The aNALoGuE Challenge: Non Aligned Language GEneration

Jekaterina Novikova and **Verena Rieser**
The Interaction Lab, Heriot-Watt University, Edinburgh, UK
j.novikova,v.t.rieser@hw.ac.uk

Abstract

We propose a shared task based on recent advances in learning to generate natural language from meaning representations using semantically unaligned data. The aNALoGuE challenge aims to evaluate and compare recent corpus-based methods with respect to their scalability to data size and target complexity, as well as to assess predictive quality of automatic evaluation metrics.

1 Relevance

Natural language generation plays a critical role for Conversational Agents (CAs) as it has a significant impact on a users impression of the system. Most CAs utilise domain-dependent methods including hand-written grammars or domain-specific language templates for surface realisation, both of which are costly to develop and maintain. Recent corpus-based methods hold the promise of being easily portable across domains, e.g. (Angeli et al., 2010; Konstas and Lapata, 2012; Mairesse and Young, 2014), but require high quality training data consisting of meaning representations (MR) paired with natural language (NL) utterances, augmented by alignments between elements of meaning representation and natural language words. Creating aligned data is a non-trivial task in its own right, see e.g. (Liang et al., 2009). This shared task aims to strengthen recent research on corpus-based NLG from unaligned data, e.g. (Dušek and Jurcicek, 2015; Wen et al., 2015; Mei et al., 2015; Sharma et al., 2016). These approaches do not require costly semantic alignment, but are based on parallel data sets, which can be collected in sufficient quality and quantity using effective crowd-sourcing techniques (Novikova and Rieser, 2016), and as such open the door for rapid development of NLG components for CAs in new domains.

In addition, we hope to attract interest from related disciplines, such as semantic parsing or statistical machine translation, which face similar challenges when learning from parallel non-aligned data sets.

Flat MR	**NL reference**
name[The Eagle], eatType[coffee shop], food[French], priceRange[moderate], customerRating[3/5], area[riverside], kidsFriendly[yes], near[Burger King]	1. There is a riverside coffee shop called The Eagle that has French food at an average price range. It is child friendly, located near Burger King, and has a 3 star customer rating. 2. The three star coffee shop, The Eagle, gives families a mid-priced dining experience featuring a variety of wines and cheeses. Find The Eagle near Burger King. 3. The Eagle coffee shop is based in the riverside area near Burger King. It serves food at mid range prices. It has a three star rating and is family friendly.

Table 1: An example of a data instance.

2 Data Description

The data provided for this shared challenge was collected by using the CrowdFlower platform and quality controlled as described in (Novikova and

Proceedings of The 9th International Natural Language Generation conference, pages 168–170,
Edinburgh, UK, September 5-8 2016. ©2016 Association for Computational Linguistics

Rieser, 2016). The dataset provides information about restaurants and consists of more than 50k combinations of a dialogue act-based meaning representation and up to 5 references in natural language, as shown in Table 1. Each MR consists of 3 - 8 attributes (labels), such as name, food or area. The detailed ontology of all attributes and values is provided in Table 2. The dataset will be split into training, validation and testing sets (70/15/15). The training and validation sets will be provided to the participants, while the testing set is used for the final evaluation of the systems. The sets are constructed to ensure a similar distribution of single-sentenced and multi-sentenced references in each set, as well as a similar distribution of MRs of different length.

Attribute	Data Type	Example value
name	verbatim string	The Eagle, ...
eatType	dictionary	restaurant, pub, ...
familyFriendly	boolean	Yes / No
priceRange	dictionary	cheap, expensive, ...
food	dictionary	French, Italian, ...
near	verbatim string	market square, ...
area	dictionary	riverside, city center, ...
customerRating	enumerable	1 of 5 (low), 4 of 5 (high), ...

Table 2: Domain ontology.

3 Evaluation

We will provide two types of baseline systems, which are frequently used by previous corpus-based methods, e.g. (Wen et al., 2015; Mairesse and Young, 2014): a challenging hand-crafted generator and n-gram Language Models, following early work by (Oh and Rudnicky, 2002). To evaluate the results, both objective and subjective metrics will be used. We will explore automatic measures, such as BLEU-4 (Papineni et al., 2002) and NIST (Doddington, 2002) scores, which are widely used in a machine translation and NLG research, and will allow comparing the results of this challenge with previous work. Since automatic metrics may not consistently agree with human perception, human evaluation will be used to assess subjective quality of generated utterances. Human judges will be recruited using CrowdFlower. Judges will be asked to compare utterance generated by different systems and score them in terms of informativeness (*"Does the utterance contains all the information specified in the MR?"*), naturalness (*"Could the utterance have been produced by a native speaker?"*) and phrasing (*"Do you like the way the utterance has been expressed?"*). Here, we will explore different experimental setups for evaluation following previous shared tasks, e.g. (Belz and Kow, 2011). The challenge will also benefit from a national research grant on Domain Independent NLG (EP/M005429/1) which will provide funds for crowd-based evaluation.

4 Research Questions

The task is set up to answer the following research questions with respect to corpus-driven methods:

• "*How much data is enough?*" So far, corpus-based methods have been trained on limited data sets, such as BAGEL (404 target utterances), Cambridge SF (5193) or RoboCup (1919). We release a data set which is almost 10-times times bigger in size than previous corpora. This allows us to test the upper quality boundary of corpus-driven NLG, as well as to determine the optimal/minimal data size per algorithm.

• "*Can they model more complex targets?*"So far, corpus-driven methods are restricted to single sentences. Our corpus contains 37% examples with multiple (2-6) sentences. We predict that longer target outputs are challenging for, e.g. neural networks due to the vanishing gradient problem. Furthermore, our crowd-sourced utterances were elicited using pictures, which makes them more varied in sentence structure and vocabulary than previously used corpora (Novikova and Rieser, 2016).

• "*How good is BLEU?*" Previous research has shown that automatic metrics like BLEU do not consistently agree with human perception (Stent et al., 2004; Belz and Gatt, 2008). We will therefore explore how well they correlate with human judgement. We will also explore how well these metrics are able to capture desired variation given a set of possible reference sentences, following similar shared tasks in machine translation, e.g. (Stanojević et al., 2015).

Acknowledgments

This research received funding from the EPSRC projects GUI (EP/L026775/1), DILiGENt (EP/M005429/1) and MaDrIgAL (EP/N017536/1).

References

Gabor Angeli, Percy Liang, and Dan Klein. 2010. A simple domain-independent probabilistic approach to generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Anja Belz and Albert Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 197–200, Columbus, Ohio, June. Association for Computational Linguistics.

Anja Belz and Eric Kow. 2011. Discrete vs. continuous rating scales for language evaluation in nlp. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 230–235, Stroudsburg, PA, USA. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Ondřej Dušek and Filip Jurcicek. 2015. Training a natural language generator from unaligned data. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 451–461, Beijing, China, July. Association for Computational Linguistics.

Ioannis Konstas and Mirella Lapata. 2012. Unsupervised concept-to-text generation with hypergraphs. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proc. of ACL-IJCNLP*.

François Mairesse and Steve Young. 2014. Stochastic language generation in dialogue using factored language models. *Comput. Linguist.*, 40(4):763–799, December.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2015. What to talk about and how? selective generation using lstms with coarse-to-fine alignment. *CoRR*, abs/1509.00838.

Jekaterina Novikova and Verena Rieser. 2016. Crowdsourcing NLG data: Pictures elicit better data. In *Proc. of the 9th International Natural Language Generation conference (INLG)*.

Alice H. Oh and Alexander I. Rudnicky. 2002. Stochastic natural language generation for spoken dialog systems. *Computer Speech and Language*, 16:387–407.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Shikhar Sharma, Jing He, Kaheer Suleman, Hannes Schulz, and Philip Bachman. 2016. Natural language generation in dialogue using lexicalized and delexicalized data. *CoRR*, abs/1606.03632.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the wmt15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal, September. Association for Computational Linguistics.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proceedings of the 42nd annual meeting on association for computational linguistics*, page 79. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal, September. Association for Computational Linguistics.