Analysing Data-To-Text Generation Benchmarks

Laura Perez-Beltrachini, Claire Gardent

Analysing Data-To-Text Generation Benchmarks

Laura Perez-Beltrachini
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
Scotland

Claire Gardent
CNRS, LORIA, UMR 7503
Vanoeuvre-lès-Nancy, F-54506
France

Abstract

A generation system can only be as good as the data it is trained on. In this short paper, we propose a methodology for analysing data-to-text corpora used for training microplanner i.e., systems which given some input must produce a text verbalising exactly this input. We apply this methodology to three existing benchmarks and we elicite a set of criteria for the creation of a data-to-text benchmark which could help better support the development, evaluation and comparison of linguistically sophisticated data-to-text generators.

1 Introduction

In some scenarios, generation datasets provide linguistic descriptions of a specific domain and application (e.g. (Reiter et al., 2005)). However, in other scenarios generation datasets aim at broader syntactic (e.g. the surface realisation shared-task (Belz et al., 2011)) or domain (Wen et al., 2015a) coverage. Recently, several datasets have been created to train data-to-text generators (Wen et al., 2015a; Liang et al., 2009; Lebret et al., 2016; Novikova et al., 2016; Chen and Mooney, 2008). It is unclear however to what extent the generation task exercised by these datasets is linguistically challenging. Do these datasets provide enough variety to support the development of high-quality data-to-text generators ? In this paper, we propose a methodology for characterising the variety and complexity of these datasets. We exemplify its use by applying it to three existing training corpora for NLG and we conclude by eliciting a set of criteria for the creation of data-to-text benchmarks which could better support the development, evaluation and comparison of linguistically sophisticated data-to-text generators.

2 Approach

Our classification aims to assess to what extent a data-to-text corpus will allow for the learning of a linguistically sophisticated microplanner i.e., a microplanner which can handle a wide range of linguistic constructions and their interaction. We focus on the following four criteria: *linguistic and computational diversity* (How complex or varied are the data and the texts?), *lexical richness* (Is the dataset lexically varied ?), *syntactic variety* (Is the dataset syntactically varied and in particular, does it include text of varied syntactic *complexity* ?) and *informational adequacy* (Does the text match the information contained in the data ?).

Linguistic and Computational Diversity. Linguistic and computational diversity can be assessed using the following metrics[1]:

Size: the number of training instances in the dataset
Nb. of Rel: the number of distinct relations
Sub.Ent: the number of distinct subject entities
Rel.Obj.Ent: the number of relation-object pairs
Da Len: the average length of the input data computed as the number of subject-relation-object triples
Da Ptns: the number of distinct relation combinations
Da Inst: the number of distinct data inputs

[1]We assume that a data-to-text corpus for NLG includes entities, concepts and binary relations. Following RDF terminology, we refer to the first argument of a binary relation as a subject entity and to the second as an object entity.

PPxDa Inst: the average (min/max) number of paraphrases per data input.

Lexical Richness. (Lu, 2012)'s system automatically measure various dimensions of lexical richness. Two measures are particularly relevant here.

Type-token ratio (TTR) is a measure of diversity defined as the ratio of the number of word types to the number of words in a text. To address the fact that this ratio tends to decrease with the size of the corpus, Mean segmental TTR (MSTTR) is computed by dividing the corpus into successive segments of a given length and then calculating the average TTR of all segments.

Lexical sophistication (LS) measures the proportion of relatively unusual or advanced word types in the text. In practice, LS is the proportion of lexical word types which are not in the list of 2,000 most frequent words from the British National Corpus.

Syntactic Variation To support the training of generators with wide syntactic coverage, a benchmark needs to show a balanced distribution of the various syntactic phenomena present in the target language. To characterise the syntactic coverage of a dataset, we use a complexity classification proposed in the domain of language learning development assessment which consists of eight levels: (0) simple sentences, including questions (1) infinitive or -ing complement with subject control; (2) conjoined noun phrases in subject position; conjunctions of sentences, of verbal, adjectival, or adverbial construction; (3) relative or appositional clause modifying the object of the main verb; nominalization in object position; finite clause as object of main verb; subject extraposition; (4) subordinate clauses; comparatives; (5) nonfinite clauses in adjunct positions; (6) relative or appositional clause modifying subject of main verb; embedded clause serving as subject of main verb; nominalization serving as subject of main verb; (7) more than one level of embedding in a single sentence.

We use (Lu, 2010)'s system for the automatic measurement of syntactic variability. Briefly, this system decomposes parse trees[2] into component sub-trees and scores each of these sub-trees based on the type of the syntactic constructions detected in it using a set of heuristics. Sentences are then assigned to a syntactic level based on the scores assigned to the sub-trees it contains as follows. If all sub-trees found in that sentence are assigned to level zero, the sentence is assigned to level 0; if one and only one non-zero level is assigned to one or more sub-trees, the sentence is assigned to that non-zero level; if two or more different non-zero scores are assigned to two or more of the sub-trees, the sentence is assigned to level 7. When evaluated against a gold standard of 500 sentences independently rated by two annotators with a very high inter-annotator agreement (kappa = 0.91), the system achieves an F-Score of 93.2% (Lu, 2010).

[2]Parses are obtained using Collins' constituent parser (Collins, 1999).

	M	A	MA	E
RNNLG$_{Laptop}$	16%	2%	0	82%
RNNLG$_{TV}$	12%	4%	0	84%
RNNLG$_{Hotel}$	0	6%	0	94%
RNNLG$_{Restaurant}$	0	6%	0	94%
IMAGEDESC	50%	6%	0	44%
WIKIBIOASTRO	30%	0	70%	0

Table 1: Match between Text and Data. M: Missing content in the text, A: Additional content in the text, MA: both additional and missing, E:Exact.

Informational Adequacy A microplanner should express all or part of the content expressed in the input data. It is therefore important to verify that this is the case through manual examination of a random subset of the dataset. A data/text pair will be considered an "Exact" match if all data is verbalised by the text. It will be labelled as "Missing" if part of the data is not present in the text (content selection) and as "Additional" if the text contains information not present in the input data.

3 Case Study

To illustrate the usage of the evaluation grid proposed in the preceding section, we apply it to three datasets recently proposed for data-to-text generation by (Lebret et al., 2016), (Wen et al., 2015b; Wen et al., 2016) and (Novikova and Rieser, 2016).

(Lebret et al., 2016)'s dataset (WIKIBIO) focuses on biographies and associates Wikipedia infoboxes with the first sentence of the corresponding article in Wikipedia. As the dataset is much larger than the other datasets and is not domain specific, we extract

two subsets of it for better comparison: one whose size is similar to the other datasets (WIKIBIO$_{16317}$) and one which is domain specific in that all biographies are about astronauts (WIKIBIOASTRO).

The other two datasets were created manually with humans providing text for dialogue acts in the case of (Wen et al., 2015b; Wen et al., 2016)'s RNNLG datasets (laptop, TV, hotel, restaurant) and image descriptions in the case of (Novikova and Rieser, 2016)'s dataset (IMAGEDESC).

We also include a text-only corpus for comparison with the texts contained in our three datasets. This corpus (GMB) consists of the texts from the Groningen Meaning Bank (Version 1.0.0, (Basile et al., 2012)) and covers different genres (e.g., news, jokes, fables).

Linguistic and Computational Diversity. Table 2 gives the descriptive statistics for each of these three datasets. It shows marked differences in terms of size (WIKIBIO$_{16317}$ being the largest and IMAGEDESC the smallest), number of distinct relations (from 16 for IMAGEDESC to 2367 for WIKIBIO$_{16317}$) and average number of paraphrases (15.11 for IMAGEDESC against 1 to 3.72 for the other two datasets). The number of distinct data inputs (semantic variability) also varies widely (from 77 distinct data inputs for the IMAGEDESC corpus to 12527 for RNNLG$_{Laptop}$). Overall the number of distinct relations is relatively small.

Lexical Richness. The WIKIBIO dataset, even when restricted to a single type of entity (namely, astronauts) has a higher MSTTR. This higher lexical variation is probably due to the fact that this dataset also has the highest number of relations (cf. Table 2): more relations brings more diversity and thus better lexical range. Indeed, there is a positive correlation between the number of relations in the dataset and MSTTR (Spearman's rho +0.385).

Again the WIKIBIO dataset has a markedly higher level of lexical sophistication than the other datasets. The higher LS might be because the WIKIBIO text are edited independently of input data thereby leaving more freedom to the authors to include additional information. It may also result from the fact that the WIKIBIO dataset, even though it is restricted to biographies, covers a much more varied set of domains than the other datasets as people's lives may be very diverse and consequently, a more varied range of topics may be mentioned than in a domain restricted dataset.

Syntactic variation. Figure 1 summarises the results for the various datasets. A first observation is that the proportion of simple texts (Level 0) is very high across the board (42% to 68%). In fact, in all data sets but two, *more than half of the sentences are of level 0 (simple sentences)*. In comparison, only 35% of the GMB corpus sentences are of level 0.

Second, levels 1, 4 and to a lesser extent level 3, are absent or almost absent from the data sets. We conjecture that this is due to the shape and type of the input data. Infinitival clauses with subject control (level 1) and comparatives (level 4) involve coreferential links and relations between entities which are absent from the simple binary relations comprising the input data. Similarly, non finite complements with their own subject (e.g., *"John saw Mary leaving the room"*, Level 3) and relative clauses modifying the object of the main verb (e.g., *"The man scolded the boy who stole the bicycle"*, Level 3) require data where the object of a literal is the subject of some other literal. In most cases however, the input data consists of sets of literals predicating facts about a single entity.

Third, datasets may be more or less varied in terms of syntactic complexity. It is in particular noticeable that, for the WIKIBIO dataset, three levels (1, 3 and 7) covers 84% of the cases. This restricted variety points to stereotyped text with repetitive syntactic structure. Indeed, in WIKIBIO, the texts consist of the first sentence of biographic Wikipedia articles which typically are of the form *"W (date of birth - date of death) was P"*. where P usually is an arbitrarily complex predicate potentially involving relative clauses modifying the object of main verb (Level 3) and coordination (Level 7).

Informational Adequacy. Each data-text pair was independently rated by two annotators resulting in a kappa score ranging between 0.566 and 0.691 depending on the dataset. The results shown in Table 1 highlight some important differences. While the RNNLG datasets have a high percentage of exact entries (82% to 94%), the IMAGEDESC dataset is less precise (44% of exact matches). The WIKIBIO datasets does not contain a single example where data and text coincide. These differences can be

Dataset	Size	Nb. of Rel	Sub.Ent‡	Rel.Obj.Ent	Da Len.	Da Ptns	Da Inst	PPxDa Inst.
WIKIBIO$_{16317}$	16317	2367	16317	149484	19.65	9990	16317	1
WIKIBIOASTRO	615	68	615	5290	15.46	293	615	1
RNNLG$_{Laptop}$	13242	34	123	451	5.86	2068	12527	1.03(1/3)
RNNLG$_{TV}$	7035	30	92	300	5.79	1024	6808	1.01(1/6)
RNNLG$_{Hotel}$	5373	22	138	535	2.66	112	940	3.72(1/149)
RNNLG$_{Restaurant}$	5192	22	223	869	2.86	182	1950	1.82(1/101)
IMAGEDESC	1242	16	33	117	5.33	21	77	15.11(8/22)

Table 2: Datasets descriptive statistics. ‡Note that we consider as distinct entities those given by the *name* relations and that in the RNNLG datasets not all dialogue acts describe entities (e.g. inform_count or ?select).

Dataset	Tokens	Types	LS	MSTTR
WIKIBIO$_{16317}$	377048	36712	**0.92**	**0.82**
WIKIBIOASTRO	14720	2335	0.81	0.8
RNNLG$_{Laptop}$	295492	1757	0.46	0.74
RNNLG$_{TV}$	141606	1171	0.48	0.71
RNNLG$_{Hotel}$	48982	967	0.43	0.59
RNNLG$_{Restaurant}$	45791	1187	0.43	0.62
IMAGEDESC	20924	598	0.47	0.56
GMB	75927	7791	0.75	0.81

Table 3: Lexical Sophistication (LS) and Mean Segmental Type-Token Ratio (MSTTR).

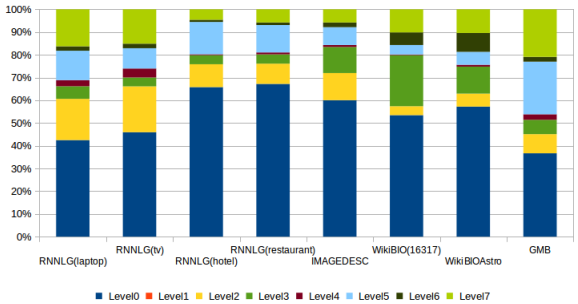

Figure 1: Syntactic complexity. D-Level sentence distribution.

traced back to the way in which each resource was created. The WIKIBIO dataset is created automatically from Wikipedia infoboxes and articles while information adequacy is not checked for. In the IMAGEDESC dataset, the texts are created from images using crowdsourcing. It seems that this method, while enhancing variety, makes it easier for the crowdworkers to omit some information.

4 Conclusion

The proposed measures suggest several key aspects to take into account when constructing a data-to-text dataset for the development and evaluation of NLG systems. Lexical richness can be enhanced by including data from different domains, using a large number of distinct relations and ensuring that the total number of distinct inputs is high. Wide and balanced syntactic coverage is difficult to ensure and probably requires input data of various size and shape, stemming from different domains. Informational adequacy is easiest to achieve using crowdsourcing which also facilitates the inclusion of paraphrases. In future work, it would be interesting to further exploit such analyses of data-to-text corpora (i) to better characterise the generators that can be learnt from a given corpus, (ii) to perform a graded analysis of generation systems on data of various syntactic complexity or (iii) to support error mining (which type of data is most often associated with generation failure ?).

More specifically, our classification could be useful to identify sources of under-performance and thus directions for improvements. For instance, BLEU results reported by (Wen et al., 2015a) on three different datasets indicate that the same systems are facing different difficulties on each of these. Indeed, lexical richness is higher (Table 3) for the RNNLG$_{Laptop}$ dataset for which (Wen et al., 2015a) reports the lowest BLEU score. But also the proportion of simple sentences is lower (Figure 1) in this dataset. A focused evaluation could report on BLEU scores aggregated on the syntactic classification of sentences into levels.

Acknowledgments

This research was partially supported by the French National Research Agency within the framework of the WebNLG Project (ANR-14-CE24-0033).

References

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In *LREC*, volume 12, pages 3196–3200.

Anja Belz, Mike White, Dominic Espinosa, Eric Kow, Deirdre Hogan, and Amanda Stent. 2011. The first surface realisation shared task: Overview and evaluation results. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, pages 217–226, Nancy, France, September. Association for Computational Linguistics.

David L Chen and Raymond J Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*, pages 128–135. ACM.

M. Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas, November. Association for Computational Linguistics.

P. Liang, M. I. Jordan, and D. Klein. 2009. Learning semantic correspondences with less supervision. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 91–99.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners oral narratives. *The Modern Language Journal*, 96(2):190–208.

Jekaterina Novikova and Verena Rieser. 2016. The analogue challenge: Non aligned language generation. In *The 9th International Natural Language Generation conference*, page 168.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing nlg data: Pictures elicit better data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 265–273, Edinburgh, UK, September 5-8. Association for Computational Linguistics.

Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167(1-2):137–169.

Tsung-Hsien Wen, Milica Gašic, Nikola Mrkšic, Lina M Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. 2015a. Toward multi-domain language generation using recurrent neural networks. In *The Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS), Workshop on Machine Learning for Spoken Language Understanding and Interaction*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015b. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. 2016. Multi-domain neural network language generation for spoken dialogue systems. *arXiv preprint arXiv:1603.01232*.