A method for the quality assessment of automated news

Laurence Dierickx, Université Libre de Bruxelles, 2015

Abstract

This paper focus on a research about the quality assessment of automated news, which was based on computational linguistics methodology (Belz and Reiter 2006, 2009)[1][2]. The approach was here complementary with the use of metrics (quantitative indicators) and of evaluations based on human judgements (qualitative indicators, Clerwall 2014[3], Krahmer and van der Kaa 2014 [15]). Automated metrics aim to analyse the language correction, while human based judgements evaluations aim to evaluate the quality of the contents.
A sample of twenty articles generated in the economic and financial field were first submitted to several automated metrics, independent of the language and based on the comparison between two strings, which implied to get a corpus composed of source texts, written by humans, and target texts, written by softwares. The three texts that got the best results were then submitted to a panel of 80 human judges. None of them knew the real aim of this experience. Human judges were chosen among native speakers experts in writing.If this panel have quoted the qualities of precision and objectivity of the texts, they did not estimate that the three articles submitted to their appreciation were pleasant to read and well written. In two cases, they did not recognize the non-human nature of the generated articles.

1 Introduction

Natural language generation (NLG) is growing in the field of journalism since 2010. Also called automated reporting or "robot journalism", it defines a process where structured data are placed in input before being turned into texts understandable in any human language [4]. The most pessimistic discourses see in this phenomenon an added danger for the employment rate of journalists (Van Dalen 2012)[14], while a more optimistic point of view sees a strong added value tool to assist professionals in their daily tasks [6].
The method of this research relies on computational linguistics tools, which gives us indicators to evaluate the quality of automatic generated texts. NLG counts a long tradition of research on quality assessments, which include evaluations focused on tasks[1], automatic evaluations and evaluation based on human judgments. Here, we have focused on automatic and human evaluations, since they are both based on language qualities and are complementary.

1.1 Corpus selection

A corpus composed of fourty texts was selected for the needs of metrics evaluations. The field of economic and business press was chosen, often considered as sensitive and which rely on a high level of data quality in the journalistic production [11]. Ten articles were generated by Quill (Narrative Science, published by Forbes) and 10 others by Wordsmith (Automated Insights, published by Associated Press). Twenty other articles were written by humans (identification of the author had to be explicit too), coming mainly from The Wall Street Journal, The Daily Mail, Dakkota Financial News and Reuters.
As the use of metrics required articles of the same length, pre-edition was necessary to reduce articles written by journalists, with giving a particular attention to keep key information. Pre-edition has consisted essentially in cancelling paragraphs in which were reported the words of a human speaker (which however is one of the added values of journalism compared to an automatically generated text).

1.2 Quantitative assessments with the use of metrics

Metrics are language independent [12].This assessment model presents several other advantages : being rapidly launched, less costly, and less heavy to organize than evaluations based on human judgements. To be achieved, metrics need to compare two types of texts of equal length: one written by a human being (source text) and another one written by a software (target text). This corpus based method, despite its assets, was exposed to some criticism due to the fact that two texts might completely be different

[1]This method relies on psychology techniques to measure the impact and performances of a system when it is used.

[7]. According to Belz and Reiter [1], metrics imply that texts quality must be sufficient to effect a realistic test, and that the corpus must be large enough to reasonably cover changes.
Five metrics were chosen for the needs of this research: BLEU, ROUGE, NIST, WER and METEOR. Both are computed from the frequency of N-grams contained in a sequence of words. Working on the basis of N-grams has several advantages: automatic capture of the most frequent roots of words, language independence, tolerance for misspellings and deformations, does not require the removal of stop-words and stemming (which reduces the word to its canonical form or lemma) [13].
Metrics are commonly used in natural language generation, for the evaluation of automated weather reports. Their results correlates well with human evaluations (Belz, Reiter citebelz2006comparing). Any of those metrics were specifically designed for NLG. Several metrics were used here to compare results, as Belz and Reiter did for generated weather forecasts. The difference, here, was the use of the Levenshtein Distance and of the Flesch-Kincaid Readability Ease. The Levenshtein Distance, or edit distance, compares similarities and differences between two strings [16]. The Flesch-Kincaid Reading Ease [8], calculates the degree of readability of a text. This score generally ranges from 0 to 100 (a higher score corresponds to a high level of readability).

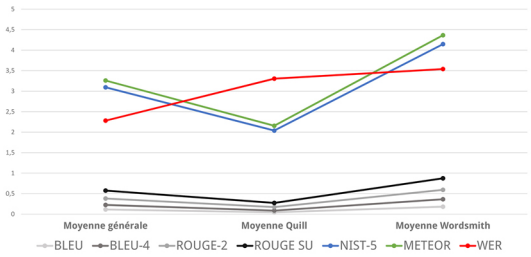

Fig.1. Comparing metrics.

If WER metric computes the distance between two strings of characters, another common method is the use of the Levenshtein Distance. This edit distance is clearly marked between articles produced by softwares and articles written by humans but due to the specific method used, it does not correlate well with WER. No impact of pre-editing was here found.
NLG systems performed generally better readability scores (Flesch-Kincaid Reading Ease).The most shorter and factual articles had recorded the highest readability scores. The brevity of generated texts (on average, 1,220 characters on a sample of twenty texts automatically generated) may explain those high scores. Pre-edited articles get generally the highest scores with an average of 68,83 (the average is 62,67 for source texts and 69,26 for targets texts). That could tend to conclude to the benefits of an interaction man-machine.

1.3 Comments about automated metric scores

Results from BLEU, ROUGE, NIST and WER metric scores shown a rather weak correspondence between the source texts/reference and target texts. That could indicate that automated texts have a relatively original character. Where scores were highest (in particular with BLEU, ROUGE and WER metrics), there were a presumption of the reuse of patterns of the target text in the source text / reference. These more favourable scores may be biased by the use of parts of texts generated for Associated Press in articles written by a journalist. The metrics indicate, in average, better performances for texts generated by Wordsmith (BLEU metrics, ROUGE metrics, NIST, METEOR, WER).

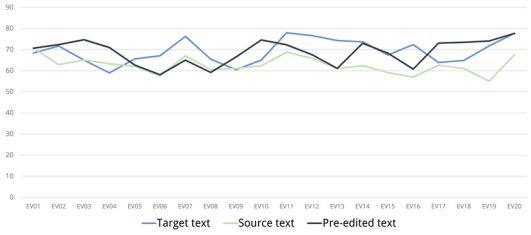

Fig.2. Measuring impact of pre-editing with the Flesch-Kincaid Reading Ease.

During the first observations, it was found that generated contents get, the most often, better readability scores. Metric evaluations has also permitted to compare the productions of two software: Quill (Narrative Science) and Wordsmith (Automated Insights). The advantage were clearly for Wordsmith, which obtained, generally, the best scores. There were “delocalised” articles in the corpus, mainly written by humans from Bangalore, in India. Those get among the worst metric scores but the corpus is not big enough to establish general conclusions.

1.4 Qualitative evaluations with human judges

If automatic metrics can provide useful measures related to the quality of the language, they say nothing about contents. Beside, metrics cannot assess important linguistics properties such as the information structure. Belz and Reiter [1] note that a corpus-based-evaluation metrics "is only sensible if they are known to be correlated with the results of human-based evaluation" but only "if the reference texts used are of high quality, or rather, can be expected to be judged high quality by the human evaluators". The corpus must also be sufficient and written by many authors. For better results, judges must be preferably monolingual (Papinieri 2002, Reiter 2009) [9] [2].
Principles of evaluations based on human judgements involve judges, who are asked to rate a corpus of generated and human written texts, by assigning them a score on a rating scale ((Belz et Reiter [2]). A human evaluation involves ensuring that subjects/judges are independent, impartial and above the familiar scope, indicate Dale and White[5]. Moreover, the human expert advice can vary considerably, put guard Belz and Reiter [1], while they note these evaluations "also depended on goodwill from participants [2].
This method of assessment, which provides a good grammatical cover, was conducted for the first time in 1997 by Lester and Porter [10]. They asked eight experts of an application domain to rate fifteen texts, regarding to different quality criteria: quality, coherence, writing style, content, organisation and accuracy. Subjects did not knew the origin of the submitted texts. A variant of this experience consists to show the subjects different versions of a same text. Another type of human evaluation covers the reading time of a text [1].
Whether automatic or human evaluations, the results of analysis measurements must be significant, well described and appropriate [5].
The three texts which obtained the best results in the first part of the research were chosen to be submitted to human judges. To obtain optimal results, it was necessary to have experts of the application domain who are native English speakers, according to results of previous researches in the field of computational linguistics. Journalists and professionals of writing were asked in UK and USA. They were not informed about the real nature of the experience, a way to not influence them. They believed that they participated to a survey about their linguistics perception of online news in the field of economy and business.

80 people have participated, while 75 have continued to completion (93,75%). Non-responses were considered as null and were not taken into account. Presumed duplicate answers were excluded of the results analysis. As the judges were also asked to indicate their profession, the panel was then divided into subgroups.

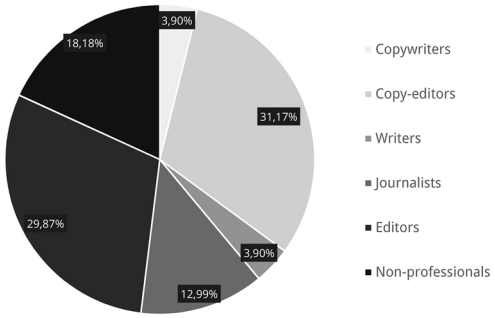

Fig.3. Expert panel.

The three automated texts with the highest scores (two from Wordsmith and one from Quill) were then submitted to the judgment of experts. Those were invited to evaluate 12 descriptors, with a methodology inspired by Clerwall. The three criteria that get highest scores are objectivity (68.46%), accuracy (65.69%) and completeness (65.17%). The worst average scores were related to the pleasure of reading (51.52%), the interest (51.51%) and the quality of writing (60.39%).

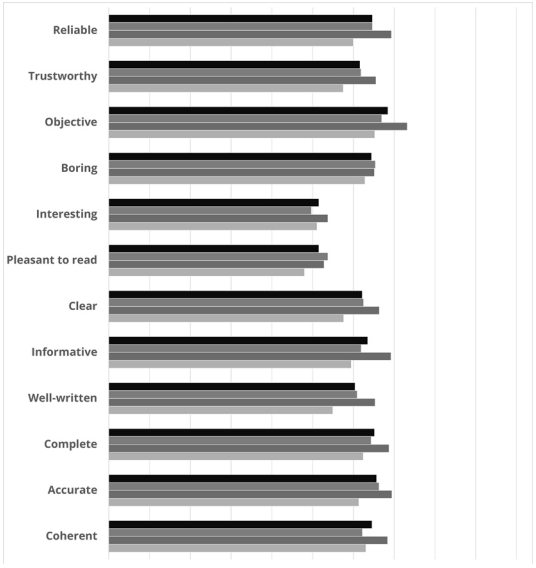

Fig.4. Human experts perception of natural language generated texts.

The lack of precise quantitative benchmarks made us compare common trends with Clerwall's research [3]. In this research, the 3 descriptors that obtained the best evaluations were for the informative, boring and trustworthy criteria. The worst scores recorded were related to criteria of reading pleasure, consistency and quality of writing.

The average results do not correlate with the highest average scores observed by Clerwall even if the boring criterion (64.46 %) get a score close to the third best result. Moreover, under the criteria used by Krahmer and Van der Kaa in their research [15], the three sample texts submitted to the human judges are considered credible (criteria of objectivity, accuracy, reliability).
Respondents had the option not to answer and, when appropriate, to comment their non-response. Critics focused mainly on sources qualified of unreliable or inadequate (32.9%). The first text has collected 32 comments, the second text, 16 comments, and the third, 19 comments (total 67 comments, including one received by e-mail). If the most frequently cited reason was about sources, the lack of warranty on the accuracy of information is then pinned (an issue related to the sources),the lack of knowledge in the domain of economic and financial information, and contradictions in the article were also quoted.
The evaluators were finally asked to guess the nature of the author of the texts: human or software? On the 75 people who has filled the whole online questionnaire, 52% thought they recognized the work of a human (average score for the 3 texts). But this result must be nuanced: one text got an excellent score for the software (77.33% had recognized it as written by a human). This result is also to be weighted depending on the subgroup to which they belong: the journalists were those who have been most sceptical with an average of 46% for the human being.

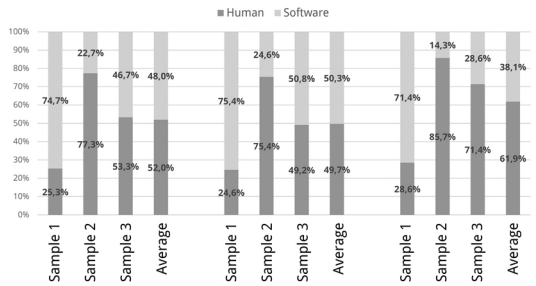

Fig.6. Perception of the author..

When looking at the size of the texts, we found that the first is the shortest (970 signs) and the second longest (1,922 characters). The panel of judges is not large enough to draw general conclusions but the question of a correlation between length of the text and improving the perception is raised. The opposite phenomenon was observed for metric evaluations, where the shorter texts tended to have, on average, the best scores. The second text submitted for evaluation (and mostly recognized as having been written by a software) is the one that has registered, in general, the less good metric scores among the three selected articles.

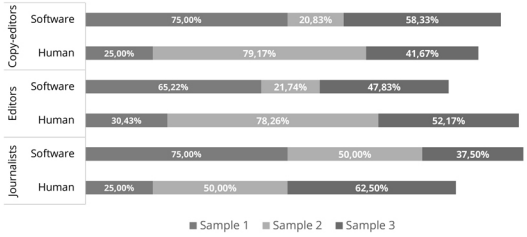

Fig.7. Detail of three subgroups of the expets panel (professionals)..

1.4.1 Points of convergence

Automatic metric on the three selected texts are following the same trend (the lowest scores for the first text and the highest scores for the second) as the evaluation of the author of the text. This observation leads to the hypothesis that a generated text with high scored metrics is more likely to be mistaken for a human production. Here, the NIST score correlates the best to the human evaluations, as Belz and Reiter has already observed in a research comparing 21 texts produced by different NLG systems in the context of weather forecasts [1].
Automatic evaluations have permitted to compare two systems, and have showed the positive impact of the prepublishing on different metrics used. It was also observed that short texts potentially lead to better measurements and that generated texts often do better readability scores. A certain originality of the generated texts was also recorded.
The evaluations based on human judgments have revealed the defects and qualities of a sample texts according to their metric performances. They have also helped to bring out some limitations related to the quality of the writing / reading, and of the data source quality. Furthermore, if the test is successful overall (with a global average of 52 %), it should be nuanced according to the subgroup to which evaluators belong, the professional group of experts was the most sceptical (journalists in particular).

1.4.2 Limits of the research

Metric and human evaluations are based on articles published online. Despite a clear identification of the author (software or journalist), no other warranty was given about the author. Moreover, when the article is coming from a press agency, it's possible that the journalist has included parts of it in his article without mentioning the source.
Another factor to take into account, is that there was no possibility to control the identity of human judges. They had the opportunity to interrupt the process at any time, or to choose not to answer. The question about the author of the article (human or software) could let them guess

that some texts could have been written by a software. It's impossible to confirm that the participants did not consider this bias. Furthermore, it's possible that some human judges have already been in contact with generated articles: since July 2014, Associated Press uses Wordsmith (Automated Insights) for business reports and the use of NLG in the field of journalism is a reality in the United States since the late 2000.

2 Conclusion

Metrics were used to compare surface generations of two NLG systems. It have shown the positive impact of the pre-edition and that generated texts often get the best readability scores. As the corpus were partially composed of news agencies, it is possible that a reusing of the texts have influences the scored. Moreover, it was observed that prepublishing activity improves metric scores, showing the complementarity between the man and the machine.
Human based judgements were focus on the quality of contents, which was considered as less well written and less pleasant to reading. Professionals have also quote a specific problem related to the source of the article (identification and or validation). In the meanwhile, they have recognized qualities as reliability and objectivity, converging with recent studies about the audiences' perception of text written by newsbots [3] [15]. Professionals were 52% to recognize an article written by a journalist, while journalists were the most skeptical subgroup.
Because of the small size of the corpus results cannot be generalized but they provide insights about the perception of journalist as well as about the way of using metrics in this particular field in order to assess the qualities of automated news.

References

[1] Anja Belz and Ehud Reiter. Comparing automatic and human evaluation of nlg systems. In *EACL*, 2006.

[2] Anja Belz and Ehud Reiter. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558, 2009.

[3] Christer Clerwall. Enter the robot journalist. *Journalism Practice*, 8(5):519–531, 2014.

[4] Robert Dale. An introduction to natural language generation. *European Summer School in Logic, Language and Information, ESSLLI'95*, 1995.

[5] Robert Dale and Michael White. Shared tasks and comparative evaluation in natural language generation, février 2012. (En ligne, site consulté le 02/02/2014). `http://www.ling.ohio-state.edu/nlgeval07/NLGEval07-Report.pdf`.

[6] Andreas Graefe. Guide to automated journalism. *Tow Center for Digital Journalism*, janvier 2016. (En ligne, consulté le 07/01/2016). `https://www.gitbook.com/book/towcenter/guide-to-automated-journalism/details`.

[7] Tomás Jesús, Mas Josep Àngel, and Casacuberta Francisco. A quantitative method for machine translation evaluation. In *Proceedings of the EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?*, pages 27–34. Association for Computational Linguistics, 2003.

[8] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch, 1975.

[9] Papineni Kishore, Roukos Salim, Ward Todd, and Zhu Wei-Jing. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[10] James C Lester and Bruce W Porter. Developing and empirically evaluating robust explanation generators: The knight experiments. *Computational Linguistics*, 23(1):65–101, 1997.

[11] Ethan Q McCallum. *Bad Data Handbook: Cleaning Up The Data So You Can Get Back To Work*. O'Reilly Media, 2012.

[12] Habash Nizar. The use of a structural n-gram language model in generation-heavy hybrid machine translation. In *Natural Language Generation*, pages 61–69. Springer, 2004.

[13] Jalam Radwan and Chauchat Jean-Hugues. Pourquoi les n-grammes permettent de classer des textes? recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques. In *6th International Conference on Textual Data Statistical Analysis, France*, pages 381–390, 2002.

[14] Arjen Van Dalen. The algorithms behind the headlines: How machine-written news redefines the core skills of human journalists. *Journalism Practice*, 6(5-6):648–658, 2013.

[15] Hille van der Kaa and Emiel Krahmer. Journalist versus news consumer: The perceived credibility of machine written news. 2014.

[16] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence (Impact Factor: 4.8)*, 29(6):1091–5, 2007.