8]. From the remaining text corpus are counted unique words – the vocabulary of the corpus. Frequencies of essential words are shown in the following graphs. Notice, that in both languages the word with highest frequency is the same and even the frequencies of use of the word are close - authors of texts (presentations on MiproCE) belong to the same category of humans who are discussing the same topics.

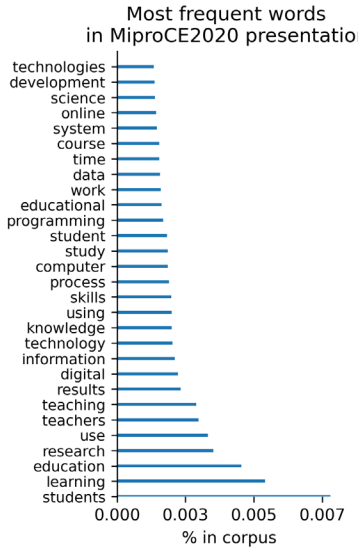


Figure 2. The most frequent word frequencies distribution in MipreCE2020 English-langauge track

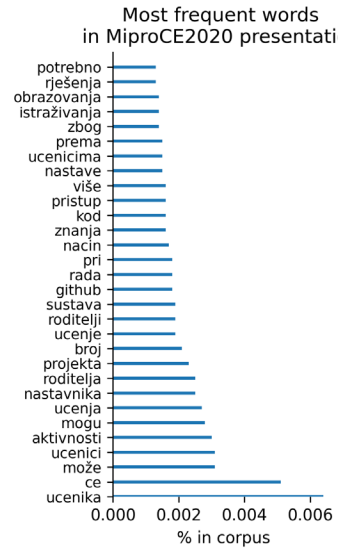


Figure 3. The word frequencies distribution in MiproCE 2020 Croatian-language track. The most frequent word is the same as in English-language track and even the frequences are close.

C. *Entropy*

The increase of disorder or entropy is what distinguishes the past from the future, giving a direction to time.
Stephen Hawking, A Brief History of Time

Text quality – how informative it is, how easy to read, its expressivity can be described by entropy of text units – entropy of single words, entropy of n-grams, entropy of phrases or sentences [9].

Entropy (on word level) of text T consisting of words w is calculated using Shannon's formula

$$H(T) = -\sum_{w \in T} p(w) \log_2(p(w))$$

Entropy shows average amount of information of words (in bits) – the most important aspect of communication and the same holds for other text units – characters (Shannon considered characters), groups of words, sentences or whole documents. Entropy of single words (unigrams), two consecutive words (bigrams), three consecutive words (trigrams) etc. show essential difference between English and Croatian languages.

	English	Croatian
Vocabulary (unique 'proper' words)	25677	27049
Stopwords:	131	181
Entropy of unigrams:	6.33	8.25
Entropy of bigrams	8.2	8.06
Entropy of trigrams	6.17	5.63
Entropy of 4-grams	4.76	4.26
Entropy of 5-grams	3.83	3.41

Research of world languages has shown, in all world languages entropy of words (unigrams) is greate than 6 bits per word. The Croatian language is more complex than English: more letters, more word forms etc., thus in studies its entropy has always been greater [10].

The above table reveals also a difference between English and Croatian languages – while entropy of single words (unigrams) in Croatian language is bigger, the relation quickly changes with longer expressions (bigrams-trigrams-… 5-grams).

Often occurring phrases can be found using the Pointwise Mutual Information (PMI) of an n-gram [11]:

$$PMI(w_1 \ldots w_n) = \log \frac{p(w_1 \ldots w_n)}{\prod_i p(w_i)}$$

The phrases with high PMI value are in English and Croatian languages different.

In English: "*does not*" (7.86), "*has been*" (7.76), "*Facebook, Instagram, Twitter*" (25.95), "*In this paper*" (15.28), "*as well as*" (14.23), "*can be seen*" (14.1), "*On the other hand,*" (24.36), "*in the field of*" (12.81) etc.

In Croatian: "*strucno usavrš1avanje*" (11.43), "*Visokog učilišta*" (11.27), "*Kljucne rijeci*" (11.02),

"*najmanje jednom, inace*" (25.81), "*izvorne engleske rijeci*" (24.57), "*Pitanje* NUM *Koliko cesto*" (29.7) etc.

In [12] was proposed keyword extraction based on entropy difference between the intrinsic and extrinsic mode, i.e. significant phrases occur in tight "dumplings' which are unevenly distributed, but common/insignificant phrases are evenly distributed in the whole text. Checking the n-grams with high PMI values did not show clear winners – the only candidate for a keyword phrase turned out the phrases "*Facebook, Instagram, Twitter*" and "*strucno usavrš1avanje*" (Professional Development).

Below are spectra of occurrence in the corpuses of both languages of some phrases with high PMI values.

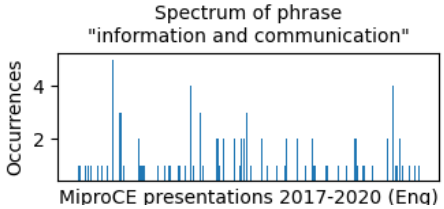


Figure 4. The phrase (E = 8.98, PMI = 11.28) also occurred nearly in every paper, but was not used as often as the previous phrase

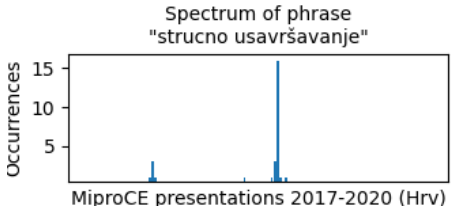


Figure 5. The phrase (E = 4.83, PMI = 15.76) is formally a good canditate for a keyword phrase

D. *Word embeddings*

The frequencies of words can be used for classification of documents, but more informative about word's meaning gives word's context.

Contexts are words which occur in a fixed length window together with the word; continuous contexts are usually called n-grams, non-continuous – bags_of_words (BOWS). In the MiproCE English corpus for the most frequent word "students" its most frequent (2,0)-contexts were: '*show that students'* – 0.125, '*percent of students*' – 0.111, '*showed that students*' – 0.091, '*with the students*' – 0.072 etc.

Most of frequent n-grams have a very limited meaning, for instance, the most frequent 5-grams created from the MiproCE English-language corpus were:

"*Facebook, Instagram, Twitter and YouTube*"

"*participants from the field of*"

"*Pitanje* NUM *Koliko često očekujete*"

"*za izraže koji se tiču*"

"za *nadprosječne učenike prosječne učenike*"

It is even for human rather difficult to get meaning of the text from these n-grams.

E. *Predicting the next word*

Information: the negative reciprocal value of probability.
Claude Shannon

The main problem of language understanding is prediction of the next word (or character). Text/corpus model is a collection of conditional probabilities of the next word in text.

Suppose we already have a sequence of words:

$$w_1, w_2, ..., w_{i-2}, w_{i-1}$$

The next word could be guessed maximizing the relative probability (the Bayesian inference [13]):

$$\arg\max(w_i \in V) P(w_i \mid w_{i-k}, ..., w_{i-1}) \qquad (*)$$

Here $P(w_i \mid w_1, ..., w_{i-1})$ is the conditional probability that after words $w_1, ..., w_{i-1}$ follows the word w_i. In practice probabilities are estimated from real-word frequencies, i.e. the relative probability of word '*students*' after the previous words '*all our*' could be calculated from the frequencies of use of these words in a (large) corpus of text where these words were already used:

$$P(w_i \mid w_1, ..., w_{i-1}) \approx \frac{Fr(w_1, ..., w_{i-1}, w_i)}{Fr(w_1, ..., w_i)}$$

In practice (to speed up calculations) the last formula is simplified even more. Using the naive Bayes conditional independence assumption that the probabilities $P(w_j \mid w_{j-1})$ are independent are in language models often used only binary probabilities (a very rough assumption), thus

$$P(w_1, ..., w_n) = P(w_1) \prod_{i=2}^{n} P(w_{i-1} \mid w_i)$$

F. *Generating (new) text*

Doing data analytics makes use of two skills: statistics, and telling a story with those statistics in ordinary words
Carl Howe, data scientist, RStudio, Inc.

The presented above argmax formula (*) can be used to create new texts based on probabilities occurring in the text corpus – give some words $w_1, ..., w_{i-1}$ as a seed and find a word w_i which maximizes probability $P(w_i \mid w_1, ..., w_{i-1})$, then shift the 'action window' one step to right and repeat the process starting with sequence $w_2, ..., w_{i-1} w_i$.

To use only binary contexts (i.e. the previous word) seems a rather rough assumption, thus we made some experiments in text prediction where for producing the most probable next word where first tried trigrams (two previous words) and if they could not produce a satisfactory continuation – the bigrams. Since the program

was maximizing the probability of the next word, i.e. used trigram/bigram with highest frequency, it quickly cycled, started to continue the same phrase (indicated below in bold), thus produced strings have a very low entropy (indicated at the end of text):

pedagogical decisions on using the Internet and the second part ***of the most important for the purpose*** *of the most important for the purpose* ***of the most important*** *...(E=3.45)*

To podrazumijeva upravljanje simulacijom putem lokalne instance GNSNUM poslužitelja. Uz postojecu VMWare Player aplikaciju, na simulacijskom racunalu omo gucena ***je lokalna propusnost "ICMP Echo Request" paketa u oba smjera aktivacijom predefiniranih pravila*** *"File and Printer Sharing (Echo Request - ICMP) – Domain /Public". Drugim rijecima, aktivacijom navedenog pravila Windows vatrozida omogucena* ***je lokalna propusnost "ICMP*** *...(E = 5.66)*

To get influence from farther earlier words are in NLP used contexts with lengths > 50 (words or characters – text processing has often done also on the level of characters), but this 'far influence' does not come from n-grams - words are considered as elements of vectors of probabilities of their context words. Mapping from words to vectors is called 'Word Embedding'. The dimensions of these vectors can be rather large, e.g. the Stanford collection of pre-trained word embeddings [14] dimensions vary from 50..600.

The word vectors are not unique – they depend on the text corpus and even with the same text corpus different NLP packages (different methods for creating text model) produce (somewhat) different results.

Representing words as vectors with real-valued coordinates allows to calculate from the cosines product of their vectors distance (i.e. similarity) between words Vocabulary of the whole corpus becomes a 'cloud' of dots in multidimensional space;

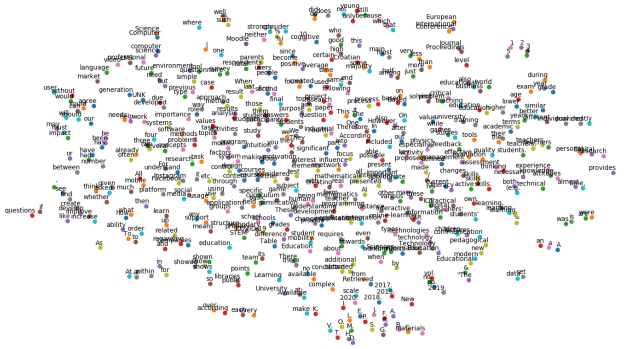


Figure 6. The word cloud created from the MiproCe 2020 English lanaguage presentations

Figure 7. The central part of the word cloud from MiproCE Croatian language presentations

Algebraic (vector) operations also show that word vectors reflect semantic relations between words, e.g. if $v(w)$ is the vector of a word w and $\approx$ denotes closeness, then

$$v(students) - v(student) + v(course) \approx v(courses)$$

- the first subtraction creates 'pure' plural which when added to word *course* creates vector for the word *courses*.

Zooming in reveals in the above picture several groups of words with close meaning:

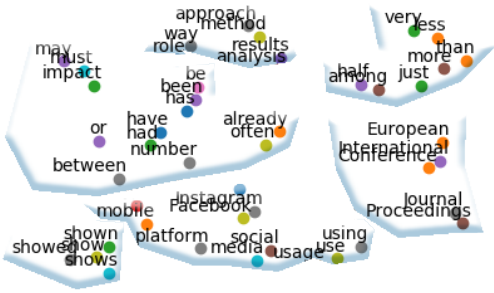


Figure 8. Some 'isles' of close words from the previous word cloud image. 'Instagram', 'Facebook', 'platform', 'social media' are semantically similar and incclination/case is not any more a problem – program has placed different lexical forms of the same word together – 'be', 'been', has', 'have', 'using', 'usage' etc.

Statistical methods become quickly too complex and computationally expensive to calculate high-dimensional word embeddings. High-dimensional embeddings are created by step-wise local information collection, imitating the method what (presumably) humans use for information processing. The high-dimensional word vectors (digital model of the text corpus) created by neural networks can be used for creating new text using word embedding's to predict next words.

There are already tens of such programs available, e.g. from the corpus of MiproCE English presentations tutorial program from [15] produced sentences, which on cursory inspection may seem already sophisticated enough to be accepted for the next MiproCE conference. The following examples are produced using text representation on character (not word) level, thus it is a miracle that grammatical concordance (case, singular/plural etc.) appears automatically (with some errors); after 3 minutes appeared (Windows 10, 4.12 GHz, 16 GB RAM):

University education should develop student's creativity. In the research activity and to the course is to explain the training of computer science courses and the number of formal education will be experiencing a basis for them to understand the tasks they are interested in the field of education and the students that will be the same as the correct solution of the 21st century.

Because of the training is constant for a solution for the teachers' point of view, where assignment needs the content of education and the students who have a very negative attitude towards the European Commission as a potential for the student and the results and analyze it statistically significant difference between the traditional teaching methods and viewed by the Fourth Industrial Revolution and its advantages and disadvantages of student motivation in the field of education and the other MS Teams are shown in Table II.

And so on and so forth – the length of the 'creation' can be set to whatever. There is no intelligence, no new ideas in produced text. Changing model parameters allows to create another 'high-scientific' output.

"*University education should develop student's creativity. Computer networking career and the comparison of the results also provides all localization in applying one system can be used for new rules. Also new ideas aimed attitudes towards understanding of the correct answer containing to access to the digital learning (local and script consumed in Croatian Science FCN game- generated experience of different CE Content-Based Programmes, Microsoft OneDDM) has been conducted on some of the most frequently evaluated e-learning software in the developed course …*

Ucenici su odgovorili na potrebe obrazovanja koje se odnosi se na potrebe studija. Sudjelovanje u svrhu potrebnih za ucenje programiranja u programiranju te na temelju podataka i studenata u nastavi informacijsko-komunikacijske tehnologije u nastavi informatike koje su nastavnici su odgovorili na pitanje koje se odnosi se na pitanje koje su potrebno postaviti standardni …

IV. WHAT IS ML ?

Every teacher knows what is learning – a process to improve, change learners behavior in order that learner can better respond to its environment, better achieve its tasks.

Computers are deterministic devices whose behavior does never change – is it does, then the computer is severely broken. When the same text corpus is re-used (with the same model structure) computer creates the same model and if it is used for text creation (with the same seed) appears the same text.

Thus the acronym 'Machine Learning' is actually a misuse of the word 'learning'.

In order to understand each other, we should have some common understanding of terms what we use, but there is lot of dissension in use of terms 'information', 'knowledge', 'learning'. Would you say that Newton learned the Law of Gravity or Einstein learned the Theory of Relativity? They did not 'learn' those laws, they discovered them setting up totally new frames of thought, performing experiments, what nobody had thought of before. They first created new mental approach, new framework, then observed, collected data in this framework and then generalized their observations data as a new Laws of Nature.

When humans speak/write, the next word also depends on all the already produced words, i.e. they use a procedure similar to rule (*) what computers use, but the process begins in their consciousness " :

$$\arg\max(w_i \in V) P(w_i \mid " , w_1, ..., w_{i-1}) \qquad (**)$$

The rule what computers use is only an approximation of the tail of the human's procedure. The premise $w_{i-k}, ..., w_{i-1}$ of the used in rule conditional probability is only a small tail of the premise $" , w_1, ..., w_{i-1}$ used by humans, thus the consequence w_i is less exact (its probability is smaller) and thus also the entropy (information content) of the whole produced phrase is smaller.

Word vectors (however long) can't express the meanings of words the way as we know them – we change them constantly. Depending on our mood, previous events, time of year/day etc. we can use the same words with quite opposite meanings: "*John, You did well!*" may mean ("*Good, we expected you to fail*") or ("*You failed, we expected you to win!*"). The current NLP research is trying to analyze sentiments (positive or negative) and some researchers even try to analyze more feelings [16], [17], [18]. But this (and many problems connected with memory) are difficult forms of verbal expression and difficult to re-produce – computers (yet) do not have feelings and do not know, what to remember - is the word *Rijeka* a name of a student, bird, virus or programming language and should it be stored in memory?

And here lays the main, most important difference between Machine Learning (ML) and Human learning (HL). Machine Learning in NLP is an approximation of the tail (visible) part of human communication.

The NLP text models can make everything looking like truth. One of (currently) biggest models, the GPT2 accepted the following fable [19]:

"In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English."

The GPT-2 system continued the fable to look like a true story from some news agency:

"The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.…"

If a program can fluently explain *four-horned silver-white English-speaking unicorns* then it certainly can also prove that *Earth is flat*, *vaccines and 5G are evil* etc. - a

perfect creator of 'fake news', but these news are 'fake' (or in modern terms. 'alternative truth') just because they are not provable.

V. DATA SCIENCE IN "COMPUTERS IN EDUCATION"

The first step in any research begins with collecting some statistics to reveal the most important features of this area and presenting this in tables.

The search for the word 'Table' returned in MiproCE2020 text corpus:

```
Search "Table" (10968 hits in 68 files of
68 searched)
```

Thus the frequency of the word '*Table*' was in average in every presentation :

431535/10968 = 39.3441247

i.e. nearly 40 times.

This evokes a question – is the topic of MIPRO CE statistics?

Earlier were here presented tables of frequencies of words in both, English and Croatian-language text corpuses. The clear indicator of data handling, the percentage sign "%" usually does not occur as a separate word, but only together with some number, thus the symbol did not appear in search for words. Therefore the MiproCE presentations 2017-2020 were separately searched also for occurrences of the "%" sign. It turned out, that more than half of presentations used the "%" tens of times – clear indicator that these presentations were dealing with Data Science:

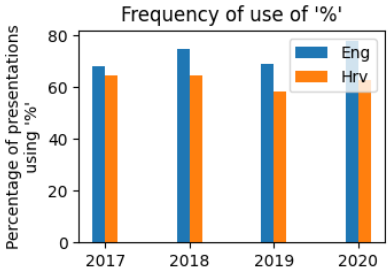


Figure 9. More than half of presenations where using the "%".

And the use was extensive – in average, ca 20 times in a presentation (at least, since often meaning of numbers in a table also was percent, if the column/row header was 'Percentage of …').

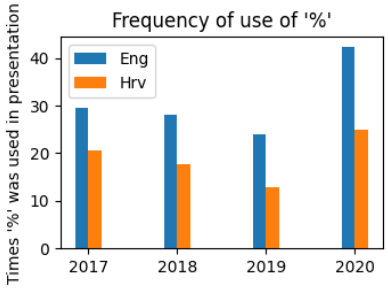


Figure 10. Number of times the character "%" was used in a presentation.

Therefore become interesting – is there a correlation between use of the word "Table" and character "%" ?

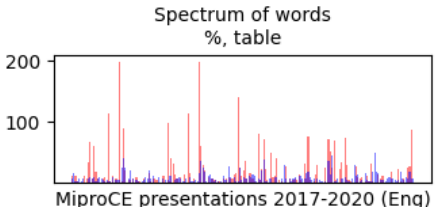
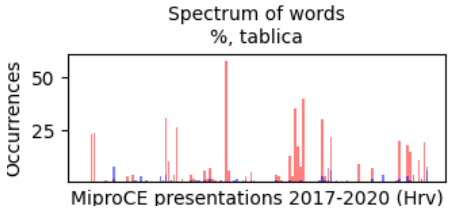


Figure 11. Spectrum of occurrence of words 'table' and character '%'

The Pearson correlation between lists of occurences was (0.173, 0.018)

In MiproCE Croatian-language presentations the picture was similar, but correlation stronger.



Correlation: (0.204, 0.023)

Data is collected in order to learn something, to introduce changes in existing practices. The tables and percentages in MiproCE presentations should also be used for introducing some improvements, changes.

To study this aspect we collected from the Wordnet [20] - a database of English words that are linked together by their semantic relationships - all words belonging to synset (synonyme set) of concept "*change*":

'change', 'alteration', 'modification', 'variety', 'alter', 'modify', 'vary', 'switch', 'shift', 'exchange', 'commute', 'convert', 'exchange', 'interchange', 'transfer', 'deepen'.

If a presentation collected some data and suggested some changes to existing procedures then it most probably should also use one of these words. But this could occur only once or twice (in Introduction and in Conclusions), what was seen also from frequency of occurrences of these words:

'change' - 0.00022, 'variety' - 0.0002, 'transfer' - 0.0001

Thus more than half of presentations in CE are dealing with Data Science.

VI. CONCLUSIONS

Achievements in development of digital communication tools have been rapid. AI bots like Siri, Cortana, Alexa, and Google Assistant in our mobile phones, chat robots in banks and large companies websites – they all provide many useful services and behave already like half-humans. Many of achievements in NLP, ML, AI etc. will be certainly applied also in education.

As with use of every very high-level technology there are also several possible risks. When the currently one of

the largest natural language models GPT-2 (1,500,000,000 parameters) was released [21], its creators warned:

"…extremist groups can use GPT-2 for misuse, specifically by fine-tuning GPT-2 models on four ideological positions: white supremacy, Marxism, jihadist Islamism, and anarchism … it's possible to create models that can generate synthetic propaganda for these ideologies".

But the truth value of texts created by NLP packages is very questionable – computer can not evaluate its statements. The same Tensorflow program which has been used for all examples in this paper produced from MipreCE English-language corpus following assertion:

Students are not included in the process of e- learning in the context of the process of teaching and learning.

And another text-creation package just revealed that *Darwin met the silver-white English-speaking unicorns with four horns on the board of 'Beagle'. But unicorns agreed to fly to 'Beagle' only on condition that Darwin signs non-disclosure agreement, thus Darwin did not mention them in his book. Handwritten (unsigned) notes about this meeting were found in Darwin's papers and sold on Christie for 50 ml USD to unnamed byer.*

REFERENCES

[1] eLife. Comprehension of computer code relies primarily on domain-general executive brain regions. https://elifesciences.org/articles/58906

[2] MIT News. https://news.mit.edu/2020/brain-reading-computer-code-1215

[3] Black, E. et al. Towards history-based grammars: Using richer models for probabilistic parsing. Proceedings of the Fifth DARPA Speech and Natural Language Workshop, Harriman, NY, 1992

[4] Some of my Best Friends are Linguists. http://www.lrec-conf.org/lrec2004/doc/jelinek.pdf

[5] P. Nulty. Semantic/Content Analysis/Natural Language Processing. DOI: https://doi.org/10.1007/978-3-319-32001-4_182-1

[6] Jefrey Lijffijt, Panagiotis Papapetrou, Kai Puolam¨ aki, and Heikki Mannila. Analyzing Word Frequencies in Large Text Corpora Using Inter-arrival Times and Bootstrapping. ECML PKDD 2011, Part II, LNAI 6912, pp. 341–357

[7] NLTK's list of English stopwords https://gist.github.com/sebleier/554280

[8] Ranks NL. https://www.ranks.nl/stopwords/croatian

[9] C. Bentz, D. Alikaniotis, M. Cysouw, R.F.Cancho. The Entropy of Words—Learnability and Expressivity across More than 1000 Languages. https://www.mdpi.com/1099-4300/19/6/275

[10] M. Ignatoski, J. Lerga, L. Stankovic, M. Dakovic. Comparison of Entropy and Dictionary Based Text Compression in English, German, French, Italian, Czech, Hungarian, Finnish, and Croatian. July 2020, DOI: 10.3390/math8071059

[11] G. Bouma. Normalized (Pointwise) Mutual Information in Collocation Extraction. Proceedings of the Biennial GSCL Conference, 2009

[12] Yang, Zhen, Lei, Jianjun & Fan, Kefeng & Lai, Yingxu. Keyword extraction by entropy difference between the intrinsic and extrinsic mode, Physica A: Statistical Mechanics and its Applications, 2013, Elsevier, vol. 392(19), pp 4523-4531.

[13] D. Jurafsky, J. H. Martin. Speech and Language Processing. web.stanford.edu/jurafsky/slp3/4.pdf

[14] GloVe: Global Vectors for Word Representation. https://nlp.stanford.edu/projects/glove/

[15] Text generation with an RNN. https://www.tensorflow.org/tutorials/text/text_generation

[16] L. Kerkeni et al. Automatic Speech Emotion Recognition Using Machine Learning. DOI: 10.5772/intechopen.84856

[17] D. Zhang et at. Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19), pp. 5415-5422

[18] M. Wadhwa et al. Speech Emotion Recognition (SER) through Machine Learning. https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine-learning/

[19] OpenAI. Better Language Models and Their Implications. https://openai.com/blog/better-language-models/

[20] WordNet. https://wordnet.princeton.edu

[21] GPT-2: 1.5B Release. https://openai.com/blog/gpt-2-1-5b-release/