How Biased Is Your NLG Evaluation?

Pavlos Vougiouklis[1†] **Eddy Maddalena**[2†] **Jonathon Hare**[1] **Elena Simperl**[2]

School of Electronics and Computer Science
University of Southampton
Southampton, United Kingdom
[1] {pv1e13, jsh2}@ecs.soton.ac.uk
[2] {e.maddalena, e.simperl}@soton.ac.uk

Abstract

Human assessments by either experts or crowdworkers are used extensively for the evaluation of systems employed on a variety of text generative tasks. In this paper, we focus on the human evaluation of textual summaries from knowledge base triple-facts. More specifically, we investigate possible similarities between the evaluation that is performed by experts and crowdworkers. We generate a set of summaries from DBpedia triples using a state-of-the-art neural network architecture. These summaries are evaluated against a set of criteria by both experts and crowdworkers. Our results highlight significant differences between the scores that are provided by the two groups.

Introduction

In the last decade, crowdsourcing has gained increased interest since it offer the methods to reach large amounts of online contributors capable of performing in a small time large amounts of short human intelligence tasks. In particular, it has served the evaluation purposes in different areas of computer science, such as information retrieval (Alonso and Mizzaro 2012), machine learning (Lease 2011), and Natural Language Processing (Marujo et al. 2013).

Human judgments are used for the evaluation of many systems employed on a variety of text generative tasks ranging from Machine Translation (Bojar et al. 2017) and conversational agents (Ritter, Cherry, and Dolan 2011; Sordoni et al. 2015) to generation of summaries (Ell and Harth 2014; Chisholm, Radford, and Hachey 2017; Vougiouklis et al. 2017) and questions (Ngonga Ngomo et al. 2013; Du, Shao, and Cardie 2017) in natural language over knowledge graphs. Depending on the task and the evaluation criteria, these judgments are collected by either a small group of "experts" or at a larger scale by crowdworkers that are recruited through a crowdsourcing platform. Especially in the case of Natural Language Generation (NLG) over knowledge graphs, human evaluation is crucial. This is attributed to the inadequacy of the automatic text similarity metrics, such as BLEU (Papineni et al. 2002) or ROUGE (Lin 2004), to objectively evaluate the generated text (Reiter 2010).

[†]The authors contributed equally to this work.

In this paper, we focus on the human evaluation of textual summaries from knowledge base triple-facts (Chisholm, Radford, and Hachey 2017; Vougiouklis et al. 2017). More specifically, we wish to investigate whether there is any similarity between the way that experts and crowdworkers perform on the same evaluation tasks. We compile a list of three criteria that are usually employed for the human evaluation of automatically generated texts (Ell and Harth 2014; Vougiouklis et al. 2017): (i) fluency, (ii) coverage, and (iii) contradictions. We use the neural network approach that has been recently proposed by Vougiouklis et al. in order to generate textual summaries from DBpedia triples. The summaries are evaluated against the selected criteria by both experts and crowdworkers using the same task interface.

Our experiments have showed that there are significant differences between the scores that are provided by experts and the crowdworkers. Our future work will focus on the methods with which the crowdworkers should be trained in order to perform more accurately on similar tasks.

Experimental design

We run a crowdsourcing task according to which we evaluate 20 summaries that have been generated with the Triples2GRU system that has been proposed by Vougiouklis et al.. We regard each summary as a a concise representation in natural language of an input set of triple-facts. Each summary is generated by Triples2GRU given a set of 8 to 18 triples[1], and is evaluated by 10 workers.

Before starting the task, the workers are presented with general instructions. They are also informed with respect to the ethics approval that we have received for the carrying out of this experiment. The task consists of three phases through which workers were required to evaluate a given summary: (i) text fluency (with an integer number between 1 and 6), (ii) information coverage, by classifying as "Present" or "Absent" each triple-fact from a given list, and (iii) contradictions, by classifying each one of the aforementioned facts as "Direct Contraction" or "Not a Contradiction". At the beginning of each phase, the workers are presented with definitions, suggestions, examples and counter-

[1]The pre-trained version of Triples2GRU that we used (https://github.com/pvougiou/Neural-Wikipedian) accepts up to 22 triples as input.

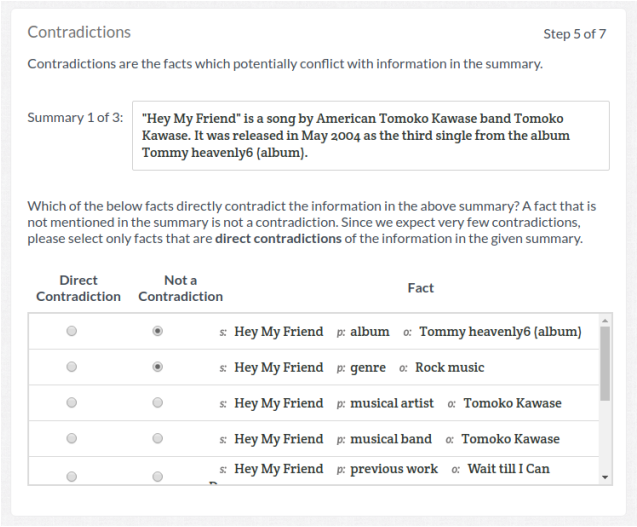

Figure 1: Task interface showing the page to identify facts whose information is contradicted in the summary.

examples. Each worker was rewarded with 0.20\$. After the carrying out of the experiment, the same 20 summaries are also evaluated under the same setup by two experts.

Results

Fluency. For each summary, (i) we computed the average of the fluency scores that have been assigned by the 10 workers. Then, (ii) we computed the average of all the values obtained in (i) resulting in an average of 4.8 out of 6. The average fluency with which the experts evaluated the 20 summaries was 5.28. The ANOVA test computed on the two fluency score series produced $p < 0.05$. Consequently, we can claim that compared to the experts, crowdworkers tend to systematically underestimate the summaries' fluency by 0.5 out of 6.
Coverage. Workers evaluated the coverage of each summary with respect to a set of triple-facts that generated it. Each summary is aligned with $8 - 18$ facts. The assessments were made by choosing between two labels: (i) "Present" for facts that are either implicitly or explicitly mentioned in the summary, and (ii) "Absent" for the rest. We compute the percentage of the "Present" facts for each summary. Then, similarly to fluency, for each summary, we first compute the average of coverage across the workers, and then the average across all the summaries. The average coverage for all the 20 summaries was 26.85%. In our second experiment, two experts repeated together the same evaluation resulting in an average of 39.71% of facts covered by the summaries. As a result, workers tend to undercount the presence of facts in the generated summaries (confirmed by ANOVA test $p < 0.05$). Finally, a positive significant correlation (Pearson $= 0.64$) pointed out that workers evaluate coverage in a consistent manner with the experts.
Contradictions. Workers were required to evaluate possible contradictions between the information in a given summary and the respective facts that generated it. Workers were required to mark as "Direct Contradiction" facts that contradict the summary, and as "Not a Contradiction" the rest. For each summary, we compute the percentage of facts that are labeled as contradictions by each single workers. Similarly to coverage, (i) for each summary, we computed the average of the percentages of contradictions of all the workers, and (ii) we averaged the contradiction scores across all the summaries. In a preliminary version of our experiments, each fact was to be marked as either "Contradiction" or "Not Contradiction". However, this proved inadequate since workers were marking facts that were not covered in the summary as contradicting, resulting in an average of $\sim 50\%$ of facts whose information is contradicted in the summaries. In order to minimize the effect of contradictions, besides changing the available labels for each triple-fact, in the contradiction instructions (shown before the third phase of the task), we explicitly noted that contradictions should be rare and that we expected many summaries without any of them. As shown in Fig. 1, we advise workers to identify as contradictions only "Direct contradictions" whose information is explicitly negated in the corresponding summary. Our final result of 30% represents the average of contradicting facts per summary. The same evaluation was made by the two expert, and the average percentage of triple-facts that are contradicted in the summaries was 0.7%. Consequently, workers tend (ANOVA test, $p < 0.05$) to significantly overestimate the presence of facts that are contradicted in the generated summaries.

Conclusion

In this paper, we presented preliminary results of a work aimed to explore the use of crowdsourcing for the evaluation of NLG systems. In particular, we focused on the evaluation of textual summaries that are generated from triple-facts. We compared the results of two studies, one that has been performed by experts and one by crowdworkers. The evaluations were conducted in three phases: (i) the fluency of the summary, (ii) the coverage, and (iii) the contradictions of a summary; the latter two are assessed with respect to the given triple-facts. Our preliminary analysis shows that crowdworkers tend to underestimate the fluency of the summaries by 0.5 out of 6. While coverage is judged consistently across both experts and crowdworkers, it is significantly underestimated by the latter. Lastly, despite the fact that we emphasised on the low number of expected contradicting facts, workers strongly overestimated their presence.

A natural extension of this work is to identify the type of facts (i.e. predicates) that influence negatively the workers' judgment. Further studies will focus on minimising this bias by both training workers on how to identify only direct contradictions, and increasing the quality control of the experiment.

Acknowledgements

This research is partially supported by the Answering Questions using Web Data (WDAqua) and QROWD projects, both of which are part of the Horizon 2020 programme under respective grant agreement Nos 642795 and 723088.

References

Alonso, O., and Mizzaro, S. 2012. Using crowdsourcing for trec relevance assessment. *Information processing & management* 48(6):1053–1066.

Bojar, O.; Chatterjee, R.; Federmann, C.; Graham, Y.; Haddow, B.; Huang, S.; Huck, M.; Koehn, P.; Liu, Q.; Logacheva, V.; Monz, C.; Negri, M.; Post, M.; Rubino, R.; Specia, L.; and Turchi, M. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, 169–214. Copenhagen, Denmark: Association for Computational Linguistics.

Chisholm, A.; Radford, W.; and Hachey, B. 2017. Learning to generate one-sentence biographies from Wikidata. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 633–642. Valencia, Spain: Association for Computational Linguistics.

Du, X.; Shao, J.; and Cardie, C. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1342–1352. Vancouver, Canada: Association for Computational Linguistics.

Ell, B., and Harth, A. 2014. A language-independent method for the extraction of RDF verbalization templates. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, 26–34. Philadelphia, Pennsylvania, U.S.A.: Association for Computational Linguistics.

Lease, M. 2011. On quality control and machine learning in crowdsourcing. *Human Computation* 11(11).

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In Marie-Francine Moens, S. S., ed., *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Marujo, L.; Gershman, A.; Carbonell, J.; Frederking, R.; and Neto, J. P. 2013. Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and coreference normalization. *arXiv preprint arXiv:1306.4886*.

Ngonga Ngomo, A.-C.; Bühmann, L.; Unger, C.; Lehmann, J.; and Gerber, D. 2013. Sorry, i don't speak SPARQL: Translating SPARQL queries into natural language. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, 977–988. New York, NY, USA: ACM.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318. Stroudsburg, PA, USA: Association for Computational Linguistics.

Reiter, E. 2010. *Natural Language Generation*. Wiley-Blackwell. chapter 20, 574–598.

Ritter, A.; Cherry, C.; and Dolan, W. B. 2011. Data-driven response generation in social media. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, 583–593. Stroudsburg, PA, USA: Association for Computational Linguistics.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.-Y.; Gao, J.; and Dolan, B. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 196–205. Denver, Colorado: Association for Computational Linguistics.

Vougiouklis, P.; ElSahar, H.; Kaffee, L.; Gravier, C.; Laforest, F.; Hare, J. S.; and Simperl, E. 2017. Neural wikipedian: Generating textual summaries from knowledge base triples. *CoRR* abs/1711.00155.