# A Turing test to evaluate a complex summarization task

Alejandro Molina[1], Eric SanJuan[1,2], and Juan-Manuel Torres-Moreno[1,2,3]

[1] LIA, Avignon University,
339 chemin des Meinajaries, Agroparc BP 1228, F-84911 Avignon Cedex 9, France
alejandro.molina-villegas@alumni.univ-avignon.fr
{eric.sanjuan,juan-manuel.torres}@univ-avignon.fr
[2] Brain & Language Research Institute,
5 avenue Pasteur, 13604 Aix-en-Provence Cedex 1, France
[3] École Polytechnique de Montréal,
2900 Bd Edouard-Montpetit Montréal, QC H3T1J4, Canada

**Abstract.** The paper deals with a new strategy to evaluate a Natural Language Processing (NLP) complex task using the Turing test. Automatic summarization based on sentence compression is a very complex task because it requires to asses informativeness and modify inner sentence structures. This is much more intrinsically related with real rephrasing than plain passage extraction so new evaluation methods are needed. We propose a novel imitation game to evaluate Automatic Summarization by Compression (ASC). Rationale of this Turing-like evaluation could be applied to many other NLP complex tasks like Machine translation or Text Generation. We show that a state of the art ASC system can pass such a test and simulate a human summary in 60% of the cases.

## 1 Introduction

Alan Turing predicted that computers will be better at playing complex board games like chess than to chat with humans in an open world. Natural Language Processing (NLP) appeared in 1951 to be one of the greatest challenges for computers. Surprisingly, some tasks like automatic summarization appeared to be easier than anticipated when considering extracts instead of abstracts [1]. Summarization by extraction often consists in segmenting the text to be summarized into sentences and to apply scoring methods to rank sentences by decreasing informativity. In this simplified task, resulting short summaries are often readable because they use real sentences. The main difficulty when dealing with longer summaries involving ten or more sentences is to avoid breaking anaphora. This is handled using simple heuristics like displaying top ranked sentences in the order they appear in the original text. Since local text grammaticality is ensured by keeping entire sentences, resulting summaries often give the illusion that they were written by a human. Moreover, under the assumption that the produced summary is readable, summary informativeness can be evaluated using measures like ROUGE given on a set of reference summaries or Jensen-Shannon metrics if no reference summary is available [2].

The task becomes much more complex if computer are asked to cut and compress sentences like humans do since this implies to understand and modify inner sentence structure. Discourse structure among other implicit semantic relations play a key

role [3]. Moreover there are usually several correct ways to compress a sentence and human experts often disagree on which is the best one. When trying to build a reference corpus of compressed sentences, inter agreement between annotators is low, even to decide if a sentence should be shortened in the summary or not. Automatic Summarization by Compression (ASC) requires to handle a high level of incertitude in the decision process since there is not a best way to compress a sentence, only observations that sometimes humans prefer one way rather than another one [4]. Not only the task itself is difficult but it cannot be evaluated using existing methods. Using sentence compression to produce a summary not always improve informativeness scores and can produce unreadable summaries. Therefore, actual state of the art evaluation metrics for automatic summarization discourage thorough investigations if a computer can handle or not ASC.

In this paper we show that coming back to original idea of a Turing test, it is possible to set up a simple imitation game to evaluate ASC. We also show that a state of the art system that learns human behavior using simple regression analysis [4] can pass this test on short summaries and give the illusion to human referee that the summary was written by a human. Moreover this imitation game is clearly adapted to crow sourcing through Internet and can be used to evaluate large amount of systems at a reasonable cost.

The rest of the paper is organized as follows. Section 2 goes back to the general definition of a Turing test. Section 3 details the imitation game that we propose to evaluate ASC in a pragmatic way. Section 4 shows statistical evidence that a state of art ASC system can pass the test. Finally, section 5 opens perspectives on how this evaluation methodology can contribute to the improvement of effective ASC systems.

## 2 Back to Turing test

As suggested by Alan Turing, a test to evaluate the ability of a computer to handle a human mind task should involve:

- an interaction with humans where the computer tries to give the illusion that it is human,
- a clear evaluation metric that allows the reproducibility of the experiment,
- a gateway to the open world to explore beyond restricted contexts and closed world assumptions.

Our main motivation relies on the fact that, to the best of our knowledge, there is no summarization evaluation methodology that encourages research on advanced NLP tasks like summarization by sentence compression. We therefore suggest to come back to Turing's initial motivations[5] when imaging imitation games to answer the controversial philosophical question "do computers have a mind ?" without having to define what "mind" means. The question then becomes "what are the common human intellectual tasks that a computer can handle? ". These are the roots of theoretical computer science where tasks almost useless for technical applications can be fundamental to understand computers's real limits. ASC can have many applications in our interconnected

world but we claim that its main interest relies on the theoretical study of computer capabilities.

In a the original imitation game defined by Turing in [5], there are two players and one assessor. The first player is a human ($A$) and the second a computer ($B$). Another human ($C$) plays the role of the assessor and has to guess the real nature (human or computer) of the two other players. The assessor cannot see the other players, he can just interact with them through a more or less restricted interface that at least allows to exchange written messages. The assessor asks questions through the interface and has to distinguish between answers given by the human player and those sent by the computer.

Turing imagined advanced imitation games to study the spectrum of Artificial Intelligence and compare it to Human mind. However, as pointed out by [6], Turing entrusted interaction through natural language. In our case, we intend to study the method of interacting itself related to NLP and its linguistic functionalities based on summary generation. Indeed, in the general case of a Turing test, the assessor is not allowed "to see" the players. This is to ensure that he focus on functional aspects and not on appearances. It then seems natural to adapt the imitation game to NLP tasks that try to reproduce human ability to handle texts like summarization. We do not consider tasks that cannot be carry out without computer assistance like Information Retrieval from large collections. Only intellectual tasks that can easily be accomplished by non experts meanwhile there are real challenges for an automatic system.

## 3 Imitation game to evaluate ASC

We consider the following imitation game involving a human player ($A$), a computer ($B$) and a human assessor ($C$). $C$ is not impartial and teams up with player $A$ against $B$ but $B$ has acquired some knowledge about the way that $A$ writes and $C$ reads. $C$ choses several texts (12 in our experiment) and ask both $A$ and $B$ to write a summary of these texts. After some time $C$ receives only one summary of each text. Half of these summaries have been written by $A$, the other half have been automatically generated by $B$. The interface between $C$ and the players has dispatched texts at random, just checking that each player receives the same number of texts. So $C$ does not know who between $A$ and $B$ wrote each summary and has to guess the correct author for each text, i.e. which are the summaries written by $A$ and those by $B$. If he correctly identifies most of the summary authors (60% in our experiment) then $A$ and $C$ win, if not $B$ wins.

This setting follows Turing's idea of an interactive game between two humans and a computer. However, one difficulty to carry it out is that human need time to write a summary meanwhile it is necessary to reproduce the same experiment at least 30 times to expect some statistical evidence if there is a regular winner between $A$ and $B$. To adapt this game to standard crow-sourcing evaluations, we decided to consider a team of $n$ extra assessors ($C_1, ..., C_n$) and six different human players $A_1, .., A_k$ to write the summaries. All humans participants expect $B$ to fail. In this new setting, $C$ choses the texts but he does not read the returned summaries. Instead he gives a copy of the summaries to the $n$ helpers. The main drawback of this crow-sourcing adaptation is the lack of real interactivity. The advantage is to allow a statistical analysis of results.

Let us give some details about the way we tried to implement this game to evaluate an ASC system. The authors of the ASC systems are not involved in this process. 12 texts were selected from the RST Spanish Tree Bank[7] at random. Summaries of these texts have been written down by post graduate students in linguistics from the UNAM. We chose as assessors $C_1, ..., C_n$, 54 other post graduate students from the UNAM that were not involved in the writing of the summaries.

As non human player $B$ we chose the ASC system derived from [4]. This systems is based on a regression analysis of the way that assessors $C_i$ agree or not with a sentence compression based on a discourse analysis. Indeed, it has been shown in [3] that humans tend to remove complete discourse units from sentences when they try to compress them. In [4] two Discourse segmentors DiSeg [8][4] and CoSeg have been used to generate compressed sentences extracted from the Spanish RST Tree bank. UNAM post graduated students were then asked to decide which compressions were acceptable. As anticipated for a so subjective task, inter agreement between assessors was very low but enough to carry out a regression analysis and learn to predict the probability of a particular sentence compression to be accepted by humans. Three summaries of different length (short, medium and long) were generated using DiSeg, and three other ones also of different length were generated using CoSeg.

## 4 Results

60 post-graduated students accepted to participated in this simulation game, 6 of them were asked to write summaries $t1, ..., t6$ and the 54 other participated as assessors $C1, ..., C54$. Two algorithms CoSeg ($c$) and DiSeg ($d$) have been used, and each of them have been used to generate summaries of three different length. All assessors read the 12 summaries and for each they tried to guess if the author of the summary was a human or a computer. They did not know that exactly half of the summaries were automatically generated. Moreover automatic generated summaries were of three different length (1:short, 2:medium, 3:expanded). Therefore, there were 66 players involved in this game, among them 6 were non human $c1, c2, c3, d1, d2, d3$. Figure 1 shows the results of the simulation. In this figure, bars for summaries written by humans $h1, ..., h6$ are expected to be higher if they are good quality summaries meanwhile bars for automatic summarizers $c1, .., c3$ and $d1, ..., d3$ are expected to be low since they intent to mislead the assessor. It appears that over the six authors of summaries, only three manage to write summaries that more that 60% of the assessors think they can not be automatically generated. Meanwhile automatic system DiSeg manage to mislead the assessors on 60% of long summaries (d3) and CoSeg system on short (c1) and medium (c2) summaries. Plot 2 shows the median normalized frequency of times that an assessor thinks the summary has been written by a human. The first boxplot shows it over the twelve summaries each assessor has to read. The second and third boxplots over the three summaries generated using CoSeg and DiSeg respectively. The fourth is over the six summaries by humans and the last one is restricted to the three best authors $h2, h3, h4$. These boxplots suggest that summary quality by three best authors (last box-

[4] http://diseg.termwatch.es

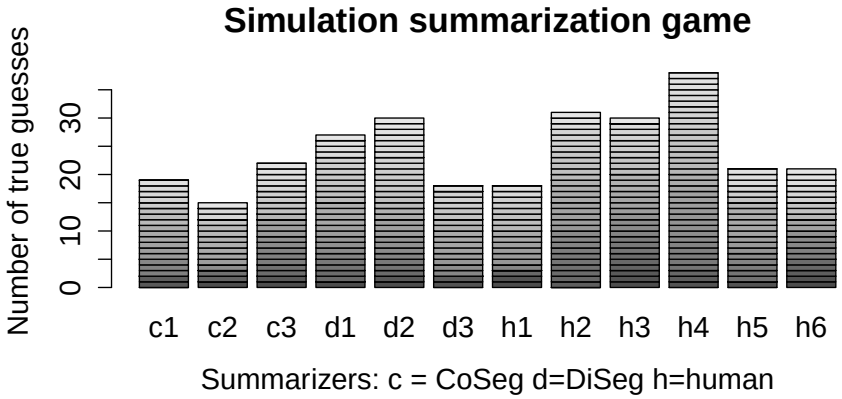


**Fig. 1.** summarization simulation game: each bar shows the number of correct guesses (Human or Computer) for each summarizer

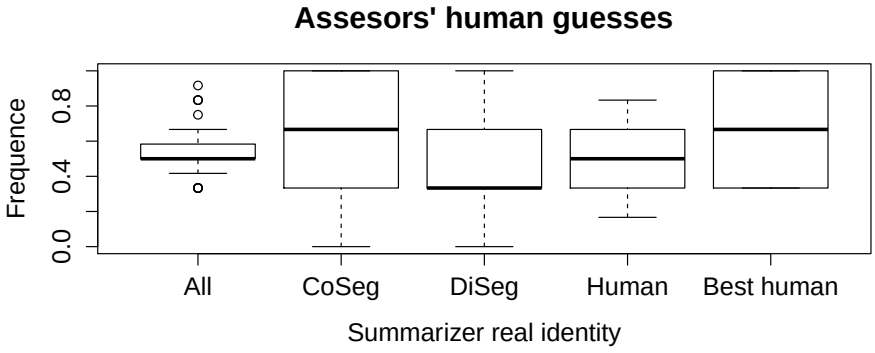


**Fig. 2.** Boxplots showing the median number of times that an assessor thought it was a summary produced bay a human for each set of six summaries and each subset of three summaries automatic/human.

plot) is above average among summaries written by real authors (fourth boxplot) and among overall summaries (first boxplot). However, according to a Wilcoxon test with a $p$-value lower than 0.01, only the differences between best human summaries and all human summaries is statistically significant. The difference between best human summaries and overall summaries is not. Similarly, CoSeg summaries outperform DiSeg summaries since the median frequency it misleads assessors is significantly higher ($p$-value $< 0.05$) meanwhile all other differences are not statistically significant. In particular there is not statistical evidence based on Wilcoxon rank sum test with continuity correction that an assessor thinks that the summary has been done by a human author when reading a summary generated by one of the automatic summarizers tested here, than one really done by a human author.

## 5 Discussion

Back to Turing's idea of simulation game, this paper uses crow-sourcing to simulate a simulation game to evaluate two state of the art automatic summarizers. Usual evaluation protocols failed to differentiate between quality levels among the two system outputs. The experiment set up here with 60 human players gives statistical evidence that one outperforms the other. But we also find out that the human ability to differentiate between a summary automatically generated and summary written by an author is less than expected on such complex task as ASC that goes beyond sentence extraction and ranking. Indeed, in ASC the machine has to built new sentences by compressing existing ones. This last finding needs to be checked out by setting up a larger crow-sourcing task. However, mixing human and machine outputs in the evaluation process seems to be a promising way to improve discriminative power of evaluation protocols based on crow-sourcing.

## References


1. Tratz, S., Hovy, E.: Summarisation Evaluation Using Transformed Basic Elements. In: Proceedings of the Workshop Text Analysis Conference (TAC'08), Gaithersburg, MD, Etats-Unis (2008)
2. Louis, A., Nenkova, A.: Automatically Evaluating Content Selection in Summarization without Human Models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Singapour, ACL (6-7 Aout 2009) 306–314
3. Molina, A., Torres-Moreno, J.M., SanJuan, E., da Cunha, I., Sierra, G., Velázquez-Morales, P.: Discourse segmentation for sentence compression. In Batyrshin, I.Z., Sidorov, G., eds.: MICAI (1). Volume 7094 of Lecture Notes in Computer Science., Springer (2011) 316–327
4. Molina, A., Torres-Moreno, J.M., SanJuan, E., da Cunha, I., Martínez, G.E.S.: Discursive sentence compression. In Gelbukh, A.F., ed.: CICLing (2). Volume 7817 of Lecture Notes in Computer Science., Springer (2013) 394–407
5. Turing, A.M.: Computing machinery and intelligence. Mind **59**(236) (1950) 433–460
6. Harnad, S.: Minds, Machines and Turing. Journal of Logic, Language and Information **9**(4) (2000) 425–445
7. da Cunha, I., Torres-Moreno, J.M., Sierra, G.: On the development of the rst spanish treebank. In: Linguistic Annotation Workshop, The Association for Computer Linguistics (2011) 1–10

8. da Cunha, I., SanJuan, E., Moreno, J.M.T., Lloberes, M., Castellón, I.: Diseg 1.0: The first system for spanish discourse segmentation. Expert Syst. Appl. **39**(2) (2012) 1671–1678