Multilingual Surface Realization Using Universal Dependency Trees

Simon Mille
UPF
Barcelona, ES
simon.mille@upf.edu

Bernd Bohnet
Google Research
London, UK
bohnetbd@google.com

Leo Wanner
ICREA and UPF
Barcelona, ES
leo.wanner@upf.edu

Anja Belz
U of Brighton
Brighton, UK
A.S.Belz@brighton.ac.uk

Abstract

We propose a shared task on multilingual Surface Realization, i.e., on mapping unordered and uninflected universal dependency trees to correctly ordered and inflected sentences in a number of languages. A second deeper input will be available in which, in addition, functional words, fine-grained PoS and morphological information will be removed from the input trees. The first shared task on Surface Realization was carried out in 2011 with a similar setup, with a focus on English. We think that it is time for relaunching such a shared task effort in view of the arrival of *Universal Dependencies* annotated treebanks for a large number of languages on the one hand, and the increasing dominance of *Deep Learning*, which proved to be a game changer for NLP, on the other hand.

1 Introduction

In 2017, three shared tasks on Natural Language Generation (NLG) take place: Task 9 of SemEval (May and Priyadarshi, 2017), WebNLG[1] and E2E[2]. The first starts from *Abstract Meaning Representations* (AMRs), the second from *RDF triples*, and the third from dialog act-based *Meaning Representations* (MRs) respectively. With these efforts, the focus is put on "real-life" generation, since the respective inputs come from existing analyzers (for AMRs) or existing databases (for RDF triples and MRs). This shows that the research on NLG is on the right track and that there is an interest in large scale "deep" NLG. However, both the 2017 and the past shared tasks (including the 2011 Surface Realization Shared Task (Belz and et al., 2011)) focus on English; multilingual generation has been neglected largely so far.

On the other side, the last years saw a push in the annotation of multilingual treebanks with so-called *Universal Dependencies* (UDs), such that nowadays resources for a number of languages are available and can be used for shared tasks.[3] Furthermore, recent years witnessed a shift of the processing paradigm in applications such as parsing and machine translation from traditional supervised machine learning techniques to deep learning.[4] This is also a chance for NLG, which could benefit from deep learning to a greater extent than it currently does.

Our objective is to set up a follow-up of the 2011 Surface Realization Shared Task (SR'11) at Generation Challenges (Belz and et al., 2011); this time with an emphasis of multilingual surface generation from UD treebanks. The success of deep learning techniques in a number of areas of natural language processing furthermore opens the avenue to a broader range of system designs than have been seen before.

As in SR'11, the proposed shared task comprises

[1] http://talc1.loria.fr/webnlg/stories/challenge.html

[2] http://www.macs.hw.ac.uk/InteractionLab/E2E/

[3] See the recent parsing shared task based on UDs (Nivre and de Marneffe et al., 2016): http://universaldependencies.org/conll17/.

[4] See for instance the 1st NMT workshop, in which the NLG topic is also addressed: https://sites.google.com/site/acl17nmt/.

two tracks with different levels of difficulty:[5]

- Shallow Track: This track will start from genuine UD structures from which word order information has been removed and the tokens have been lemmatized, i.e., from unordered dependency trees with lemmatized nodes that hold PoS tags and morphological information as found in the original annotations. It will consist in determining the word order and inflecting words.
- Deep Track: This track will start from UD structures from which functional words (in particular, auxiliaries, functional prepositions and conjunctions) and surface-oriented morphological information have been removed. In addition to what has to be done for the Shallow Track, the Deep Track will thus consist of the introduction of the removed functional words and morphological features.

The participating teams will be expected to produce outputs at least for the Shallow Track.

2 Data

Universal Dependencies[6] (UD) have attracted in recent years interest from many researchers across different fields of NLP. Currently, 70 treebanks covering about 50 languages can be downloaded freely[7].

UD Treebanks facilitate the development of an application that works potentially across all of the UD treebank languages in a uniform fashion, which is a big advantage for system developers. These treebanks are also a good basis for a multilingual shared task: a system that has been built for some of the languages may work for most of the other languages as well.

For the SR'18 Task, we will use a subset of the UD treebanks, selecting about 10 languages with an annotation of high quality, which provides PoS tags and morphological annotation (number, tense, verbal finiteness, etc.). A subset of at least 4 treebanks will be used for the Deep Track. The treebanks will be selected according to (i) the expertise of the task organizers in the corresponding language, (ii) the availability of native speakers for conversion and evaluation, (iii) the size of the treebank, (iv) the feasibility of the format conversion, (v) the variety of linguistic features captured in the annotation.

For the input to the Shallow Track, the UD structures will be processed as follows:

1. the information on word order will be removed by randomized scrambling;
2. the words will be replaced by their lemmas or stems, depending on the availability of lemmatization and stemming tools, respectively.

For the Deep Track, additionally:

3. functional prepositions and conjunctions that can be inferred from other lexical units or from the syntactic structure will be removed, as e.g., "by" and "of" in Figure 2;
4. determiners and auxiliaries will be replaced (when needed) by attribute/value pairs, as, e.g., "Definiteness" and "Aspect" in Figure 3;
5. edge labels will be generalized into predicate argument labels, following the PropBank/NomBank edge label nomenclature (Meyers and et al., 2004; Palmer et al., 2005), with three main differences: (i) there will be no special label for external arguments (i.e., no "A0"), which means that all first arguments of a predicate will be mapped to A1, and the rest of the arguments will be labeled starting from A2; (ii) all modifier edges "AM-..." will be generalized to "AM"; (iii) there will be a coordinative relation; and (iv) any relation that does not fall into the first three cases will be assigned an underspecified edge label.
6. morphological information coming from the syntactic structure or from agreements will be removed; in other words, only "semantic" information such as nominal number and verbal tense will be maintained in the Deep input, as opposed to verbal finiteness (which comes from the structure) or verbal number (which comes from agreement with the subject);

[5] In what follows, we refer to the proposed task as 'SR'18'.

[6] http://universaldependencies.org/#en

[7] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1983

7. fine-grained PoS labels found in some treebanks (as, e.g., column 5 in Figure 2) will be removed, and only coarse-grained ones will be maintained (column 4 in Figures 2 and 3).

The idea beyond the Deep Track is to make the input closer to a real-life input to NLG systems, in which no syntactic or language-specific information is available (see, e.g., the inputs in the SemEval, WebNLG, E2E shared tasks), while keeping it relatively simple. The main differences between the proposed Deep input and AMRs are the following: (i) no linking with NE databases; (ii) no abstraction of nominal VS verbal events; (iii) no OntoNotes labeling; (iv) no shared arguments; (v) no typed circumstancials.

The inputs to the Shallow and Deep Tracks will be distributed in the CoNLL-U format[8], and in the Human-Friendly Graph (HFG) format, as in SR'11 (Belz and et al., 2011). Figures 1, 2 and 3 show a sample original UD annotation for English, a sample input for the Shallow Track, and a sample input for the Deep Track respectively, in the 10-column CoNLL-U format.

3 Evaluation

We will perform both automatic and manual evaluations of the outputs of the systems.

For the automatic evaluation, we will compute scores with the following metrics:

1. BLEU as geometric mean of 1 to 4-grams with smoothing to compute sentence level scores,
2. NIST n-gram similarity weight,
3. METEOR lexical similarity based on stem, synonym and paraphrase matches.

We will apply text normalization before scoring. For n-best ranked system outputs, we will compute a single score for all outputs by computing the weighted sum of their individual scores, with a weight assigned to an output in inverse proportion to its rank. For a subset of the test data we may obtain additional alternative realizations via Mechanical Turk for use in the automatic evaluations.

For the human-assessed evaluation, we are planning to use a type of evaluation that is based on preference judgements (Kow and Belz, 2012, p.4035), using the existing evaluation interface described in Kow and Belz's paper. As in SR'11, we plan to use students in the third year of an undergraduate degree, from Cambridge, Oxford and Edinburgh. Two candidate outputs[9] will be presented to the evaluators, who will assess them for Clarity, Fluency and Meaning Similarity. For each criterion, they will be asked not only to state which system output they prefer, but also how strong is their preference.

We plan to organize a workshop collocated with ACL '18, COLING '18, or EMNLP '18 at which the results of the SR'18 will be presented. To ensure a smooth setup of the Shared Task and a swift evaluation of the system outputs, the organizers will contribute with their research funds. Furthermore, Google sponsorship will be solicited.

4 Conclusion

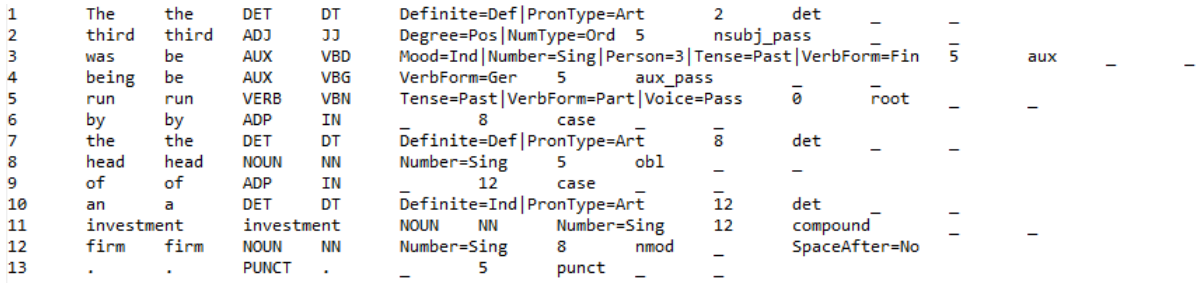
With this shared task, we aim to continue a very successful first shared task on surface realization. We think it is a good moment to take this topic up again due to emerging new techniques and system designs, new available data sets that can be used as basis for data-preparation, and a broad interest in deep generation techniques that emerges from new applications such as chat bots and personal assistants. We hope to attract a number of submissions within these application contexts (not only from the generation, but also, for instance, from the parsing community) and deepen the interest in text generation.

Beyond the possible impact of the tools developed in the context of this shared task due to the standard input sets and thus their easier reuse, we also see the shared task as an interesting experiment on the usability of UDs in the context of NLG. Our secondary objective is to assess how feasible it is to connect UD representations to predicate argument structures commonly used in deep NLG systems.

A valuable by-product of the shared task will be a set of input structures derived from UD data on a shallow and deep levels, which will be useful for further system development, application and research.

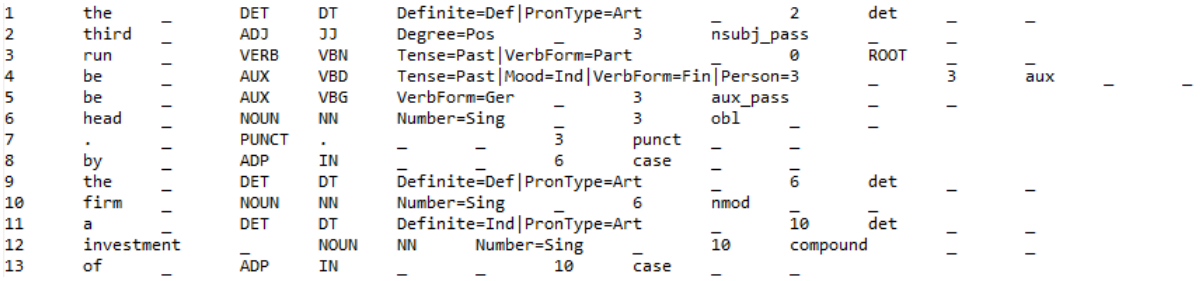
[8] http://universaldependencies.org/format.html

[9] Candidate outputs can also include a gold sentence, in addition to the system output.



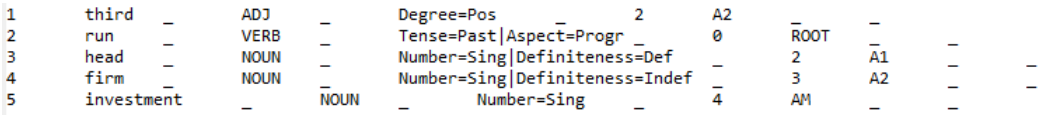
1	The	the	DET	DT	Definite=Def|PronType=Art	2	det	_	_		
2	third	third	ADJ	JJ	Degree=Pos|NumType=Ord	5	nsubj_pass	_	_		
3	was	be	AUX	VBD	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	5	aux	_	_		
4	being	be	AUX	VBG	VerbForm=Ger	5	aux_pass	_	_		
5	run	run	VERB	VBN	Tense=Past|VerbForm=Part|Voice=Pass	0	root	_	_		
6	by	by	ADP	IN	_	8	case	_	_		
7	the	the	DET	DT	Definite=Def|PronType=Art	8	det	_	_		
8	head	head	NOUN	NN	Number=Sing	5	obl	_	_		
9	of	of	ADP	IN	_	12	case	_	_		
10	an	a	DET	DT	Definite=Ind|PronType=Art	12	det	_	_		
11	investment	investment	NOUN	NN	Number=Sing	12	compound	_	_		
12	firm	firm	NOUN	NN	Number=Sing	8	nmod	_	SpaceAfter=No		
13	.	.	PUNCT	.	_	5	punct	_	_		

Figure 1: A sample UD structure in English



1	the	_	DET	DT	Definite=Def|PronType=Art	_	2	det	_	_	
2	third	_	ADJ	JJ	Degree=Pos	_	3	nsubj_pass	_	_	
3	run	_	VERB	VBN	Tense=Past|VerbForm=Part	_	0	ROOT	_	_	
4	be	_	AUX	VBD	Tense=Past|Mood=Ind|VerbForm=Fin|Person=3	_	3	aux	_	_	
5	be	_	AUX	VBG	VerbForm=Ger	_	3	aux_pass	_	_	
6	head	_	NOUN	NN	Number=Sing	_	3	obl	_	_	
7	.	_	PUNCT	.	_	_	3	punct	_	_	
8	by	_	ADP	IN	_	_	6	case	_	_	
9	the	_	DET	DT	Definite=Def|PronType=Art	_	6	det	_	_	
10	firm	_	NOUN	NN	Number=Sing	_	6	nmod	_	_	
11	a	_	DET	DT	Definite=Ind|PronType=Art	_	10	det	_	_	
12	investment	_	NOUN	NN	Number=Sing	_	10	compound	_	_	
13	of	_	ADP	IN	_	_	10	case	_	_	

Figure 2: A sample Shallow input



1	third	_	ADJ	_	Degree=Pos	_	2	A2	_	_
2	run	_	VERB	_	Tense=Past|Aspect=Progr	_	0	ROOT	_	_
3	head	_	NOUN	_	Number=Sing|Definiteness=Def	_	2	A1	_	_
4	firm	_	NOUN	_	Number=Sing|Definiteness=Indef	_	3	A2	_	_
5	investment	_	NOUN	_	Number=Sing	_	4	AM	_	_

Figure 3: A sample Deep input

5 Proposed Timeline

Assuming that the presentation of the results will not take place before mid-July 2018, the proposed timeline for the shared task would be the following:

- **Oct 1, 2017**: Completion of the consultation process regarding SR'18 input specifications and concerned languages.
- **Oct 1–Dec 8, 2017**: Implementation of conversion scripts and production of new inputs.
- **Oct 6, 2017**: Announcement of SR'18 and website.
- **Nov 13, 2017**: Call for interest in participation in SR'18.
- **Nov 13, 2017**: SR'18 Trial datasets and documentation.
- **Dec 11, 2017**: Registration for the task.
- **Dec 11, 2017**: SR'18 training and development sets.
- **April 2, 2018**: Evaluation scripts available.
- **May 14, 2018**: SR'18 test sets available.
- **May 18, 2018**: SR'18 system outputs collected.
- **May 21–June 30, 2018**: Evaluation period.

References

Anja Belz and Mike White et al. 2011. The first Surface Realisation Shared Task: Overview and evaluation results. In *Proceedings of ENLG*, pages 217–226, Nancy, France.

Eric Kow and Anja Belz. 2012. Lg-eval: A toolkit for creating online language evaluation experiments. In *Proceedings of LREC*, pages 4033–4037, Istambul, Turkey.

Jonathan May and Jay Priyadarshi. 2017. Semeval-2017 task 9: Abstract meaning representation parsing and generation. In *Proceedings of SemEval*, pages 534–543, Vancouver, Canada, August. Association for Computational Linguistics.

Adam Meyers and Ruth Reeves et al., 2004. *The NomBank project: An interim report*, pages 24–31.

Joakim Nivre and Marie-Catherine de Marneffe et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*, Portorož, Slovenia.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.