The Surface Realisation Task: Recent Developments and Future Plans

Anja Belz
Computing, Engineering and Maths
University of Brighton
Brighton BN1 4GJ, UK
a.s.belz@brighton.ac.uk

Bernd Bohnet
Institute for Natural Language Processing
University of Stuttgart
70174 Stuttgart
bohnet@ims.uni-stuttgart.de

Simon Mille, Leo Wanner
Information and Communication Technologies
Pompeu Fabra University
08018 Barcelona
<firstname>.<lastname>@upf.edu

Michael White
Department of Linguistics
Ohio State University
Columbus, OH, 43210, US
mwhite@ling.osu.edu

Abstract

The Surface Realisation Shared Task was first run in 2011. Two common-ground input representations were developed and for the first time several independently developed surface realisers produced realisations from the same shared inputs. However, the input representations had several shortcomings which we have been aiming to address in the time since. This paper reports on our work to date on improving the input representations and on our plans for the next edition of the SR Task. We also briefly summarise other related developments in NLG shared tasks and outline how the different ideas may be usefully brought together in the future.

1 Introduction

The Surface Realisation (SR) Task was introduced as a new shared task at Generation Challenges 2011 (Belz et al., 2011). Our aim in developing the SR Task was to make it possible, for the first time, to directly compare different, independently developed surface realisers by developing a 'common-ground' representation that could be used by all participating systems as input. In fact, we created two different input representations, one shallow, one deep, in order to enable more teams to participate. Correspondingly, there were two tracks in SR'11: In the Shallow Track, the task was to map from shallow syntax-level input representations to realisations; in the Deep Track, the task was to map from deep semantics-level input representations to realisations.

By the time teams submitted their system outputs, it had become clear that the inputs required by some types of surface realisers were more easily derived from the common-ground representation than the inputs required by other types. There were other respects in which the representations were not ideal, e.g. the deep representations retained too many syntactic elements as stopgaps where no deeper information had been available. It was clear that the input representations had to be improved for the next edition of the SR Task. In this paper, we report on our work in this direction so far and relate it to some new shared task proposals which have been developed in part as a response to the above difficulties. We discuss how these developments might usefully be integrated, and outline plans for SR'13, the next edition of the SR Task.

2 SR'11

The SR'11 input representations were created by post-processing the CoNLL 2008 Shared Task data (Surdeanu et al., 2008), for the preparation of which selected sections of the WSJ Treebank were converted to syntactic dependencies with the Pennconverter (Johansson and Nugues, 2007). The resulting dependency bank was then merged with Nombank (Meyers et al., 2004) and Propbank (Palmer et al., 2005). Named entity information from the BBN Entity Type corpus was also incorporated. The SR'11 shallow representation was based on the Pennconverter dependencies, while the deep representation was derived from the merged Nombank, Propbank and syntactic dependencies in a pro-

INLG 2012 Proceedings of the 7th International Natural Language Generation Conference, pages 136–140,
Utica, May 2012.

cess similar to the graph completion algorithm outlined by Bohnet (2010).

Five teams submitted a total of six systems to SR'11 which we evaluated automatically using a range of intrinsic metrics. In addition, systems were assessed by human judges in terms of Clarity, Readability and Meaning Similarity.

The four top-performing systems were all statistical dependency realisers that do not make use of an explicit, pre-existing grammar. By design, statistical dependency realisers are robust and relatively easy to adapt to new kinds of dependency inputs which made them well suited to the SR'11 Task. In contrast, there were only two systems that employed a grammar, either hand-crafted or treebank-derived, and these did not produce competitive results. Both teams reported substantial difficulties in converting the common ground inputs into the 'native' inputs required by their systems.

The SR'11 results report pointed towards two kinds of possible improvements: (i) introducing (additional) tasks where performance would not depend to the same extent on the relation between common-ground and native inputs, e.g. a text-to-text shared task on sentential paraphrasing; and (ii) improving the representations themselves. In the remainder of this paper we report on developments in both these directions.

3 Towards SR'13

As outlined above, the first SR Shared Task turned up some interesting representational issues that required some in-depth investigation. In the end, it was this fact that led to the decision to postpone the 2nd SR Shared Task until 2013 in order to allow enough time to address these issues properly. In this section, we describe our plans for SR'13 to the extent to which they have progressed.

3.1 Task definition

As in the first SR task, the participating teams will be provided with annotated corpora consisting of common-ground input representations and their corresponding outputs. Two kinds of input will be offered: deep representations and surface representations. The deep input representations will be semantic graphs; the surface representations syntactic trees. Both will be derived from the Penn Treebank. The task will consist in the generation of a text starting from either of the input representations.

3.2 Changes to the input representations

During the working group discussions which followed SR'11, it became apparent that the CoNLL syntactic dependency trees overlaid with Propbank/Nombank relations had turned out to be inadequate in various respects for the purpose of deriving a suitable semantic representation. For instance:

- **Governed prepositions** are not distinguished from semantically loaded prepositions in the CoNLL annotation. In SR'11, only strongly governed prepositions such as *give something TO someone* were removed, but in many cases the meaning of a preposition which introduces an argument (of a verb, a noun, an adjective or an adverb) clearly depends on the predicate: *believe IN something*, *account FOR something*, etc. In those cases, too, the preposition should be removed from the semantic annotation, since the relisers have to be able to introduce non-semantic features un-aided. On the contrary, semantically loaded governed prepositions such as *live IN a flat/ON a roof/NEXT TO the main street* etc. should be retained in the annotation. These prepositions all receive argumental arcs in PropBank/NomBank, so it is not easy to distinguish between them. One possibility would be to target a restricted list of prepositions which are void of meaning most of the time, and remove those prepositions when they introduce arguments.

- The annotation of **relative pronouns** did not survive the conversion of the original Penn Treebank to the CoNLL format unscathed: the antecedent of the relative pronoun is sometimes lost or the relative pronoun is not annotated, predominantly because the predicate which the relative pronoun is an argument of was not considered to be a predicate by annotators, as in *the degree TO WHICH companies are irritated*. However, in the original constituency annotation, the traces allow for retrieving antecedents and semantic governors, hence using this orig-

inal annotation could be useful in order to get a clean annotation of such phenomena.

Agreement has been reached on a range of other issues, although the feasibility of implementing the corresponding changes might have to be further evaluated:

- **Coordinations** should be annotated in the semantic representation with the conjunction as the head of all the conjuncts. This treatment would allow e.g. an adequate representation of sharing of dependents among the conjuncts.
- The inversion of 'modifier' arcs and the introduction of **meta-semantemes** would avoid anticipating syntactic decisions such as the direction of non-argumental syntactic edges, and allow for connecting unconnected parts of the semantic structures.
- In order to keep the **scope** of various phenomena intact after inverting non-argumental edges, we should explicitly mark the scope of e.g. negations, quantifiers, quotation marks etc. as attribute values on the nodes.
- **Control arcs** should be removed from the semantic representation since they do not provide information relevant at that level.
- **Named entities** will be further specified adding a reduced set of named entity types from the BBN annotations.

Finally, we will perform automatic and manual quality checks in order to ensure that the proposed changes are adequately introduced in the annotation.

3.3 Evaluation

We will once again follow the main data set divisions of the CoNLL'08 data (training set = WSJ Sections 02–21; development set = Section 24; test set = Section 23), with the proviso that we have removed 300 randomly selected sentences from the development set for use in human evaluations. Of these, we used 100 sentences in SR'11 and will use a different 100 in SR'13.

Evaluation criteria identified as important for evaluation of surface realisation output in previous work include Adequacy (preservation of meaning), Fluency (grammaticality/idiomaticity), Clarity, Humanlikeness and Task Effectiveness. We will aim to evaluate system outputs submitted by SR'13 participants in terms of most of these criteria, using both automatic and human-assessed methods.

As in SR'11, the automatic evaluation metrics (assessing Humanlikeness) will be BLEU, NIST, TER and possibly METEOR. We will apply text normalisation to system outputs before scoring them with the automatic metrics. For n-best ranked system outputs, we will again compute a single score for all outputs by computing their weighted sum of their individual scores, where a weight is assigned to a system output in inverse proportion to its rank. For a subset of the test data we may obtain additional alternative realisations via Mechanical Turk for use in the automatic evaluations.

We are planning to expand the range of human-assessed evaluation experiments (assessing Adequacy, Fluency and Clarity) to the following methods:

1. Preference Judgement Experiment (C2, C3): Collect preference judgements using an existing evaluation interface (Kow and Belz, 2012) and directly recruited evaluators. We will present sentences in the context of a chunk of 5 consecutive sentences to the evaluators, and ask for separate judgements for Clarity, Fluency and Meaning Similarity.

2. HTER (Snover et al., 2006): In this evaluation method, human evaluators are asked to post-edit the output of a system, and the edits are then categorised and counted. Crucial to this evaluation method is the construction of clear instructions for evaluators and the categorisation of edits. We will categorise edits as relating to Meaning Similarity, Fluency and/or Clarity; we will also consider further subcategorisations.

We will once again provide evaluation scripts to participants so they can perform automatic evaluations on the development data. These scores serve two purposes. Firstly, development data scores must be included in participants' reports. Secondly, partici-

pants may wish to use the evaluation scripts in developing and tuning their systems.

We will report per-system results separately for the automatic metrics (4 sets of results), and for the human-assessed measures (2 sets of results). For each set of results, we will report single-best and n-best results. For single-best results, we may furthermore report results both with and without missing outputs. We will rank systems, and report significance of pairwise differences using bootstrap resampling where necessary (Koehn, 2004; Zhang and Vogel, 2010). We will separately report correlation between human and automatic metrics, and between different automatic metrics.

3.4 Assessing different aspects of realisation separately

In addition, we will consider measuring different aspects of the realisation performance of participating systems (syntax, word order, morphology) since a system can perform well on one and badly on another. For instance, a system might perform well on morphological realisation while it has poor results on linearisation. We would like to capture this fact. This may involve asking participating teams to submit intermediate representations or identifiers to identify the reference words. This more fine-grained approach should help us to obtain a more precise picture of the state of affairs in the field and could help to reveal the respective strengths of different surface realisers more clearly.

4 Related Developments

4.1 Syntactic Paraphrase Ranking

The new shared task on syntactic paraphrase ranking described elsewhere in this volume (White, 2012) is intended to run as a follow-on to the main surface realisation shared task. Taking advantage of the human judgements collected to evaluate the surface realisations produced by competing systems, the task is to automatically rank the realisations that differ from the reference sentence in a way that agrees with the human judgements as often as possible. The task is designed to appeal to developers of surface realisation systems as well as machine translation evaluation metrics. For surface realisation systems, the task sidesteps the thorny issue of converting inputs to a common representation. Developers of realisation systems that can generate and optionally rank multiple outputs for a given input will be encouraged to participate in the task, which will test the system's ability to produce acceptable paraphrases and/or to rank competing realisations. For MT evaluation metrics, the task provides a challenging framework for advancing automatic evaluation, as many of the paraphrases are expected to be of high quality, differing only in subtle syntactic choices.

4.2 Content Selection Challenge

The new shared task on content selection has been put forward (Bouayad-Agha et al., 2012) to initiate work on content selection from a common, standardised semantic-web format input, and thus provide the context for an objective assessment of different content selection strategies. The task consists in selecting the contents communicated in reference biographies of celebrities from a large volume of RDF-triples. The selected triples will be evaluated against a gold triple selection set using standard quality assessment metrics.

The task can be considered complementary to the surface realisation shared task in that it contributes to the medium-term goal of setting up a task that covers all stages of the generation pipeline. In future challenges, it can be explored to what extent and how the output content plans can be mapped onto semantic representations that serve as input to the surface realisers.

5 Plans

We are currently working on the new improved common-ground input representation scheme and converting the data to the new scheme.

The provisional schedule for SR'13 looks as follows:

Announcement and call for expressions of interest:	6 July 2012
Preliminary registration and release of description of new representations:	27 July 2012
Release of data and documentation:	2 Nov 2012
System Submission Deadline:	10 May 2013
Evaluation Period:	10 May– 10 Jul 2013
Provisional dates for results session:	8–9 Aug 2013

6 Conclusion

For a large number of NLP applications (among them, e.g., text generation proper, summarisation, question answering, and dialogue), surface realisation (SR) is a key technology. Unfortunately, so far in nearly all of these applications, idiosyncratic, custom-made SR implementations prevail. However, a look over the fence at the language analysis side shows that the broad use of standard dependency treebanks and semantically annotated resources such as PropBank and NomBank that were created especially with parsing in mind led to standardised high-quality off-the-shelf parser implementations. It seems clear that in order to advance the field of surface realisation, the generation community also needs adequate resources on which large-scale experiments can be run in search of the surface realiser with the best performance, a surface realiser which is commonly accepted, follows general transparent principles and is thus usable as plug-in in the majority of applications.

The SR Shared Task aims to contribute to this goal. On the one hand, it will lead to the creation of NLG-suitable resources in that it will convert the PropBank into a more semantic and more completely annotated resource. On the other hand, it will offer a forum for the presentation and evaluation of various approaches to SR and thus help us to search for the best solution to the SR task with the greatest potential to become a widely accepted off-the-shelf tool.

Acknowledgments

We gratefully acknowledge the contributions to discussions and development of ideas made by the other members of the SR working group: Miguel Ballesteros, Johan Bos, Aoife Cahill, Josef van Genabith, Pablo Gervás, Deirdre Hogan and Amanda Stent.

References

Anja Belz, Michael White, Dominic Espinosa, Deidre Hogan, Eric Kow, and Amanda Stent. 2011. The first surface realisation shared task: Overview and evaluation results. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG'11)*, pages 217–226. Association for Computational Linguistics.

Bernd Bohnet, Leo Wanner, Simon Mille, and Alicia Burga. 2010. Broad coverage multilingual deep sentence generation with a stochastic multi-level realizer. In *Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, China.

Nadjet Bouayad-Agha, Gerard Casamayor, Leo Wanner, and Chris Mellish. 2012. Content selection from semantic web data. In *Proceedings of the 7th International Natural Language Generation Conference (INLG'12)*.

Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for English. In Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek, and Mare Koit, editors, *Proceedings of NODALIDA 2007*, pages 105–112, Tartu, Estonia.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July. Association for Computational Linguistics.

Eric Kow and Anja Belz. 2012. LG-Eval: A toolkit for creating online language evaluation experiments. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank project: An interim report. In *NAACL/HLT Workshop Frontiers in Corpus Annotation*.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: A corpus annotated with semantic roles. In *Computational Linguistics Journal*, pages 71–105.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL'08)*, Manchester, UK.

Michael White. 2012. Shared task proposal: Syntactic paraphrase ranking. In *Proceedings of the 7th International Natural Language Generation Conference (INLG'12)*.

Ying Zhang and Stephan Vogel. 2010. Significance tests of automatic machine translation evaluation metrics. *Machine Translation*, 24:51–65.