A Snapshot of NLG Evaluation Practices 2005 - 2014

Dimitra Gkatzia
Department of Computer Science
Heriot-Watt University
EH14 4AS Edinburgh, UK
d.gkatzia@hw.ac.uk

Saad Mahamood
Department of Computing Science
University of Aberdeen
Scotland, United Kingdom
s.mahamood@abdn.ac.uk

Abstract

In this paper we present a snapshot of end-to-end NLG system evaluations as presented in conference and journal papers[1] over the last ten years in order to better understand the nature and type of evaluations that have been undertaken. We find that researchers tend to favour specific evaluation methods, and that their evaluation approaches are also correlated with the publication venue. We further discuss what factors may influence the types of evaluation used for a given NLG system.

1 Introduction

Evaluation plays a crucial role in helping to understand whether a given approach for a text generating Natural Language Generation (NLG) system has expressed particular properties (such as quality, speed, etc.) or whether it has met a particular potential (domain utility). Past work within the NLG community has looked at the issues of evaluating NLG techniques and systems, the challenges unique to the NLG context in comparison to Natural Language Analysis (Dale and Mellish, 1998), and the comparisons between evaluation approaches (Belz and Reiter, 2006). Whilst there has been a better understanding of the types of evaluations that can be conducted for a given NLG technique or system (Hastie and Belz, 2014) there is little understanding on the frequency or types of evaluation that is typically conducted for a given system within the NLG community.

In this paper, we shed some light on the frequency of the types of evaluations conducted for NLG systems. In particular, we have focused only on end-to-end complete NLG system as opposed to NLG components (referring expression generation, surface realisers, etc.) in our meta-analysis of published NLG systems from a variety of conferences, workshops, and journals for the last ten years since 2005. For the purpose of this research, we created a corpus consisting of these papers (Section 3). We then investigated three questions 4: (1) which is the most preferred evaluation method; (2) how does the method use change over time; and (3) whether the publication venue influences the evaluation type. In Section 5, we discuss the results of the meta analysis and finally in Section 6 we conclude the paper and we discuss directions for future work.

2 Background

NLG evaluation methodology has developed considerably over the last several years. Work by Dale and Mellish (1998) initially focused on the role that evaluation methods should play for a given NLG system and how they are different from the kind of evaluations undertaken by the natural language understanding community.

Traditional NLG evaluations have typically fell into one of two types: *intrinsic* or *extrinsic* (Belz and Reiter, 2006). Intrinsic evaluations of NLG systems seek to evaluate properties of the system. Past NLG systems have typically been evaluated using human subjects (Dale and Mellish, 1998). Humans have been involved in either reading and rating texts and comparing the ratings for NLG generated texts against human written texts for metrics such as quality, correctness, naturalness, understandability, etc. Extrinsic evaluations, on the other hand, have typically consisted of evaluating the impact of a given system such as its effectiveness for a given application (Belz and Reiter, 2006). These can include measuring correctness of decisions made in a task based evaluation, measuring the number of post-edits by experts, or measuring usage/utility of a given system.

The intrinsic evaluation of text output quality for NLG systems has seen different evaluation approaches. Recently, NLG systems have evaluated

[1] Dataset available from here: https://github.com/Saad-Mahamood/ENLG2015

Proceedings of the 15th European Workshop on Natural Language Generation (ENLG), pages 57–60,
Brighton, September 2015. ©2015 Association for Computational Linguistics

this particular property using comparisons to corpus text through the use of automatic metrics (Reiter and Belz, 2009). The use of automatic metrics, such as BLEU and ROUGE, have been shown to correlate with human judgements for text quality and are an attractive way of performing evaluations for NLG applications due to being fast, cheap, and repeatable (Reiter and Belz, 2009). Nevertheless, questions remain with regards to the quality and representativeness of corpora (Reiter and Sripada, 2002) used for these metrics and whether these metrics are appropriate for measuring other factors such as content selection, information structure, appropriateness, etc. (Scott and Moore, 2007).

Whilst there is an understanding of the types of evaluations that can be conducted, other unresolved issues remain. Issues such as having realistic input, having an objective criterion for assessing the quality of the NLG output, deciding on what aspects to measure for a given NLG system, what controls to use, acquiring adequate training and test data, and finally, handling disagreements between human judges (Dale and Mellish, 1998). These unresolved issues of evaluating NLG systems could be related to the fact that language is inherently context dependant. What is relevant for on NLG application task in a given domain may not be relevant to another system in a different domain (Paris et al., 2007). Thus, making direct quantitative NLG system or component evaluation comparisons is difficult outside of shared task evaluations. Additionally, whilst there has been speculation that evaluations based on human ratings and judgements are the most popular way of evaluating NLG systems (Reiter and Belz, 2009) we are not aware of any quantitative measures that supports this supposition.

3 Corpus Creation

To better understand the current nature of NLG system evaluations we performed a meta-analysis. We started by assembling a corpus consisting of as many peer reviewed papers as they could be retrieved which described end-to-end systems published at a variety of NLG conferences and workshops (ENLG, INLG, ACL, NAACL, EACL, EMNLP and COLING) and some journals (e.g. JAIR). We specifically chose a period of the last 10 years of publications to limit the scope of the corpus collection. In total, a corpus of 79 papers was assembled (consisting of: ENLG - 17, INLG - 12, ACL - 20, NAACL - 5, EACL - 7, EMNLP - 10, COLING - 3, Journals - 5). Each paper within the collected corpus was annotated using the intrinsic and extrinsic evaluation classification categories of Hastie and Belz (2014). Hastie and Belz broke down intrinsic and extrinsic evaluation methods into the following types:

Intrinsic Methods

1. *Output Quality Measures*: These assess the similarity of the systems' output to a reference model or assess quality criteria using BLUE, NIST, ROUGE, etc.
2. *User Like Measures*: For this type of evaluation, users/participants are asked questions such as "How useful did you find the generated text?" and they usually use Likert or rating scales.

Extrinsic Methods

1. *User Task Success Metrics*: A form of evaluation that measures anything that has to do with what the user gains from the systems' output, such as decision making, comprehension accuracy etc.
2. *System Purpose Success Metrics*: An evaluation type where a given system is evaluated by measuring whether it can fulfil its initial purpose.

The collected 79 papers were annotated by two annotators. To agree on the annotation procedure a set of 5 papers was annotated by both annotators. Thereafter, each annotated 33 and 49 papers including an overlapping set of 22 papers. From this overlapping set the Cohen's kappa agreement score of $\kappa = .824$ ($p < .001$) was computed.

4 Meta-analysis

Using the collected corpus of papers we investigated whether there were significant differences between the evaluation methods used. In particular we focused on the following three qualitative aspects: (1) proportionally of evaluation methods, (2) method use over time, and (3) with regard to the publication venue.

4.1 Proportions of Evaluation Methods

It was found that the majority of papers report an intrinsic evaluation method (74.7%), whereas a very small proportion of the papers report an extrinsic (15.1%) or both types of evaluation (10.1%), see also Table 1.

Regarding intrinsic evaluation, we further observed that papers report *User like measures* significantly more often than *Output Quality measures* (see also Table 2). With regard to extrinsic

Intrinsic	Extrinsic	Both
59	12	8
74.7*%	15.2*%	10.1*%

Table 1: High level descriptive statistics. * denotes significance at $p < .016$, using Z-test (after Bonferroni correction).

evaluation, most papers report a *User Task Success* evaluation setup as opposed to *System Purpose Success* methods (Table 2).

Intrinsic		Extrinsic	
Output Quality	User Like	User Task Success	System purpose success
42 38.2*%	50 **45.4*%**	13 11.8*%	5 4.6*%

Table 2: Detailed descriptive statistics. * denotes significance at $p < .008$, using Z-test (after Bonferroni correction).

We speculate that intrinsic methods are inherently easier, cheaper and quicker to be performed than extrinsic evaluations (Belz and Reiter, 2006), and therefore researchers opt for these significantly more often than extrinsic methods. In addition, intrinsic methods can be domain-independent which allows comparisons between methods. Finally, not all systems can be assessed for user task or system purpose success, e.g. commercial weather forecast systems.

4.2 Evaluation Trends over Time

Next, we investigated whether there was a change in the selection of evaluation metrics between the present and the past. For this analysis, the data was separated into three groups. The first group consisted of papers published between 2005 - 2008 (25 papers), the second group consists of publications between 2009 - 2011 (24 papers) and the last one contains papers published from 2012 to 2015 (30 papers). We used only the first and the last group in order to identify whether there are differences in the application of evaluation methods.

We observed that papers published after 2012 are significantly ($p < 0.04$) more likely to include *System Purpose* evaluations. We can also observe a trend towards intrinsic evaluations, as well as a reduction in using *User Task Success* evaluations, however the differences are not statistically significant (see also Table 3).

	2005-2008	2012-2015
Output Quality	44%	60%
User Like	56%	70%
User Task Success	24%	6.6%
System Purpose	0%	**13.4*%**

Table 3: Proportions of evaluation metrics in papers. Note that some papers contain more than one type of evaluation. * denotes significance at $p < .05$, using T-test in pair-wise comparisons.

We assume that this shift in evaluation metrics is correlated with the system design - more specific systems with well defined end users. In addition, more general purpose systems such as adult humour generation systems (Valitutti et al., 2013) have been recently developed which can be evaluated with a *System Purpose* metric in a straightforward way.

4.3 Correlations between Evaluation Methods and Publication Venue

Finally, we looked into whether papers published in specific venues "prefer" specific types of evaluation. We used Pearson's χ^2 to identify relations between the publication venues and the evaluation methods. Table 4 presents for each conference the percentages of papers that use specific evaluation metrics.

	Output Quality	User Like	User Task Success	System Purpose
ACL	70*%	65%	15%	5%
COLING	66*%	33%	33%	0%
EACL	43*%	71%	14%	0%
EMNLP	80*%	40%	20%	0%
NAACL	80*%	60%	0%	0%
ENLG	35*%	64%	12%	12%
INLG	25*%	75%	17%	17%

Table 4: Proportions of papers that report specific evaluation metrics. Note that some papers contain more than one type of evaluation. * denotes significance at $p < .05$, using Pearson's χ^2 test.

We found that more than half of the papers published at ACL, COLING, EMNLP and NAACL contain an *Output Quality* study, whereas for EACL, ENLG and INLG these percentages are below 50%. Most papers published at ACL, EACL, NAACL, ENLG and INLG also contain a "User Like" study. Extrinsic evaluation seems not to be popular across all venues (see also Table 4).

We further investigated whether there was a difference between ACL (including EACL, COLING, NAACL and EMNLP) publications and NLG publications (including ENLG and INLG). Table 5 shows the results obtained. From this analysis, journal papers have been omitted due to their low frequency.

Possible speculation for this significant difference in the use of the *Output Quality* evaluation type between the two sets of conference venues could be related to the fact that the ACL venues are patronised by a majority NLU audience. Therefore, NLG papers submitted to these conferences would be more likely to use automatic metrics

	Output Quality	User Like	User Task Success	System Purpose
ACL	68*%	57%	15%	2%
NLG	31*%	68%	13%	13%

Table 5: Proportions of ACL vs NLG papers that report specific evaluation metrics. Note that some papers contain more than one type of evaluation. * denotes significance at $p < .05$, using Pearson's χ^2 test.

(such as BLEU or ROUGE) as these measures are widely used by the NLU community as well.

5 Discussion

Output quality evaluations using automatic metrics can be repeatable (Belz and Reiter, 2006). However, automatic evaluations require large aligned corpora (input and output data), which are not always available for NLG. In such cases, other types of evaluations are preferred. In addition, Reiter and Sripada (2002) argue that the information presented in aligned corpora might not be always true, due to the fact that text from experts can be erroneous. *Output quality* metrics are sensitive to this, therefore, the research community often uses automatic metrics paired with other types of evaluations (55%) in order to overcome this barrier.

User like metrics are straightforward and easily applicable, therefore it is not surprising that these are the most popular measures among researchers. These metrics can evaluate an NLG system quickly and thus can be incorporated in any stage of a system's design. User likeness is one indication of whether a system is going to be used, as users will not use a system that they do not like. However, success on user like metrics does not equate with *system purpose success* and *user task success*. Although there are studies discussing the relation between *output quality metrics* and *user like metrics* e.g. (Foster, 2008; Belz and Reiter, 2006), to our knowledge there are not any studies discussing the relation between *user like metrics* and extrinsic metrics.

Finally, extrinsic metrics have been the least popular among researchers, due to their time-consuming nature and their complication to be organised. In addition, extrinsic metrics can be also expensive. For instance, the STOP evaluation cost £75,000 over 20 months; the SKILLSUM and BT45 evaluations cost about £20,000 over six months (Reiter and Belz, 2009).

6 Conclusion

At present NLG evaluation does not include a standardised approach for evaluating systems. Although papers tend to use automatic methods to overcome this limitation (especially papers at ACL conferences), extrinsic methods are more thorough than intrinsic and they can provide useful insights of the domains' needs, and thus they provide better indications of the systems' usefulness and utility. However, quicker and less resource intensive means are needed to allow for more systems to be evaluated with extrinsic methods.

In future, we will expand the scope of the survey by adding a greater number of journal papers for analysis and secondly and by looking at the quantitative evaluation differences between NLG systems and components. In addition, we will look into whether specific organisation and/or groups of researchers have influenced the evaluation approaches. Finally, it would be interesting to investigate whether the influential papers (for instance papers with high number of citations) have played a role in the selection of the evaluation methods.

References

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

Robert Dale and Chris Mellish. 1998. Towards the Evaluation of Natural Language Generation. In *Proceedings of the First International Conference on Language Resources and Evaluation*, volume 562.

Mary Ellen Foster. 2008. Automated Metrics That Agree With Human Judgements On Generated Output for an Embodied Conversational Agent. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 95–103.

Helen Hastie and Anja Belz. 2014. A Comparative Evaluation Methodology for NLG in Interactive Systems. In *Proceedings of the Language Resources and Evaluation Conference*.

Cécile Paris, Donia Scott, Nancy Green, Kathy McCoy, and David McDonald. 2007. Desiderata for Evaluation of Natural Language Generation. In *Shared Tasks and Comparative Evaluation on Natural Language Generation - Workshop Report*, pages 9–15.

Ehud Reiter and Anja Belz. 2009. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics*, 35(4):529–558.

Ehud Reiter and Somayajulu Sripada. 2002. Should corpora texts be gold standards for NLG? In *INLG*, pages 97–104, Harriman, NY.

Donia Scott and Johanna Moore. 2007. An NLG evaluation competition? Eight reasons to be cautious. In *Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, pages 22–23, Arlington, VA.

Alessandro Valitutti, Hannu Toivonen, Antoine Doucet, and Jukka M. Toivanen. 2013. "Let Everything Turn Well in Your Wife": Generation of Adult Humor Using Lexical Constraints. In *ACL*.