LORIA / Lorraine University at Multilingual Surface Realisation 2019

Anastasia Shimorina
LORIA / Lorraine University
anastasia.shimorina@loria.fr

Claire Gardent
LORIA / CNRS
claire.gardent@loria.fr

Abstract

This paper presents the LORIA / Lorraine University submission at the Multilingual Surface Realisation shared task 2019 for the shallow track. We outline our approach and evaluate it on 11 languages covered by the shared task. We provide a separate evaluation of each component of our pipeline, concluding on some difficulties and suggesting directions for future work.

1 Introduction

SR'19 (Mille et al., 2019) is the second edition of the multilingual surface realisation task ran in 2018 (Mille et al., 2018). It aims at developing surface realisers in the multilingual setting. Given an input tree, a well-formed sentence should be produced. The input tree can be either an unordered dependency tree (shallow track), or a tree with a predicate-argument structure (deep track).

The predecessor of the task is SR'11 (Belz et al., 2011), which dealt with surface realisation for English using data from Penn Treebank. Then, most approaches were based on statistical and rule-based methods. In SR'18, most participants used neural-based components, however, most teams (7 out of 8) used a pipeline approach, where they dealt separately with word ordering and morphological inflection (Basile and Mazzei, 2018; Castro Ferreira et al., 2018; Elder and Hokamp, 2018; King and White, 2018; Madsack et al., 2018; Puzikov and Gurevych, 2018; Singh et al., 2018; Sobrevilla Cabezudo and Pardo, 2018).

In this paper, we present a brief overview of the LORIA / Lorraine University system. We participated in the shallow track, and delivered solutions for all the languages proposed by the organisers. We also participated in generating output for all the types of corpora: in-domain, out-of-domain, and predicted by syntax parsers. Results on the development set are presented, and the system performance for each step of surface realisation is evaluated and discussed. All the code and experiments are available at https://gitlab.com/shimorina/msr-2019.

2 Data

We used only the data provided by the organisers. If several corpora were available for a language, they were mixed to the one training and development dataset. We used original UD files for creating target files for training the word ordering component, i.e. we extracted a sequence of tokens (the field `token` in the CoNLL format) instead of using a reference sentence.

3 Model

We made use of the model introduced in Shimorina and Gardent (2019) with some slight modifications. This model, developed for the SR'18 shared task data, is a pipeline approach to the surface realisation task, which has separate modules for word ordering, morphological inflection, and contraction generation. A brief outline is provided below; for more details about the model, we refer the reader to Shimorina and Gardent (2019).

3.1 Word Ordering (WO)

Word ordering is modelled as a sequence-to-sequence task, where an input tree is linearised. Linearisation differs from our previous approach in that it was augmented with information about the relative order of some elements, a feature that was introduced for this year edition of the shared task. So nodes were linearised using the depth-first search, and then elements with the relative order feature were reordered to match the added information.

Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019), pages 88–93
Hong Kong, China, November 3rd, 2019. ©2019 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

All input lemmas were delexicalised, i.e. replaced by identifiers both in the source and target, and enriched with features, or factors. A neural, factored encoder-decoder model was trained for each language, where factors are dependency relations, POS tags, and parent node identifiers (Elder and Hokamp, 2018; Alexandrescu and Kirchhoff, 2006).

During relexicalisation, all the identifiers were replaced by inflected lemmas. For the word ordering evaluation, we also relexicalised identifiers using the corresponding lemmas (see Section 4).

3.2 Morphological Realisation (MR)

Morphological paradigms were learned from pairs of (lemma, POS+features) extracted from the training data (the `upos` and `features` fields from CoNLL) using Aharoni and Goldberg (2017)'s model. Lemmas with no morphological features were not used. Since features are not provided for Chinese, Japanese, and Korean treebanks, the morphological realisation module was not trained for those languages. Instead, during the inflection phase (a) for Chinese, analytic language, lemmas were copied verbatim to the ouput; (b) for Korean, agglutinative language, morphemes in a lemma were glued together, and then the lemma was copied; (c) for Japanese, synthetic language, a dictionary of the form (lemma+POS: wordform) was constructed from the training data and looked up. If a key 'lemma+POS' was not present in the dictionary, the lemma was copied to the output verbatim. The same rule applies for any other lemma with no morphological features in any treebank (e.g., URLs, foreign words, numbers, punctuation signs, etc.)[1].

3.3 Contraction Generation (CG)

Contraction generation was implemented for French and Portuguese to handle clitic attachment, contractions, and elision. In the following, we will refer to the MR component as including the contraction generation module as well.

Eventually, one may also include detokenisation, a task of glueing tokens together, in this last step, as each language requires specific detokenisation rules to produce a final well-formed sentence, which can be shown to an end-user. We used the `sacremoses`[2] library to perform detokenisation. Besides, it was also used to tokenise reference sentences; we need that for the automatic scoring.

lang	Acc.	Amb. %	Amb. count
ar	90.87	7.29	1,815
en	96.35	0.84	226
es	98.85	0.85	418
fr	98.40	1.48	430
hi	89.95	6.46	569
id	98.52	0.55	47
ja	NA	3.62	800
ko	NA	0.86	945
pt	98.95	0.85	233
ru	97.25	0.72	933
zh	NA	0	0

Table 1: Accuracy of the morphological realisation component. NA: no MR component was developed. Percentage and count of lemmas with ambiguous forms found in the training data.

4 Results and Discussion

We evaluate each module separately. For WO, we compared a generated sequence of lemmas with a gold sequence of lemmas extracted from UD (Section 4.1). For MR, we calculated wordform prediction accuracy, and also applied MR to a gold sequence of lemmas instead of predicted sequence of lemmas (Section 4.2). Finally, we performed the overall evaluation, where our system predictions were compared to reference sentences (Section 4.3)[3].

4.1 WO Evaluation

Table 3 shows the results of WO. BLEU scores vary from 30 to 66 depending on the language and corpus ($mean = 56.98, median = 60.01$).

We surmised that low scores for Arabic, Chinese, Indonesian are due to small sizes of training corpora (6K, 4K, 4.5K, respectively), which are not enough for neural systems. Other languages' scores show a smaller variation, ranging from 51 to 66; we conjecture that the variations between languages are due to different syntactic phenom-

[1] We deleted features for foreign words in ru_gsd_ud for it to be consistent with ru_syntagrus_ud.

[2] https://github.com/alvations/sacremoses

[3] After the official submission, we fixed a bug in MR for Japanese and Korean. In this paper we are reporting improved results.

Corpus	BLEU	DIST	NIST
ar_padt-ud	40.07	47.23	8.25
en_ewt-ud	80.88	80.22	12.90
en_gum-ud	90.73	98.74	12.80
en_lines-ud	86.78	97.03	12.74
en_partut-ud	86.31	96.46	10.23
es_ancora-ud	93.82	98.47	14.88
es_gsd-ud	89.31	99.23	13.93
fr_gsd-ud	90.53	98.12	14.04
fr_partut-ud	87.07	96.61	9.82
fr_sequoia-ud	91.00	96.38	12.38
hi_hdtb-ud	91.88	96.89	13.67
id_gsd-ud	94.46	98.91	12.90
ja_gsd-ud	77.85	99.70	11.40
ko_gsd-ud	60.38	94.38	9.45
ko_kaist-ud	97.13	99.67	13.44
pt_bosque-ud	94.09	99.10	12.68
pt_gsd-ud	57.04	91.12	10.54
ru_gsd-ud	86.87	96.89	12.18
ru_syntagrus-ud	91.25	98.15	15.53
zh_gsd-ud	99.16	99.81	13.33

Table 2: The MR module applied to the gold word ordering input. Predictions and reference sentences are both tokenised. Results on the development set.

ena occurring in each language and the variations between corpora are due to different annotation guidelines.

4.2 MR+CG Evaluation

The inflection module was initially measured by accuracy of producing a correct word form given a lemma and its POS together with morphological features (cf. Table 1, second column). The average accuracy is 96.14 across 8 languages, which corresponds to the state-of-the-art results in inflection tasks (Cotterell et al., 2016).

We also calculated a number of lemmas, which can have different word forms, given the same set of POS and morphological features (Table 1, third column). For example, the lemma *people* with `pos=NOUN, Number=Plur` as features have two word forms in the training data: *people* and *peoples*. Those ambiguous forms may stem from different sources: language variation (as in the example above) including spelling, non-standard forms and typos; annotation mistakes; underspecified morphological features. The example of the latter is an adjective in Russian, which can have different forms in the accusative case depending on animacy of the noun it modifies (animacy in that case is an underspecified feature).

To measure the effect on scores, when converting a sequence of lemmas into a sentence, we applied MR+CG to gold sequences of lemmas (they have the same word order as the reference). Results are shown in Table 2. In general, high accuracies of MR alone (word level, Table 1) do not guarantee good performance while evaluating on the sentence level.

That type of evaluation enabled us to have more insight into the data used. Some of our findings are listed below.

- English: a discrepancy in performance across datasets. The sources of the en_ewt_ud corpus are blogs, social networks, reviews, emails, where the use of contractions (*isn't*, *ain't*, etc) is dominant comparing to formal style. Since the contraction generation was not applied for English, scores for this particular dataset are lower than for others.

- Arabic. We conjecture low scores for the high variability of forms (cf. Table 1) and contractions (We did not develop a module for handling contractions in Arabic.). For instance, some diacritics are optional (e.g., hamza with alif), so a word form can be written with or without them, being a valid word form in both cases.

- Japanese. MR module was not developed for Japanese, so a look-up dictionary based on training data was not sufficient to handle the morphology. The high number of ambiguous forms also impacted the scores, as in the case of Arabic.

- Portuguese. The pt_gsd_ud corpus is not annotated with morphological features, hence 57.04 score in BLEU compared to 94.09 in pt_bosque-ud.

- Korean. We do not read Korean, so we were not able to explain the difference between the two Korean corpora (97.13 vs. 60.38 BLEU). Some annotation disparity may well be the explanation.

4.3 Surface Realisation Evaluation

The performance of the overall surface realisation model is shown in Table 4. Automatic scores

Corpus	BLEU	DIST	NIST
ar_padt-ud	30.45	54.72	8.86
en_ewt-ud	66.71	84.18	12.57
en_gum-ud	62.92	80.61	11.53
en_lines-ud	61.89	75.76	11.67
en_partut-ud	62.38	75.87	9.82
es_ancora-ud	59.43	75.03	12.69
es_gsd-ud	61.83	74.94	12.80
fr_gsd-ud	60.58	78.66	12.74
fr_partut-ud	61.24	82.37	9.35
fr_sequoia-ud	55.22	74.18	10.86
hi_hdtb-ud	63.07	59.87	11.74
id_gsd-ud	46.09	76.07	9.99
ja_gsd-ud	56.53	62.41	10.33
ko_gsd-ud	53.73	53.01	11.69
ko_kaist-ud	66.43	63.19	12.85
pt_bosque-ud	52.88	81.98	11.13
pt_gsd-ud	51.01	72.59	11.82
ru_gsd-ud	59.11	62.30	11.72
ru_syntagrus-ud	62.37	67.88	14.03
zh_gsd-ud	45.67	56.01	10.23

Table 3: WO component performance on the development set. Predictions and references (sequences of lemmas) are both tokenised.

Corpus	BLEU	DIST	NIST
ar_padt-ud	18.06	43.86	6.49
en_ewt-ud	54.45	65.45	11.32
en_gum-ud	58.17	79.68	11.16
en_lines-ud	52.53	73.48	10.79
en_partut-ud	54.79	73.98	9.10
es_ancora-ud	56.99	73.78	12.53
es_gsd-ud	59.63	74.07	12.32
fr_gsd-ud	51.94	70.43	11.63
fr_partut-ud	51.72	74.74	8.47
fr_sequoia-ud	49.08	70.27	10.13
hi_hdtb-ud	58.48	61.13	11.42
id_gsd-ud	45.28	75.50	9.88
ja_gsd-ud	46.30	62.31	9.37
ko_gsd-ud	32.21	49.43	8.73
ko_kaist-ud	64.58	61.21	12.69
pt_bosque-ud	59.35	83.84	11.20
pt_gsd-ud	35.44	69.47	9.17
ru_gsd-ud	52.90	60.59	11.13
ru_syntagrus-ud	57.97	66.87	13.70
zh_gsd-ud	45.48	55.91	10.21

Table 4: Automatic metrics on the development set (WO + MR). Predictions and reference sentences are both tokenised.

show a drop compared to the WO component performance (Table 3), which is consistent with the errors of the MR+CG module, described in Section 4.2.

Figure 1 aggregates the BLEU scores, shown in Tables 2, 3, 4. For each corpus, BLEU for each module (X axis) is mapped to the final BLEU score (Y axis). The scatterplots show a strong, positive association between the two variables: Pearson's $\rho = 0.83$ and $\rho = 0.86$ for WO and MR on gold data respectively.

During test time, we also ran our system on out-of-domain and machine-generated data. For all languages concerned, automatic scores remain stable, which demonstrates the portability of our approach.

5 Conclusion

We presented the LORIA / Lorraine University submission to the SR'19 shared task. Our main takeaways are as follows. The WO component is easily transferrable between languages, and it will not require much effort for applying it to unseen languages. In contrast, the MR component requires a lot of attention, and needs to be tuned for each language separately. That is mainly due to the different approaches for language annotation across UD treebanks, and, what is more unexpected, across UD treebanks for the same language, not to speak of the detokenisation process, which is different for each language, and which should also be implemented separately.

Having those particularities in mind, we think that for future work MR (including contraction generation, and possibly detokenisation) would benefit for including context information, i.e. doing inflection and necessary character transformations on a whole sentence, rather than word by word. As for word ordering, it remains a tough problem for sequence-to-sequence architectures, and it is worth exploring other ways of encoding tree structure.

We also would like to highlight the importance of modular evaluation. If a system design allows it, system outputs may be tested against a sequence of lemmas, not only a reference sentence, thanks to the UD annotations. We encourage future participants not to neglect this type of evaluation to gain deeper insight into their system and data.

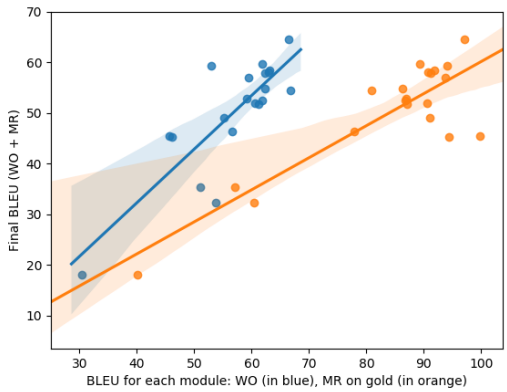

Figure 1: Linear regression between BLEU scores for each module and final BLEU scores. Data points are corpora. In orange: MR on gold data vs. final BLEU (WO + MR); in blue: WO vs. final BLEU (WO + MR).

References

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada. Association for Computational Linguistics.

Andrei Alexandrescu and Katrin Kirchhoff. 2006. Factored neural language models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 1–4. Association for Computational Linguistics.

Valerio Basile and Alessandro Mazzei. 2018. The dipinfo-unito system for srst 2018. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 65–71. Association for Computational Linguistics.

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 13th European Workshop on Natural Language Generation*, pages 217–226. Association for Computational Linguistics.

Thiago Castro Ferreira, Sander Wubben, and Emiel Krahmer. 2018. Surface realization shared task 2018 (sr18): The tilburg university approach. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 35–38. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Henry Elder and Chris Hokamp. 2018. Generating high-quality surface realizations using data augmentation and factored sequence models. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 49–53. Association for Computational Linguistics.

David King and Michael White. 2018. The osu realizer for srst '18: Neural sequence-to-sequence inflection and incremental locality-based linearization. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 39–48. Association for Computational Linguistics.

Andreas Madsack, Johanna Heininger, Nyamsuren Davaasambuu, Vitaliia Voronik, Michael Käufl, and Robert Weißgraeber. 2018. Ax semantics' submission to the surface realization shared task 2018. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 54–57. Association for Computational Linguistics.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018. The first multilingual surface realisation shared task (sr'18): Overview and evaluation results. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 1–12. Association for Computational Linguistics.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, and Leo Wanner. 2019. The Second Multilingual Surface Realisation Shared Task (SR'19): Overview and Evaluation Results. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR), 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, China.

Yevgeniy Puzikov and Iryna Gurevych. 2018. Binlin: A simple method of dependency tree linearization. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 13–28. Association for Computational Linguistics.

Anastasia Shimorina and Claire Gardent. 2019. Surface Realisation Using Full Delexicalisation. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, Hong Kong, China.

Shreyansh Singh, Ayush Sharma, Avi Chawla, and A.K. Singh. 2018. Iit (bhu) varanasi at msr-srst 2018: A language model based approach for natural language generation. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 29–34. Association for Computational Linguistics.

Marco Antonio Sobrevilla Cabezudo and Thiago Pardo. 2018. Nilc-swornemo at the surface realization shared task: Exploring syntax-based word ordering using neural models. In *Proceedings of*

the First Workshop on Multilingual Surface Realisation, pages 58–64. Association for Computational Linguistics.