International Journal of Artificial Intelligence and Neural Networks– IJAINN
Volume 5 : Issue 1 [ISSN : 2250-3749]
Publication Date: 30 April, 2015

Limits of information retrieval using basic natural language processing aproaches

Reasearch focused on used words for summarizing the key message of given text by readers

Radim Brixí

***Abstract*—The paper presents original results from research that analyses words that are used by the readers of given text to describe the key message in the entire text. Those readers' words are then lemmatized and compared to lemmatized words from the original main text. The research reveals that for texts with deeper meaning are used surprisingly and also significantly such words by the readers that cannot be found in the entire original text. The percentage of such words increases when forced to express the key message using less words.**

***Keywords*—word, limit, natural language processing, lemmatization, meaning, summary**

I. Introduction

Even though we live in the age of technology, understanding the written text is still a main domain of man although significant improvements are emerging in complex information systems. Specific human domains like [1] seem to trouble architects of algorithms simulating human behaviour because some are not algorithmizable at all, similar issues can be discussed with consciousness topics when designing artificial intelligence approaches [2]. Hardly possible algorithmization of specific issues applies also to areas like information retrieval and natural language processing (i.e. irony).

This paper focuses on specific summarizing topic. That is an analysis of summarization of the most important key message of a text with deeper meaning by humans and identifying limits for current natural language processing approaches. The research was performed using the novel "The little prince" from Antoine de Saint-Exupéry [3]. Although the entire research was designed and realized in the Czech language with Czech version [4] of the novel with group of 176 native Czech probands (experimentee person), it is very reasonable to have an assumption that the output from such a research should have very similar results and outputs even if it would be realized in other languages. The similarity is assumed in a sense that for every reader might be important different thing according to one's knowledge, background, preference and therefore the research should focus on detection whether there is a difference (and also how big) among the words used by readers to describe the key message and the whole original text.

Radim Brixí
University of Economics, Prague
Czech Republic

The main reason for the research is assumption that when one is forced to express the main idea, key message or most important thing or essence of the whole text in a limited number of words, than one is forced to pick up the main core idea that seems to be somehow hidden in the text behind the lines instead of the description. This is based on the assumption that readers use their individual knowledge and may accent different things to be important for them. Cultural effect might slightly shift some meanings but the revealed principle from this research is assumed to be applicable internationally and therefore we assume that the output of this research can be generalized to other languages only with insignificant exceptions.

The other reason for the research is also the fact that if we would use much different words that are not derivable from the original text itself, it should lead the research interest of natural language processing approaches more to such approaches focusing more on the deep understanding or at least pseudo-understanding, eventually on machine learning with analogy of humans knowledge gathering and meaning understanding.

In the real word complex applications where the intention is to use natural language processing to obtain the summary or main information from larger texts we face some difficulty when the text has deeper meaning hidden behind the lines such as in the novel The little prince [3]. If we rely on machine/computer based summarization in such cases we can easily miss "the point", because the humans may see "the point" in a very different thing or idea (also very differently among each other) than the algorithmically summarized version would offer (basically based on any kind of approach excluding meaning or understanding).

The main research idea is to compare what is important for humans from the given text and whether all the words used by humans to identify the key message can be found in the original text or just part of them. The research was designed to measure whether we use the same words as given in the entire text to express the key message (most important meaning) or whether we use some percentage of such words that identify the most important meaning and that are not in the entire original text and therefore cannot be obtained by basic operations like filtering, subset, synonyms or any kind of combination of these basic operations or similar in principle. The research output should give push to new approaches in the natural language processing area that do not rely only on basic text operations as described above with desire to prove that for some texts' summarizations especially for texts with deeper meaning performed by humans will be unreachable by basic

natural language processing algorithms because of incapability to understand the main most important idea actually somehow hidden in the text as humans would understand it. The research was done quantitatively with applicable statistics, but also qualitatively in a very large scale analysing all words from every person for every version of summary. This paper can provide due to page limits only aggregate output from this deep research, but detailed data are of course available for anyone who is interested for specific details of this research.

II. Construction of research

A. Hypothesis

The main hypothesis of the research can be declared in the following form: If we implement such a function of a program that will make subset of the whole entire given text by predefined algorithm (using sorting, filtering, selection, synonyms, subsets or similar operations in principle in any combination) in order to find the essential (key message or most important), there is no guarantee that the function will be able to return such information from the whole text that would be important for the human reader (because important portion of such words may not be in the entire text at all).

Other research questions can be that forcing the reader to summarize large text in a very sort form like 13 words forces one significantly to skip the longer describing approach of summarization and forces one to formulate directly the key essence or message hidden behind the lines. (Such an important thing may vary across readers but may be of course especially valuable and exceptionally important.)

Finally the quantitative research question is how many words (in percentage) that were used by readers cannot be found in the entire original text and how it differs in case of forced summarization in longer version of max 111 words summarization and shorter version of max 13 words summarization.

B. Design of performed research

The research was designed in such a way that 176 readers were supposed to read the entire novel "The little prince" [4] and write in 111 words what was it about. Later all should write what was it about but only with maximum 13 words (additionally other tens of another probands were supposed to read another 4 different novels with the same task to prove that the observed principle being researched emerges automatically in other similar cases to prove back the principle in order to prevent this novel to be an exception).

Then the goal was to compare the words used by each proband with the entire text of "The little prince" [4]. The comparison in the Czech language faces many difficulties like falls, conjugation, word-formation, folding, intensification, homonyms, etc. [5]. For that reason all sentences must be processed through lemmatization process and for that transformation of sentences and words was determined online lemmatization tool [6][7] as adequate.

Before the short and long versions of probands can be lemmatized, several initial corrections need to be done (like correcting typographic and evident mistakes in spelling, remove special characters, removing invalid samples from different set of reasons, etc.

After lemmatization of all short and long versions, it is necessary to lemmatize also the entire original text of novel.

When lemmatization of each text is done, all words are compared in all versions with the original text and those, that are used by readers but are not present in the entire original text are marked bold and counted.

Finally the percentage of used words for each variant (short with 13 words and long with 111 words) that are not present in the entire original text is calculated and the results are compared. Supporting the numbers of quantitative research also qualitative analysis of texts is performed to understand the reasons for the results with abduction (in the sense of scientific concluding/reasoning method) of the reasons for such results.

Aggregate set of words used especially only in the short version that are neither in the longer version nor in the entire original text is constructed using implemented programs for deeper analysis.

Conclusions and statements regarding hypothesis and research questions are formulated.

C. Course and results of the experiment

According to the previous section the number of responses that were valid and usable for research was reduced to 157 responses with two variants (13 and 111 words) per proband's response due to clearing and isolating invalid responses (like responses in Slovak language, removing incorrect forms of answer, partly saved questionnaires, etc.).

After lemmatization and comparison with original text surprisingly 19,06% of all words in the 111 words version of human summarization cannot be found in the entire original text. In case of the 13 words version the percentage rapidly rises to 25,92% and therefore every fourth word describing the important essence of the given text cannot be found in the original text (we must also face the fact that although a big portion of the rest of the words can be found in the entire text, when analysing more deeply, we observe that the words were used in different perspective and meaning so deriving the key message from these words would be also difficult). This percentage supports the hypothesis very much indeed because when qualitative analysis is performed on every single sentence from the readers we realize that it is not evidently possible to algorithmically get to such words that were used by human readers in case of methods manipulating with text described in the hypothesis. The percentage rise also supports the other research question that forcing the reader to express in 13 words forces one significantly to express the essence or main idea instead of description what happened in the story.

The following words were translated from Czech language from quite specific forms to most equivalent English meaning with some loss in translation in order to capture the set of

those words that were used only in the 13 words version by the readers and are not present in the Czech version of the novel in the entire original text. These words are according to one of the research outputs: "knower, morale, exceptionality, say regression, faith, couple, look, meritorious, step, indicating, emotion, shapes, forgive, insignificance, egocentrism, materialism, rationalism, racing, viewing, talks, not only, lived, ET, more, approaches, readable, terraced, served, incorruptibility, brink, pointless, telling the, aspects, melodrama, self-knowledge, integrity, reflects, oppressed, magic, encourages, nice, meaningful, met, discrimination, continues, contradictions, awareness, incorruptibility, humility, telling, alien, revealing, dispatching, weigh, invincibility, resilience, illustrated, sensitive, cognitive, excessive, self-referentiality, narrowness, drives, cognition, introduction, existential, autobiography, survival, limitations, futility, childish, task, working, child, does not make, liberate, use, touch, interpret, sorrows, uncomprehending". All of these words are marked and identified in all the 157 thirteen words versions proving that the human key point is mainly focused on the idea behind the lines of text and doesn't follow the story that much, because the story seems to be just a tool for humans to come up with the revealed hidden message.

If we pick up a random example of 13 words variant, we get after translation into English: "**Significance** of beauty, love, true **friendship** and **integrity** of **human** life in endless space." Another random example can be: "The book **criticizes** the **current lifestyle** and **encourages** a **stronger perception** of the world around us." Please note that the original Czech version is really based on 13 words. It changes due to translation process. The bold (originally Czech) words determine words that were not in the entire text of the novel in the Czech version.

If we analyze deeply what is accented by the humans then we must agree that the construction of the key message is based on individual knowledge and preferences based on what is essential (substantial) for each reader. In other words lots of the short variants tend to be some kind of provoking to action or to see different perspective or in some way something is pointed out based on a wide text with a story where behind the lines is hidden another message for which interpretation process has many attributes (like age when the novel is read) and plays an important role and therefore is so different for every person. At this point humans are very original at pointing out the message behind the lines although it varies depending on process of interpretation of the metaphor.

This leads us to serious conclusions regarding limits of natural language processing approaches based on the methods described above within the hypothesis, because for decision making for example in business area facing problems with fast technology grow, innovations and the difference between humans and computer algorithms [8] it is important to deeply understand increasingly growing amount of data, large scale texts, documentation, multimedia, etc. in order to be able to do good managerial deaccessioning based on true understanding with ability to point out the truly important aspect of given texts/data in the information systems.

III. Conclusions

The potential of natural language processing and all similar semantic approaches that are not based on the true and deep understanding and meaning of the text has limits of impossibility to get to the human way of understanding the meaning as described and experimentally proved by the large scale qualitative and also quantitative experiment with reasonably acceptable sample of 157 versions of human summarizations of novel "The little prince" [3] especially in those cases, where other meanings flow hidden within the text and the meanings are also different relatively to age and knowledge background of the reader. The results of experiments were verified also on several different novels, so in principle the key message for humans seems to be unreachable by approaches based on methods similar or analogic to those defined in the hypothesis section in similar cases.

Although common summarization tools seem to be very successful and widely used even by consumers and end users, the limits of the approach or algorithmic summarization were proved to be real because the really essential point can be missed due to too much "focused on the trees without seeing the point of the forest" in another words focusing on the story of the given text by algoritmization approaches may not inform us about the key essential message of the given text, because the message itself emerges in the head of the readers based on the individual knowledge and preferences of the reader and is not stored in the data itself in a kind of point of view where we use our brain, consciousness and knowledge that cannot be used during algorithmic approaches based on basic operations with the text.

Individual knowledge is very powerful reason (age and other attributes are also important) for having so much different key messages from all probands that we can clearly claim that we did not obtain two same responses. So our personality is very much included in the process of interpretation of the text and therefore we must admit that all the texts saved in all databases represent just a partial subset of all the components that are necessary to obtain the understanding of information. For summarization it is a higher level of process of interpretation and therefore the variability of responses rises.

The results of this experiment should be taken in account in case of more explicit texts as well as the process of interpretation based on experience and knowledge is similar in principle although hidden messages behind the lines might not be that much present in other texts. In case of decision making based on information, humans attach importance differently to different information and different people may react in the manager role differently based on the same information where the meaning might be interpreted differently. This is another aspect that challenges natural language programming approaches not to be rigid in only one context we encourage to implement variability of interpretation mechanisms based on different of sets of knowledge, background and parameterized preferencing.

Acknowledgment

Detailed data and other experimental results are available for deeper analysis at.

References

[1] R. Brixí, "Human Prediction of Computer Generated Value Based on Statistical Experimental Approach Intuition in Human-Computer Interaction Research," in *2nd International Symposium on Computer, Communication, Control and Automation*, 2013.
[2] R. Brixí and S. Brixí, "The Association Reactor in a Human-like Robot Architecture Concept: An Interdisciplinary Approach," *Spons. IDIMT-2009*, p. 297, 2009.
[3] A. de Saint-Exupéry and I. Testot-Ferry, *The Little Prince*. Wordsworth, 1995.
[4] A. de Saint-Exupéry, *Malý princ*. Albatros Praha, 1972.
[5] "Lematizace," 18-Jul-2014. [Online]. Available: http://www.eridanus.cz/id32402/jazyk/Lematizace.htm.
[6] J. Hajič, "Morphological Analysis of Czech Word Forms," 2013. [Online]. Available: http://quest.ms.mff.cuni.cz/morph.
[7] J. Hajič, *Czech Morphological Analyzer v1*. Charles University in Prague, UFAL, 2014.
[8] S. Mildeova and R. Brixi, "Technical Limits of ICT for Enterprises' Innovations," *J. Syst. Integr.*, vol. 3, no. 1, pp. 45–53, 2012.

About Author (s):

It is quite impossible to find words used for expressing the summarization of most important things from the given text in the text itself! Because the interpretation process within our brain and knowledge reveals the true key information from data even though it is not present in the data. What a limit and also challenge for NLP.