The Third Multilingual Surface Realisation Shared Task (SR'20): Overview and Evaluation Results

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

Anya Belz
University of Brighton
a.s.belz@brighton.ac.uk

Bernd Bohnet
Google Inc.
bohnetbd@google.com

Thiago Castro Ferreira
Fed. Univ. Minas Gerais
thiago.castro.ferreira@gmail.com

Yvette Graham
ADAPT, Trinity College Dublin
ygraham@tcd.ie

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

Abstract

This paper presents the results of the Third Shared Task on Multilingual Surface Realisation (SR'20) which was organised as part of the COLING'20 Workshop on Multilingual Surface Realisation. As in SR'18 and SR'19, the shared task comprised two tracks: (1) a Shallow Track where the inputs were full UD structures with word order information removed and tokens lemmatised; and (2) a Deep Track where additionally, functional words and morphological information were removed. Moreover, each track had two subtracks: (a) restricted-resource, where only the data provided or approved as part of a track could be used for training models, and (b) open-resource, where any data could be used. The Shallow Track was offered in 11 languages, whereas the Deep Track in 3. Systems were evaluated using automatic metrics and direct assessment by human evaluators of Readability and Meaning Similarity to reference outputs. We present the evaluation results, along with descriptions of the SR'20 tracks, data and evaluation methods, as well as brief summaries of the participating systems. Full descriptions of the participating systems can be found in separate system reports elsewhere in this volume.

1 Introduction

SR'20 is the fourth in a line of shared tasks focused on surface realisation, the name originally given to the last stage in the first-generation (pre-statistical and pre-neural) Natural Language Generation (NLG) pipeline, mapping from semantic representations to fully realised surface word strings. When we ran the first Surface Realisation Shared Task in 2011 (Belz et al., 2011), it was to address a situation where there were many different approaches to SR but none of them were comparable. We developed a common-ground input representation that different approaches could map their normal inputs to, making results directly comparable for the first time. Most SR'11 systems (and all top performing ones) were statistical dependency realisers that did not make use of an explicit, pre-existing grammar. However, the question of how inputs to the realisers were going to be provided in an embedding system was left open.

By the time we proposed the second SR Task (Mille et al., 2017), Universal Dependencies (UDs) had emerged as a convenient standard in parsing, with many associated data sets, that we were able to pick up and use as the common-ground representation. By now, the third, neural generation of NLG methods was beginning to dominate the field, and systems participating in SR'18 were all trained to map directly from the UD inputs to the surface strings by some form of neural method. The question of how inputs to the realisers were going to be supplied remained open; moreover, most current approaches to NLG no longer even distinguished a separate surface realisation stage. Nevertheless, the community enthusiastically participated in SR'18 and SR'19 (Mille et al., 2018; Mille et al., 2019) as we expanded tracks to 11 languages.

This year, things look different again. There is much discussion in the field of how to control the vexed tendencies of neural generators to 'hallucinate' content (Dušek et al., 2019), and how to instil some order

Proceedings of the Third Workshop on Multilingual Surface Realisation (MSR'20), pages 1–20
Barcelona, Spain (Online), December 12, 2020.

and coherence over longer texts. Multi-hop approaches are increasingly proposed to address such issues (Hua and Wang, 2019; Zhai et al., 2019; Zhao et al., 2020), and are beginning to look somewhat like the old NLG pipeline. In this context, surface realisation is very much back on the agenda, and the term is coming back into frequent use (Zhai et al., 2019; Zhao et al., 2020). Our aim for future editions of the SR Shared Task is to test whether multi-hop gives better results overall than single-hop, but also to link up with content selection modules capable of supplying the inputs required by SR systems.

For this year, our main objective is to explore the impact of restricted vs. unrestricted resources in system training, and cross-domain generalisability. We start below with an overview of the shared task and tracks (Section 2), followed by descriptions of the participating systems (Section 3), the data (4), evaluation methods (Section 5), and results (Section 6).

2 Overview of Shared Task and Tracks

SR'20 uses the same languages and datasets as SR'19. There is a shallow and a deep track, as before; however, each track divides into two subtracks, one of which is retricted-mode, meaning only the data provided or approved for the given track may be used to train systems, the other an open track where any resources may be used in building systems. We have also created new test data sets derived from Wikipedia articles by the method described in Section 4.2 below. This year's set-up allows us to explore topline system performance and generalisability of results to a new domain.

The two main tracks are as follows:

T1 Shallow Track: The inputs in this track are UD structures in which most of the word order information has been removed and tokens have been lemmatised. In other words, it starts from unordered dependency trees with lemmatised nodes that hold PoS tags and morphological information as found in the original treebank annotations. The outputs are the fully realised sentences. The task in this track therefore amounts to determining the word order and inflecting words.

- **a. Restricted-resources subtrack (same as SR'19 Track 1)**: Teams built models trained on the provided T1 dataset(s), but use of external task-specific data was not permitted. However, teams were allowed to use external generic resources. For example, available parsers such as UUParser (Smith et al., 2018) could be run to create a silver standard versions of provided datasets and use them as additional or alternative training material. Also permitted was the use of generic publicly available off-the-shelf language models such as GPT-2 (Radford et al., 2019), ELMo (Peters et al., 2018), polyglot (Al-Rfou et al., 2013). Alternatively, BERT (Devlin et al., 2018) could be fine-tuned with publicly available datasets such as WikiText (Merity et al., 2016) or the DeepMind Q&A Dataset (Hermann et al., 2015).
- **b. Open subtrack**: In this track, teams built models trained on the provided T1 dataset(s), also using any additional resources, without restrictions. Teams could even use the SR conversion tool to produce data with the exact same specifications as the data provided in the track, by applying the converter to a parsed output (see Section 4.2).

T2 Deep Track: Inputs in this track are UD structures as in T1 from which functional words (in particular, auxiliaries, functional prepositions and conjunctions) and surface-oriented morphological and syntactic information have additionally been removed. The task in the Deep Track thus also involves introduction of functional words and morphological features, in addition to what is required for the Shallow Track.

- **a. Restricted-resources subtrack (same as SR'19 Track 2)**: Teams built models trained on the provided T2 dataset(s) using resources restricted exactly as described for T1-a above.
- **b. Open subtrack**: Teams built models trained on the provided T2 dataset(s), using additional resources without restrictions as described for T1-b above.

	2019 test sets				2020 Wiki test sets			
	T1		T2		T1		T2	
	a	b	a	b	a	b	a	b
ADAPT 20	english (ewt)	english (ewt)	–	–	english	english	–	–
BME-TUW 20	all (all)	–	–	–	all	–	–	–
Concordia 20	english (all)	–	english (all)	–	english	–	english	–
IMS 20	all (all)	all (all)	all (all)	all (all)	all	all	all	all
NILC 20	–	–	english (all)	–	–	–	english	–
BME-UW 19	*	N/A	*	N/A	all	N/A	–	N/A
IMS 19	*	N/A	*	N/A	all	N/A	all	N/A
RALI 19	–	N/A	–	N/A	english	N/A	english	N/A
Tilburg 19	–	N/A	–	N/A	all	N/A	–	N/A

Table 1: Teams submitting in each track/language (dataset); * = 2019 outputs included in human eval.

3 Participating Systems

There were two distinct sets of participating systems this year. Firstly, there were the new participants who built systems specifically for SR'20 (each system briefly summarised in Section 3.1). In addition, we asked the 2019 participants to run their systems on the new test sets, and 4 teams were able to do so (Section 3.2), two of which also submitted new systems to SR'20. Table 1 provides an overview of which teams submitted outputs in which (sub)tracks, languages and datasets. The 2019 systems can only contribute to the restricted track columns ('a') since it was the only mode of participation last year. We also indicate in the table which systems we reevaluated in the human evaluation (* in the table).

3.1 SR'20 new systems

The **ADAPT** system was trained using a custom fork of the OpenNMT-py framework, the only change made was to the beam search decoding code. The model used was a bidirectional recurrent neural network (BRNN) with long short term memory (LSTM) cells. Two variants of the ADAPT system were submitted; one trained with just the EWT dataset and one with both the EWT dataset and an augmented dataset constructed from the WikiText 103 and CNN stories corpora. (For all datasets, see Section 4.)

The **BME-TUW** system performs word order restoration by learning rules of an Interpreted Regular Tree Grammar (IRTG) that encodes the correspondence between UD-subgraphs and word orderings. The grammars build strings and UD graphs simultaneously, using pairs of operations that each connect some set of dependents to their common head while concatenating the corresponding words. The approach extends the team's 2019 system by allowing rules to reference lemmas in addition to POS-tags and by giving preference to derivations that use a smaller number of more specific rules to construct a particular UD graph. Word order restoration is performed separately for each clause. For the inflection step, a standard sequence-to-sequence model with biLSTM encoder and LSTM decoder with attention is used.

Concordia uses a text-to-text model to tackle graph-to-text surface realisation. The approach fine-tunes the pre-trained BART (Lewis et al., 2020) language model on the task of surface realisation where the model receives the linearised representation of the dependency tree and generates the surface text.

The **IMS** system builds on their system from the previous year with a substantial change in the linearisers proposed in (Yu et al., 2020), which models the task of word ordering as a Traveling Salesman Problem, and uses a biaffine attention model to calculate the bigram scores for the output sequence. To remedy the restriction of projectivity, it uses a transition system to reorder the sentence. Furthermore, model ensembling and data augmentation is applied to push the performance.

The **NILC** submission explores different ways to represent a UD structure linearly, and models the generation task by using the small version of GPT-2.

3.2 SR'19 systems run on the SR'20 new test sets

The **BME-UW** system (Kovács et al., 2019) performs word order restoration by learning Interpreted Regular Tree Grammar (IRTG) rules encoding the correspondence between UD-subgraphs and word orderings. The grammars build strings and UD graphs simultaneously, using pairs of operations each connecting a set of dependents to their common head while concatenating the corresponding words. Rule

Data type	Dataset	Track	train	dev	test
In-domain	arabic_padt (ar)	T1	6,075	909	680
	chinese_gsd (zh)	T1	3,997	500	500
	english_ewt (en)	T1, T2	12,543	2,002	2,077
	english_gum (en)	T1, T2	2,914	707	778
	english_lines (en)	T1, T2	2,738	912	914
	english_partut (en)	T1, T2	1,781	156	153
	french_gsd (fr)	T1, T2	14,450	1,476	416
	french_partut (fr)	T1, T2	803	107	110
	french_sequoia (fr)	T1, T2	2,231	412	456
	hindi_hdtb (hi)	T1	13,304	1,659	1,684
	indonesian_gsd (id)	T1	4,477	559	557
	japanese_gsd (ja)	T1	7,133	511	551
	korean_gsd (ko)	T1	4,400	950	989
	korean_kaist (ko)	T1	23,010	2,066	2,287
	portuguese_bosque (pt)	T1	8,328	560	477
	portuguese_gsd (pt)	T1	9,664	1,210	1,204
	russian_gsd (ru)	T1	3,850	579	601
	russian_syntagrus (ru)	T1	48,814	6,584	6,491
	spanish_ancora (es)	T1, T2	14,305	1,654	1,721
	spanish_gsd (es)	T1, T2	14,187	1,400	426
Out-of-domain	english_pud (en)	T1, T2	-	-	1,000
	japanese_pud (ja)	T1	-	-	1,000
	russian_pud (ru)	T1	-	-	1,000
Automatically parsed	english_ewt$_{HIT}$ (en)	T1, T2	-	-	1,795
	english_pud$_{LAT}$ (en)	T1, T2	-	-	1,032
	hindi_hdtb$_{HIT}$ (hi)	T1	-	-	1,675
	korean_kaist$_{HIT}$ (ko)	T1	-	-	2,287
	portuguese_bosque$_{STF}$ (pt)	T1	-	-	471
	spanish_ancora$_{HIT}$ (es)	T1, T2	-	-	1,723
Automatically parsed Wikipedia	english_wiki$_{STZ}$(en)	T1, T2	-	-	1,313
	french_wiki$_{STZ}$ (fr)	T1, T2	-	-	1,313
	korean_wiki$_{STZ}$ (ko)	T1	-	-	530
	portuguese_wiki$_{STZ}$ (pt)	T1	-	-	1,135
	russian_wiki$_{STZ}$ (ru)	T1	-	-	1,291
	spanish_wiki$_{STZ}$ (es)	T1, T2	-	-	1,280

Table 2: SR'20 dataset sizes for training, development and test sets (number of sentences).

weights are proportional to the observed frequency of each pattern in the training data. The inflection step uses a standard sequence-to-sequence model with biLSTM encoder and LSTM decoder with attention.

IMS (Yu et al., 2019) uses a pipeline approach for both tracks, consisting of linearisation, completion (for T2 only), inflection, and contraction. All models use the same bidirectional Tree-LSTM encoder architecture. The linearisation model orders each subtree separately with beam search, then combining the trees into a full projective tree; the completion model generates absent function words sequentially given the linearised tree of content words; the inflection model predicts a sequence of edit operations to convert lemmas to word forms character by character; the contraction model predicts BIO tags to group words to be contracted, and then generates the contracted word form of each group with a seq2seq model.

The **RALI** system (Lapalme, 2019) uses a symbolic approach to transform the dependency tree into a tree of constituents that is transformed into an English sentence by an existing English realiser, JSrealB (Molins and Lapalme, 2015). This realiser was then slightly modified for the two tracks.

The **Tilburg** approach (Ferreira and Krahmer, 2019), based on Ferreira et al. (2018), realises texts by first preprocessing the dependency tree into a preordered linearized form, which is then converted into its textual counterpart using a rule-based approach together with a statistical machine translation (SMT) model. A singular version of the model was trained for each language considered in the experiment.

4 Data Sets

4.1 T1 and T2 training and test sets (same as in SR'19)

There are 42 datasets in 11 languages, 29 datasets for T1, and 13 for T2 (for a summary overview, see Table 2, top 3 sections of the table). The datasets were selected from the available collection of

Language	Dataset	Performance (LAS)	Performance (lemmas)	Best CoNLL'18 (LAS)	Best CoNLL'18 (lemmas)
English	ewt	83.59	97.21	84.57	97.23
French	gsd	89.05	97.64	86.89	97.03
Korean	gsd	83.53	92.69	85.14	94.02
Portuguese	bosque	87.57	97.8	87.81	97.54
Russian	syntagrus	90.06	97.51	92.48	98.19
Spanish	ancora	90.01	99.19	90.93	99.02

Table 3: Datasets used to train the Stanza parsing models.

UD datasets mainly based on the completeness of annotations in terms of PoS tags and morphologically relevant markup (number, tense, verbal finiteness, etc.). The test data sets can be grouped into three types: (i) in-domain test data, in the same domains as the training and development data; (ii) Out-of-domain, which are test sets of parallel sentences in different languages in domains not covered by the training and development data; and (iii) silver standard data, which consists of automatically parsed sentences.

The in-domain and out-of-domain data is provided in the UD release V2.3.[1] The silver standard data was processed using the best CoNLL'18 parsers for the chosen datasets: the Harbin HIT-SCIR (*HIT*) parser (Che et al., 2017) for English_ewt, Hindi_hdtb, Korean_kaist and Spanish_ancora; the LATTICE (*LAT*) parser (Lim et al., 2018) for English_pud and the Stanford (*STF*) parser (Qi et al., 2019) for Portuguese_bosque.[2] A detailed description of all SR'19 datasets and how they were processed can be found in the SR'19 report paper (Mille et al., 2019).

4.2 SR'20 new test sets

To obtain new test sets,[3] we selected sentences from Wikipedia in six out of the eleven SR'19 languages for which it was possible to get a good quantity of clean texts on the same topics. The used articles contain mostly landmarks and some historical figures. On the extracted sentences, we applied extensive filtering to achieve reasonably good text quality. We skipped sentences that include special characters, contain unusual tokens (e.g. ISBN), or have unbalanced quotation marks or brackets. Furthermore, we took only sentences with more than 5 tokens and shorter than 50 tokens. After the initial filtering, quite a few malformed sentences remained. In order to remove those, we scored the sentences with BERT and kept only the best scored half. Finally, via manual inspection we identified patterns and expressions to reduce the number of malformed sentences still further.

We parsed the cleaned Wikipedia sentences with the Stanza parser (Qi et al., 2020), using the trained models provided for the respective languages; the Stanza parser gets very competitive results on a large set of languages (see Table 3). For each language, we executed the parser with the processors for Tokenisation and Sentence Split, Multi-word Token Expansion, Part-of-Speech and Morphological Tagging, Lemmatisation and Dependency Parsing. The performance of the parser for all six languages in terms of Labelled Attachment Score and lemmatisation, two of the crucial aspects for our task, is provided in Table 3; for reference, we also provide the LAS and lemma scores of the best parser on each dataset according to the CoNLL'18 shared task results. All the datasets and their respective sizes are summarised in Table 2; the *STZ* extension in the last 6 rows of the table indicate a reference to the Stanza parser.

As it was the case in the previous editions of the task, Shallow Track inputs were generated with the aid of Python scripts from the UD structures, using all available input sentences (inflected forms and most word order information are removed), and Deep Track inputs were then generated by automatically processing the Shallow Track structures using a series of graph-transduction grammars for removing functional nodes and other superficial features, and generalising the dependency relations; see SR'19 report (Mille et al., 2019) for details. The code for converting the UD trees into SR'19/SR'20 Shallow

[1] https://universaldependencies.org/

[2] The CoNLL'18 shared task submissions were downloaded from https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2885.

[3] The complete SR'20 datasets can be downloaded from https://sites.google.com/site/genchalrepository/surface-realisation/sr-20-multilingual

1	Will	will	AUX	MD	VerbForm=Fin	6	aux
2	it	it	PRON	PRP	Case=Nom|Gender=Neut|Number=Sing|Person=3|PronType=Prs	6	nsubj
3	not	not	PART	RB	_	6	advmod
4	also	also	ADV	RB	_	6	advmod
5	be	be	AUX	VB	VerbForm=Inf	6	cop
6	grandiose	grandiose	ADJ	JJ	Degree=Pos	0	root
7	in	in	ADP	IN	_	9	case
8	its	its	PRON	PRP$	Gender=Neut|Number=Sing|Person=3|Poss=Yes|PronType=Prs	9	nmod:poss
9	way	way	NOUN	NN	Number=Sing	6	obl
10	?	?	PUNCT	.	_	6	punct

Figure 1: Sample UD structure (without the last two columns).

1	not	_	PART	RB	_	10	advmod
2	its	_	PRON	PRP$	Gender=Neut|Number=Sing|Person=3|Poss=Yes|PronType=Prs	7	nmod:poss
3	?	_	PUNCT	.	lin=+1	10	punct
4	be	_	AUX	VB	VerbForm=Inf	10	cop
5	it	_	PRON	PRP	Case=Nom|Gender=Neut|Number=Sing|Person=3|PronType=Prs	10	nsubj
6	also	_	ADV	RB	_	10	advmod
7	way	_	NOUN	NN	Number=Sing	10	obl
8	in	_	ADP	IN	_	7	case
9	will	_	AUX	MD	VerbForm=Fin	10	aux
10	grandiose	_	ADJ	JJ	Degree=Pos	0	root

Figure 2: Sample T1 input structure (without the last two columns).

1	not	_	PART	_	_	2	A1INV
2	grandiose	_	ADJ	_	Tense=Fut|Degree=Pos|ClauseType=Int	0	ROOT
3	it	_	PRON	_	Number=Sing|Person=3|PronType=Prs	2	A1
4	also	_	ADV	_	_	2	A1INV
5	way	_	NOUN	_	Number=Sing	2	AM
6	its	_	PRON	_	Number=Sing|Poss=Yes|Person=3|PronType=Prs	5	AM

Figure 3: Sample T2 input structure (without the last two columns).

and Deep Track inputs is available on GitLab.[4] Figures 1, 2 and 3 shown sample UD, Track 1 and Track 2 structures respectively, taken from the parsed Wikipedia English dataset.

5 Evaluation Methods

5.1 Automatic methods

We used BLEU, NIST, BERT, and inverse normalised character-based string-edit distance (referred to as DIST, for short, below) to assess submitted systems. BLEU (Papineni et al., 2002) is a precision metric that computes the geometric mean of the n-gram precisions between generated text and reference texts and adds a brevity penalty for shorter sentences. We use the smoothed version and report results for $n = 4$. NIST[5] is a related n-gram similarity metric weighted in favor of less frequent n-grams which are taken to be more informative. DIST starts by computing the minimum number of character inserts, deletes and substitutions (all at cost 1) required to turn the system output into the (single) reference text. The resulting number is then divided by the number of characters in the reference text, and finally subtracted from 1, in order to align with the other metrics. Spaces and punctuation marks count as characters; output texts were otherwise normalised as for all metrics (see below). BERTScore (Zhang et al., 2020) computes a token-based similarity score by comparing each token of the generated texts with each token of the reference sentence. BERTScore uses contextual embeddings rather than exact matches, and has been shown to correlate better with human judgments than other commonly used metrics. The figures in the tables below are the system-level scores for BLEU, NIST and BERTScore, and the mean sentence-level scores for DIST.

Output texts were normalised prior to computing metrics by lower-casing all tokens, removing any extraneous whitespace characters. Missing outputs were scored 0. We only report results for all sentences (incorporating the missing-output penalty), rather than also separately reporting scores for just the in-coverage items.

[4] https://gitlab.com/talnupf/ud2deep

[5] http://www.itl.nist.gov/iad/mig/tests/mt/doc/ngram-study.pdf; http://www.itl.nist.gov/iad/mig/tests/mt/2009/

5.2 Human-assessed methods

For the human evaluation, we selected a subset of language/dataset combinations based on number of submissions received and availability of evaluators: three in-domain datasets (English_ewt, Russian_syntagrus, Spanish_ancora), and the three corresponding Wikipedia datasets for these languages. All submitted Track 1 and Track 2 outputs for these datasets were evaluated, plus two 2019 outputs (IMS, BME-UW) for each in-domain dataset, which were already evaluated in 2019.

We adopted the same approach to human evaluation as in SR'18 (Mille et al., 2018) and SR'19 (Mille et al., 2019). The evaluation method is Direct Assessment (DA) (Graham et al., 2016), as used by the WMT competitions to produce the official ranking of machine translation systems (Barrault et al., 2020) and video captioning systems at TRECvid (Graham et al., 2018; Awad et al., 2019). We ran the evaluation on Mechanical Turk,[6] assessing two quality criteria, in separate evaluation experiments, but using the same method: *Readability* and *Meaning Similarity*. We used continuous sliders as rating tools, the evidence being that raters tend to prefer them (Belz and Kow, 2011). Slider positions were mapped to values from 0 to 100 (best).

Raters were given brief instructions, including the direction to ignore formatting errors, superfluous whitespace, capitalisation issues, and poor hyphenation. The statement to be assessed in the Readability evaluation was: *The text reads well and is free from grammatical errors and awkward constructions.*

The corresponding statement in the Meaning Similarity evaluation, in which system outputs ('the black text') were compared to reference sentences ('the gray text'), was:[7,8] *The meaning of the gray text is adequately expressed by the black text.*

The DA method involves quality assurance techniques as follows. System outputs are randomly assigned to HITs (following Mechanical Turk terminology) of 100 outputs, of which 20 are used solely for quality assurance (QA) (i.e. do not count towards system scores): (i) some are repeated as-is, (ii) some are repeated in a 'damaged' version and (iii) some are replaced by their corresponding reference texts. In each case, a minimum threshold has to be reached for the HIT to be accepted: for (i), scores must be similar enough, for (ii) the score for the damaged version must be worse, and for (iii) the score for the reference text must be high. For full details of how these additional texts are created and thresholds applied, please refer to Barrault et al. (2019). We report QA figures for the MTurk evaluations below.

Test set sizes out of the box varied for the different languages. For the human test sets we selected a subset of at least 500 sentences for each language, motivated by the power analysis provided by Graham et al. (2019). For subsets, test set items were selected randomly.

We follow the same format for reporting results as WMT adopt when reporting DA method results, i.e. we report both average raw scores and average standardised scores per system in the tabular form shown in the results tables below. In order to produce standardised scores we simply map each individual evaluator's scores to their standard scores (or z-scores) computed on the set of all raw scores by the given evaluator using each evaluator's mean and standard deviation. For both raw and standard scores, we compute the mean of sentence-level scores.

6 Results

In this section, we present evaluation results produced with all evaluation methods from the preceding section, for all test set outputs received from participants. The best scores are generally better this year compared to 2019, both according to automatically computed metrics and human assessments. By way of introduction, Table 4 shows a sample output for one of the English Wikipedia sentences, as generated by each participating system. T1 and T2 inputs for these sample outputs are shown above in Figures 2 and 3 respectively. Interestingly, the outputs show a lot of variation, and none of them gets the target exactly. The last column shows the percentage of sentences exactly matching their human-written reference for each system, as calculated on the English_wiki dataset (1,313 sentences).

[6] We were able to reuse, with minor adaptations, the code produced for the WMT'17 evaluations: `https://github.com/ygraham/segment-mteval`

[7] Since a main proportion of workers on Mechanical Turk are located in the US, we employ US spelling in evaluations.

[8] Past work in machine translation has investigated the degree to which the presence of a reference sentence might introduce

System	Sample output	Exact (%)
HUMAN	Will it not also be grandiose in its way?	
ADAPT20aT1	In its way it will alson't be grandiose?	15.8
ADAPT20bT1	Will it also not be grandiose in its way?	27.11
BME19T1	It will also be not grandiose in its way ?	0.57
BME20aT1	Also be it will not grandiose in its way ?	0.57
Concordia20aT1	Will it also not be grandiose in its way ?	0.46
Concordia20aT2	It will also not be grandiose in its way .	0.42
IMS19T1	Will not it also be grandiose in its way?	16.41
IMS19T2	It should also not grandiose in its way?	1.18
IMS20aT1	It will not be also grandiose in its way?	18.74
IMS20aT2	It will also not be grandiose in its way?	1.56
IMS20bT1	It will also not be grandiose in its way?	21.59
IMS20bT2	Will it also not be grandiose in its way?	2.17
NILC20aT2	It will also be n't in its way;	1.56
RALI19T1	It? will not be grandiose also in its way.	0.46
RALI19T2	It is grandiose not also its way.	0.46
Tilburg19T1	it will also be not grandiose in its way.	0.23

Table 4: Sample system outputs for the inputs in Figures 2 and 3, and % of exact matches on English_wiki (systems in alphabetical order).

6.1 Automatic Evaluation Results

6.1.1 Overview and of metric results provided

Tables 5, 6, 7, 8, 9, 10, 11 and 12 show the results of the automatic evaluations with BLEU, NIST, DIST, and BERTScore on all test sets. We have grouped results tables together by metric, so that the first page of results shows all BLEU results, in Tables 5 and 6; the second page of results shows all NIST results, in Tables 7 and 8; the third page of results shows all DIST results, in Tables 9 and 10; and the fourth page shows all BERT results, in Tables 11 and 12. In each case, the first, larger, table shows results for the 2019 test sets, whereas the second, smaller table shows results for the new 2020 Wikipedia test sets. In each table, the column headings show the system team, system year (20, 19) and subtrack (a, b) for which results are shown in a column, while the row labels in the first column show which test set and track (T1, T2) results are for. Rows are shown in alphabetical order of the test set name (ar_padt, en_ewt, etc.). For an overview of test sets, see Table 2.

In Section 6.3, we furthermore provide comparisons between automatic and human evaluations.

6.1.2 Discussion of metric results

Considering all metric results tables together, scores are generally (but not always) higher for 2020 systems than for comparable 2019 systems (e.g. BLEU for **BME20a** is higher for most test sets than for **BME19**, Table 5); and T1 results are higher in all cases than directly comparable T2 results. The picture for 'a' subtracks (restricted) compared to 'b' subtracks (unrestricted) is different. Here we have two teams who submitted comparable outputs for both subtracks, **ADAPT** and **IMS**. **ADAPT** submitted just for T1 en_ewt, and here, all results for 'b' are higher than comparable results for 'a'. **IMS** submitted for all datasets in both subtracks 'a' and 'b' in both T1 and T2. For the 2019 tests, the picture is mixed, and there is no clear, consistent benefit from additional resources. However, for the 2020 out-of-domain Wikipedia test sets, 'b' scores are always greater than (or in one case equal to) comparable 'b' scores, with the margin bigger for T2 scores than T1.

Taking a closer look at improvements this year compared to 2019, we see for instance, on the **English_ewt** test set, last year's top BLEU score in T1 (the Shallow Track) was 82.98 (**IMS**); in 2020, it goes up to 86.16 in the restricted track (**IMS**), and 87.5 in the open track (**ADAPT**). In T2 (the Deep Track), top BLEU scores also increased, from 54.75 (**IMS**) to 58.84 in the restricted track, and 58.66 in the unrestricted track (both **IMS**).

We next look at overall improvements of team submissions across all test sets they submitted outputs

bias into the evaluation revealing no significant evidence of reference-bias (Ma et al., 2017).

–BLEU-4–	ADAPT		BME		Concordia	IMS			NILC
	20a	20b	20a	19	20a	20a	20b	19	20a
T1_ar_padt			26	26.4		69.56	69.71	64.9	
T1_en_ewt	80.4	87.5	57.25	59.22	70.71	86.16	85.67	82.98	
T2_en_ewt					58.44	58.84	58.66	54.75	45.19
T1_en_gum			60.77	57.57	66.98	88.89	89.7	83.84	
T2_en_gum					53.92	53.98	56.33	52.45	40.94
T1_en_lines			55.98	48.78	62.7	85.05	85.3	81	
T2_en_lines					47.96	50.23	50.45	47.29	41.04
T1_en_partut			57.96	61.37	67.05	89.72	89.37	87.25	
T2_en_partut					50.54	46.87	50.11	45.89	43.41
T1_es_ancora			59.32	61.09		87.42	87.34	83.7	
T2_es_ancora						56.67	55.64	53.13	
T1_es_gsd			54.6	53.74		84.61	84.52	82.98	
T2_es_gsd						55.1	55.99	51.17	
T1_fr_gsd			43.21	43.8		86.08	85.08	84	
T2_fr_gsd						58.86	56.95	53.62	
T1_fr_partut			52.46	49.17		87.09	89.22	83.38	
T2_fr_partut						51.11	57.62	46.95	
T1_fr_sequoia			45.25	46.72		87.25	87.29	85.01	
T2_fr_sequoia						59.37	60.26	57.41	
T1_hi_hdtb			57.2	63.63		84.53	84.77	80.56	
T1_id_gsd			59.16	54.22		87.53	88.33	85.34	
T1_ja_gsd			50.89	49.53		89.36	89.54	87.69	
T1_ko_gsd			58.37	46.08		81.14	82.52	74.19	
T1_ko_kaist			57.05	47.23		79.96	80.28	73.93	
T1_pt_bosque			39.89	39.53		82.92	83.36	77.75	
T1_pt_gsd			30.68	30.39		80.59	80.69	75.93	
T1_ru_gsd			54.28	54.58		77.43	78.93	71.23	
T1_ru_syntagrus			54.79	50.91		81.82	79.78	76.95	
T1_zh_gsd			50.58	58.72		86.36	88.05	83.85	
T1_en_pud			58.67	60.42	74.47	85.37	85.65	86.61	
T2_en_pud					58.45	50.59	53.7	51.01	42.6
T1_ja_pud			51.08	53.65		88.88	89.21	86.64	
T1_ru_pud			46.07	10.15		69.07	69.35	58.38	
T1_en_ewt$_{HIT}$			55.5	58.07	67.12	84.72	84.31	81.8	
T2_en_ewt$_{HIT}$					56.01	57.1	57.14	53.54	43.15
T1_en_pud$_{LAT}$			54.76	53.46	73.41	79.74	80.14	82.6	
T2_en_pud$_{LAT}$					56.69	47.96	50.15	47.6	42.64
T1_es_ancora$_{HIT}$			59.7	61.26		86.97	86.81	83.31	
T2_es_ancora$_{HIT}$						56.92	55.96	53.54	
T1_hi_hdtb$_{HIT}$			56.83	64.27		84.42	84.78	80.19	
T1_ko_kaist$_{HIT}$			56.74	46.72		81.01	81.42	74.27	
T1_pt_bosque$_{STA}$			41.86	40.42		84.35	85.18	78.97	
Macro-avg	80.4	87.5	51.96	50.04	61.75	74.56	75.13	71.13	42.71

Table 5: BLEU scores on the 2019 datasets, with indicative average scores on the submitted outputs.

–BLEU-4–	ADAPT		BME		Concordia	IMS			NILC	RALI	Tilburg
	20a	20b	20a	19	20a	20a	20b	19	20a	19	19
T1_en_wiki$_{STZ}$	84.11	94.32	60.8	63.37	74.66	88.34	90.85	86.54		43.73	61.54
T2_en_wiki$_{STZ}$					57.49	52.8	56.26	49.74	39.68	25.23	
T1_es_wiki$_{STZ}$			61.13	16.34		87.85	88.28	84.97			58.18
T2_es_wiki$_{STZ}$						52.86	56.19	51.69			
T1_fr_wiki$_{STZ}$			46.33	51.14		89.22	90.6	87.79			56.59
T2_fr_wiki$_{STZ}$						59.22	60.68	55.74			
T1_ko_wiki$_{STZ}$			53.61	40.9		76.57	81.65	73.81			1.72
T1_pt_wiki$_{STZ}$			42.22	11.78		83.57	84.74	79.69			36.88
T1_ru_wiki$_{STZ}$			51.89	13.54		77.91	76.31	73.86			34.01
Macro-avg	84.11	94.32	52.66	32.85	66.08	74.26	76.17	71.54	39.68	34.48	41.49

Table 6: BLEU scores on the Wikipedia datasets, with indicative average scores on the submitted outputs.

–NIST–	ADAPT		BME		Concordia	IMS			NILC
	20a	20b	20a	19	20a	20a	20b	19	20a
T1_ar_padt			8.29	8.29		12.63	12.62	12.22	
T1_en_ewt	13.47	13.81	12.52	12.62	12.7	13.78	13.74	13.61	
T2_en_ewt					11.61	12.13	12.14	11.79	9.96
T1_en_gum			12.1	11.99	11.62	12.96	12.98	12.69	
T2_en_gum					10.51	11.09	11.25	11.04	9
T1_en_lines			11.78	11.54	11.3	12.98	12.97	12.71	
T2_en_lines					9.93	10.82	10.89	10.63	9.09
T1_en_partut			10.22	10.34	9.83	11.07	11.05	11.01	
T2_en_partut					8.57	8.92	9.26	9.03	8.24
T1_es_ancora			13.57	13.52		14.9	14.89	14.69	
T2_es_ancora						12.66	12.66	12.38	
T1_es_gsd			11.6	11.44		12.86	12.86	12.77	
T2_es_gsd						11.05	11.13	10.82	
T1_fr_gsd			10.41	10.33		12.58	12.55	12.45	
T2_fr_gsd						11.14	11.06	10.79	
T1_fr_partut			9.05	8.99		10.56	10.58	10.36	
T2_fr_partut						8.38	8.88	8.27	
T1_fr_sequoia			10.56	10.55		12.65	12.64	12.53	
T2_fr_sequoia						11.02	11.21	11	
T1_hi_hdtb			11.98	12.26		13.32	13.34	13.07	
T1_id_gsd			12.14	11.82		12.89	12.91	12.83	
T1_ja_gsd			10.08	9.99		12.53	12.54	12.42	
T1_ko_gsd			12.12	11.98		12.44	12.46	12.27	
T1_ko_kaist			12.77	12.65		13.18	13.19	13	
T1_pt_bosque			9.89	9.77		12.4	12.41	12.15	
T1_pt_gsd			8.97	8.85		13.33	13.34	13.07	
T1_ru_gsd			11.9	11.91		12.4	12.42	12.15	
T1_ru_syntagrus			14.17	13.8		15.45	15.34	15.08	
T1_zh_gsd			11.62	11.85		12.86	12.91	12.78	
T1_en_pud			12.48	12.6	12.62	13.41	13.41	13.47	
T2_en_pud					11.43	11.36	11.59	11.45	9.68
T1_ja_pud			10.34	10.56		13.2	13.23	13.02	
T1_ru_pud			11.26	9.64		11.85	11.89	10.91	
T1_en_ewt$_{HIT}$			12.36	12.49	12.4	13.61	13.59	13.46	
T2_en_ewt$_{HIT}$					11.19	11.88	11.92	11.55	9.64
T1_en_pud$_{LAT}$			12.27	12.29	12.52	13.16	13.16	13.26	
T2_en_pud$_{LAT}$					11.18	11.02	11.19	11.08	9.59
T1_es_ancora$_{HIT}$			13.57	13.51		14.81	14.8	14.61	
T2_es_ancora$_{HIT}$						12.63	12.63	12.36	
T1_hi_hdtb$_{HIT}$			11.97	12.29		13.31	13.33	13.05	
T1_ko_kaist$_{HIT}$			12.77	12.63		13.22	13.23	13.02	
T1_pt_bosque$_{STA}$			9.89	9.73		12.4	12.43	12.14	
Macro-avg	13.47	13.81	11.47	11.39	11.24	12.4	12.44	12.21	9.31

Table 7: NIST scores on the 2019 datasets, with indicative average scores on the submitted outputs.

–NIST–	ADAPT		BME		Concordia	IMS			NILC	RALI	Tilburg
	20a	20b	20a	19	20a	20a	20b	19	20a	19	19
T1_en_wiki$_{STZ}$	13.77	14.3	12.76	12.94	12.9	14	14.14	13.91		10.99	12.72
T2_en_wiki$_{STZ}$					11.57	11.82	12.12	11.56	9.31	8.74	
T1_es_wiki$_{STZ}$			13.03	10.5		14.09	14.1	13.95			12.81
T2_es_wiki$_{STZ}$						11.72	12.04	11.22			
T1_fr_wiki$_{STZ}$			11.25	11.42		13.98	14.09	14			12.04
T2_fr_wiki$_{STZ}$						12.06	12.28	11.72			
T1_ko_wiki$_{STZ}$			11.12	10.98		11.5	11.55	11.37			3.24
T1_pt_wiki$_{STZ}$			10.64	9.26		13.57	13.61	13.25			9.61
T1_ru_wiki$_{STZ}$			12.04	10.87		12.86	12.86	12.65			10.04
Macro-avg	13.77	14.3	11.81	11	12.24	12.84	12.98	12.63	9.31	9.87	10.08

Table 8: NIST scores on the Wikipedia datasets, with indicative average scores on the submitted outputs.

–DIST–	ADAPT		BME		Concordia	IMS			NILC
	20a	20b	20a	19	20a	20a	20b	19	20a
T1_ar_padt			44.78	43.06		75.8	76.51	73.71	
T1_en_ewt	85.5	90.35	65.23	62.69	77.94	88.48	87.74	86.72	
T2_en_ewt					73.66	78.99	79.23	76.3	64.83
T1_en_gum			62.86	56.07	69.87	91.41	91.97	83.49	
T2_en_gum					67.02	73.22	76.47	73.07	60.42
T1_en_lines			61.44	52.77	68.62	85.89	86.48	82.21	
T2_en_lines					64.33	73.06	73.1	71.93	61.18
T1_en_partut			58.39	61.22	71.59	90.38	88.73	85.68	
T2_en_partut					62.39	69.75	72.98	67.45	59.74
T1_es_ancora			55.66	58.15		85.66	85.26	79.82	
T2_es_ancora						71.85	71.52	68.58	
T1_es_gsd			55.5	59.03		82.6	82.53	79.45	
T2_es_gsd						71.01	72.53	68.85	
T1_fr_gsd			55.48	59.35		84.64	83.24	84.15	
T2_fr_gsd						72.38	71.94	68.82	
T1_fr_partut			62.26	56.87		85.84	87.67	82.32	
T2_fr_partut						68.43	75.04	68.99	
T1_fr_sequoia			57.61	59.28		85.65	85.12	85.13	
T2_fr_sequoia						73.71	73.3	72.06	
T1_hi_hdtb			57.55	64.04		83.03	83.14	79.07	
T1_id_gsd			59.62	55.57		86.41	87.11	83.92	
T1_ja_gsd			60.57	57.03		87	87.83	87.17	
T1_ko_gsd			66.14	52.1		85.49	86.82	80.95	
T1_ko_kaist			62.88	50.9		84.52	84.9	78.69	
T1_pt_bosque			55.63	58.72		84.59	85.45	79.8	
T1_pt_gsd			53.49	54.93		87.86	87.7	79.33	
T1_ru_gsd			53.78	52.67		78.89	81.54	73.04	
T1_ru_syntagrus			56.72	55.6		83.02	81.07	78.66	
T1_zh_gsd			54.56	59.29		83.89	85.19	83.18	
T1_en_pud			61.85	59.84	76.46	85.75	84.52	87	
T2_en_pud					68.83	70.43	73.31	72.31	59.85
T1_ja_pud			55.77	56.72		85.67	86.29	84.04	
T1_ru_pud			56.17	32.08		82.4	82.67	77.12	
T1_en_ewt$_{HIT}$			62.82	60.36	74.47	86.65	86.24	85.35	
T2_en_ewt$_{HIT}$					69	77.23	76.98	74.99	62.2
T1_en_pud$_{LAT}$			59.93	56.13	77.68	82.19	82.23	86.18	
T2_en_pud$_{LAT}$					67.79	70.53	72.76	71.65	60.43
T1_es_ancora$_{HIT}$			56.14	58.38		86.96	86.36	81.14	
T2_es_ancora$_{HIT}$						73.27	73.06	70.02	
T1_hi_hdtb$_{HIT}$			57.43	64.58		83.52	84.05	78.88	
T1_ko_kaist$_{HIT}$			62.31	50.16		85.36	85.82	79.12	
T1_pt_bosque$_{STA}$			56.6	59.72		87.35	87.98	81.56	
Macro-avg	85.5	90.35	58.25	56.11	70.69	81.21	81.77	78.38	61.24

Table 9: DIST scores on the 2019 datasets, with indicative average scores on the submitted outputs.

–DIST–	ADAPT		BME		Concordia	IMS			NILC	RALI	Tilburg
	20a	20b	20a	19	20a	20a	20b	19	20a	19	19
T1_en_wiki$_{STZ}$	84.93	94.56	62.73	60.94	76.93	86.51	89.21	86.52		58.42	71.69
T2_en_wiki$_{STZ}$					67.26	72.53	74.94	71.46	57.56	50.68	
T1_es_wiki$_{STZ}$			57.66	35.26		87.35	87.68	81.39			63.27
T2_es_wiki$_{STZ}$						73.5	75.84	71.46			
T1_fr_wiki$_{STZ}$			58.82	66.53		91.64	92.58	86.44			70.12
T2_fr_wiki$_{STZ}$						73.68	76.64	72.53			
T1_ko_wiki$_{STZ}$			62.55	49.04		80.2	85.99	79.87			47.12
T1_pt_wiki$_{STZ}$			59.08	34.89		84.58	86.4	81.54			62.93
T1_ru_wiki$_{STZ}$			56.67	33.07		82.37	80.09	77.26			55.85
Macro-avg	84.93	94.56	59.59	46.62	72.1	81.37	83.26	78.72	57.56	54.55	61.83

Table 10: DIST scores on the Wikipedia datasets, with indicative average scores on the submitted outputs.

–BERT–	ADAPT		BME		Concordia	IMS			NILC
	20a	20b	20a	19	20a	20a	20b	19	20a
T1_ar_padt			0.9595	0.9618		0.9835	0.9836	0.9812	
T1_en_ewt	0.9815	0.9924	0.9473	0.958	0.9565	0.9849	0.9844	0.9819	
T2_en_ewt					0.9635	0.965	0.9643	0.9607	0.941 1
T1_en_gum			0.943	0.9503	0.9555	0.9897	0.9904	0.9829	
T2_en_gum					0.9572	0.9586	0.9624	0.9573	0.939 9
T1_en_lines			0.9341	0.9321	0.9565	0.9819	0.9826	0.9765	
T2_en_lines					0.9487	0.9519	0.9535	0.9491	0.938 1
T1_en_partut			0.934	0.9474	0.9565	0.9857	0.985	0.9829	
T2_en_partut					0.9505	0.9464	0.9528	0.9433	0.943 8
T1_es_ancora			0.9509	0.9612		0.9893	0.9891	0.9843	
T2_es_ancora						0.9644	0.9642	0.9601	
T1_es_gsd			0.9498	0.9584		0.9743	0.9738	0.9798	
T2_es_gsd						0.9532	0.9545	0.9561	
T1_fr_gsd			0.9414	0.9584		0.9853	0.9845	0.9835	
T2_fr_gsd						0.964	0.9622	0.9583	
T1_fr_partut			0.9452	0.9501		0.9886	0.9902	0.9854	
T2_fr_partut						0.954	0.9649	0.9525	
T1_fr_sequoia			0.9452	0.9562		0.9869	0.9872	0.9857	
T2_fr_sequoia						0.966	0.9666	0.9629	
T1_hi_hdtb			0.9808	0.9836		0.992	0.992	0.99	
T1_id_gsd			0.9589	0.9581		0.9707	0.9708	0.9843	
T1_ja_gsd			0.8833	0.9679		0.9917	0.9921	0.9914	
T1_ko_gsd			0.9814	0.9784		0.9933	0.9936	0.9903	
T1_ko_kaist			0.9811	0.976		0.9919	0.9921	0.9887	
T1_pt_bosque			0.9463	0.9546		0.9856	0.9857	0.98	
T1_pt_gsd			0.9297	0.9374		0.9845	0.9844	0.9785	
T1_ru_gsd			0.9702	0.9755		0.7624	0.7626	0.9828	
T1_ru_syntagrus			0.9718	0.9769		0.991	0.9899	0.9882	
T1_zh_gsd			0.887	0.9688		0.9894	0.9906	0.9884	
T1_en_pud			0.9385	0.9482	0.9605	0.9812	0.9812	0.983	
T2_en_pud					0.9613	0.9515	0.9557	0.9607	0.938 2
T1_ja_pud			0.8761	0.9666		0.9905	0.9909	0.9892	
T1_ru_pud			0.9686	0.9612		0.9876	0.9878	0.9833	
T1_en_ewt$_{HIT}$			0.9431	0.9535	0.9562	0.9824	0.982	0.9806	
T2_en_ewt$_{HIT}$					0.9593	0.9614	0.9611	0.9578	0.939 2
T1_en_pud$_{LAT}$			0.9333	0.9383	0.9596	0.9734	0.9738	0.9785	
T2_en_pud$_{LAT}$					0.9584	0.9475	0.9505	0.9474	0.936 8
T1_es_ancora$_{HIT}$			0.9506	0.9607		0.9886	0.988	0.9843	
T2_es_ancora$_{HIT}$						0.9641	0.9636	0.9601	
T1_hi_hdtb$_{HIT}$			0.9806	0.984		0.9922	0.9923	0.9898	
T1_ko_kaist$_{HIT}$			0.9768	0.9757		0.9924	0.9927	0.9892	
T1_pt_bosque$_{STA}$			0.9475	0.9544		0.9856	0.9868	0.9795	
Macro-avg	0.9815	0.9924	0.9468	0.9605	0.9572	0.972	0.9728	0.9755	0.9396

Table 11: BERT scores on the 2019 datasets, with indicative average scores on the submitted outputs.

–BERT–	ADAPT		BME		Concordia	IMS			NILC	RALI	Tilburg
	20a	20b	20a	19	20a	20a	20b	19	20a	19	19
T1_en_wiki$_{STZ}$	0.9826	0.9849	0.9413	0.9435	0.9614	0.9856	0.9882	0.9837		0.9396	0.938
T2_en_wiki$_{STZ}$					0.9473	0.9555	0.959	0.9515	0.9317	0.9171	
T1_es_wiki$_{STZ}$			0.952	0.9089		0.9893	0.9892	0.9853			0.9477
T2_es_wiki$_{STZ}$						0.9638	0.9671	0.9613			
T1_fr_wiki$_{STZ}$			0.9435	0.9508		0.9902	0.9917	0.9872			0.9512
T2_fr_wiki$_{STZ}$						0.9637	0.9667	0.9606			
T1_ko_wiki$_{STZ}$			0.9756	0.9683		0.9895	0.9919	0.9612			0.9354
T1_pt_wiki$_{STZ}$			0.948	0.9107		0.9801	0.9857	0.9801			0.9385
T1_ru_wiki$_{STZ}$			0.9707	0.9604		0.9883	0.987	0.986			0.9645
Macro-avg	0.9826	0.9849	0.9552	0.9404	0.9826	0.9784	0.9807	0.973	0.9317	0.9284	0.9459

Table 12: BERT scores on the Wikipedia datasets, with indicative average scores on the submitted outputs.

for, which we roughly approximate with macro-averages of all scores for a given metric as indicated in the bottom row of each metric results table. **BME-UW** and **IMS** improved their 2019 average scores on all datasets by 1.92 and 3.43 BLEU points respectively (+4 points in the open track for **IMS**), while **ADAPT** increased its score by 0.7 BLEU points, and almost 8 points in the open track. **Concordia** substantially improved their system compared to last year, now generating also from T2 inputs, and obtaining the highest BLEU scores on half of the English T2 datasets.

We now turn to comparisons between results for gold standard annotated datasets, and datasets with automatically generated inputs (silver standard data), for the same language. As in 2019, for the English and Spanish gold standards datasets (ewt and ancora), systems score equally high or higher than on the silver standard datasets (pred_{HIT}) for the same language; this year this is also true for Korean and Hindi, but not for Portuguese. On the new 2020 datasets, all of which are silver standard (see Section 4.2), the general tendency is that systems score higher than on the gold standard datasets (this is the case for 6 out of 9 test sets), even for systems that were not further developed compared to last year (**BME-UW**, **RALI**, **Tilburg**).[9] For 2 of the 9 Wikipedia datasets, namely Korean and Russian, all systems score lower than on any 2019 dataset in the same language, and on English-T2, only a few scores are higher than on the 2019 silver-standard datasets (but never higher than on the gold standard datasets). One explanation to the generally higher scores on the Wikipedia datasets might be that these contain cleaner sentences, or at least sentences are easier to parse and generate from, than those in the other data sets, which contain more varied and less standard sentences.

In terms of highest scores, bearing in mind scores aren't directly comparable across different test sets, we note that stand-out highest scores were achieved on the new 2020 Wikipedia English dataset in T1 by **ADAPT** which reaches over 94 BLEU / 14.3 NIST / 94 DIST, in subtrack 'b' (unrestricted resources). **IMS** matches this performance in terms of BERT score only. **IMS** also has the highest scores on all other new 2020 Wikipedia test sets in terms of all metrics (in some cases being the only team that submitted), except for the T2_en_wiki 'b' BLEU score, where **Concordia** reaches 57.49. As mentioned in the human evaluation section, **ADAPT**'s very high scores on T2_en_wiki 'b' subtrack are in part due to the fact that it used models trained on WikiText-103 (Merity et al., 2016).

6.2 Human Evaluation Results

Tables 13 and 14 show results from the human evaluations with Direct Assessment (DA) for English, Russian and Spanish (see Section 5.2 for details of the evaluation method). The datasets included were as shown in the results tables, and included all new SR'20 test sets. For each dataset, system outputs in the Shallow (T1) and Deep (T2) Tracks were evaluated in the same experiment.

Results from DA quality control were as follows. A total of 183,000 human assessments were collected on Mturk.[10] A lower rate of bad data was incurred with a higher proportion of Mturk workers, 48% passing quality control, compared to previous years, but still a large proportion, 52%, who did not meet this criterion, were omitted from computation of the official DA results above. High levels of low quality workers are consistent with what we have seen in DA used for crowd-sourced Machine Translation (Graham et al., 2016) and Video Captioning evaluations (Graham et al., 2017).

Results in Tables 13 and 14 are laid out as six separate tables one for each experiment run. The rank column indicates groups of systems where all systems have significantly higher scores than all systems in the next group below. Note that these groupings obscure some significant differences between systems within the same group. But because groups cannot be further subdivided in the sense above, systems within each group are of the same rank. Column 'Ave.' gives the average raw scores, 'Ave. z' the corresponding standard scores, *n* is the number distinct test sentences, and *N* the number of evaluators.

As can be seen from the tables, we included the 2019 submissions from two teams who submitted new systems in 2020: **BME-UW**'s and **IMS**'s English_ewt, Russian_syntagrus and Spanish_ancora outputs, in order to have comparable results. Results from the 2019 human evaluations are otherwise not comparable to the 2020 results, because a different set of systems was evaluated in each case. Absolute

[9] However, Tilburg and BME-UW show unexpected drops on some test sets that remain to be explained.

[10] www.mturk.com

English (ewt)

Rank	Ave.	Ave. z	n	N	System
1	92.6	0.540	1,698	1,931	ADAPT20BT1
	92.7	0.534	1,693	1,919	IMS20AT1
	92.3	0.520	1,683	1,912	IMS20BT1
	91.5	0.504	1,706	1,943	IMS19T1
	90.7	0.476	1,685	1,914	ADAPT20AT1
6	87.0	0.332	1,679	1,915	CONCORDIA20AT1
7	85.1	0.272	1,667	1,927	IMS20BT2
	84.7	0.259	1,701	1,942	IMS20AT2
	84.7	0.245	1,675	1,897	CONCORDIA20AT2
10	82.7	0.201	1,692	1,920	IMS19T2
11	79.3	0.086	1,679	1,925	BME20AT1
	77.4	0.024	1,690	1,933	BME19T1
13	75.6	−0.079	1,657	1,892	NILC20AT2

English (Wiki)

Rank	Ave.	Ave. z	n	N	System
1	94.7	0.638	699	1,043	ADAPT20BT1
2	93.7	0.535	719	1,072	IMS20BT1
	92.3	0.475	708	1,052	IMS20AT1
	91.6	0.444	715	1,061	IMS19T1
	91.6	0.441	718	1,115	ADAPT20AT1
6	88.7	0.275	707	1,038	CONCORDIA20AT1
7	87.3	0.157	700	1,016	IMS20BT2
8	85.6	0.057	755	1,078	IMS20AT2
	85.5	0.025	698	1,023	IMS19T2
	84.7	−0.029	715	1,036	CONCORDIA20AT2
	83.4	−0.033	698	1,033	TILBURG19T1
	81.8	−0.050	724	1,055	BME20AT1
	82.4	−0.074	721	1,056	BME19T1
	81.5	−0.118	689	1,021	RALI19T1
15	76.0	−0.463	720	1,044	RALI19T2
	76.6	−0.491	721	1,088	NILC20AT2

Russian (syntagrus)

Rank	Ave.	Ave. z	n	N	System
1	90.3	0.375	4,403	5,269	IMS20AT1
2	89.4	0.319	4,377	5,260	IMS20BT1
3	88.9	0.285	4,429	5,332	IMS19T1
4	81.2	−0.166	4,461	5,392	BME20AT1
	81.3	−0.177	4,466	5,371	BME19T1

Russian (wiki)

Rank	Ave.	Ave. z	n	N	System
1	89.1	0.490	722	963	IMS20AT1
2	88.1	0.396	724	947	IMS19T1
	87.2	0.382	728	972	IMS20BT1
4	78.2	−0.079	738	963	BME20AT1
5	73.7	−0.256	703	929	TILBURG19T1
6	68.2	−0.493	722	968	BME19T1

Spanish (ancora)

Rank	Ave.	Ave. z	n	N	System
1	89.8	0.518	1,233	1,491	IMS20AT1
	89.5	0.498	1,251	1,479	IMS20BT1
3	85.9	0.383	1,252	1,504	IMS19T1
4	77.9	0.037	1,184	1,421	IMS20AT2
	77.1	0.017	1,225	1,467	IMS20BT2
	76.8	−0.029	1,199	1,436	IMS19T2
7	70.6	−0.271	1,223	1,490	BME19T1
	70.2	−0.276	1,230	1,489	BME20AT1

Spanish (Wiki)

Rank	Ave.	Ave. z	n	N	System
1	87.6	0.557	879	1,167	IMS20BT1
	86.3	0.524	909	1,192	IMS20AT1
3	85.2	0.465	897	1,162	IMS19T1
4	76.9	0.123	847	1,088	IMS20BT2
	75.9	0.092	910	1,197	IMS20AT2
	75.4	0.063	880	1,124	IMS19T2
7	71.0	−0.119	894	1,173	TILBURG19T1
	69.8	−0.170	877	1,139	BME20AT1
9	55.5	−0.726	908	1,181	BME19T1

Table 13: Human evaluation results for **Meaning Similarity**. Ave. = average score received by systems; Ave. z = corresponding average standardized score; systems ranked according to Ave. z score; horizontal lines indicate groups, such that systems in a group all significantly outperform all systems in lower ranked groups; n = total number of distinct test sentences assessed; N = total number of human judgments.

scores (*Ave.*) are particularly affected by differences even from different sets of evaluators and output samples. Pairwise rankings for the same systems however, can be expected to be more robust.

The 2020 evaluations are indeed consistent with last year's in terms rankings according to z-scores: when a system was in a higher cluster than another in 2019, it is still the case in 2020. There are two exceptions, **IMS19T2** and **BME19T1** were together in the same cluster on Meaning Similarity and Readability for English_ewt, but this year appear in consecutive clusters. One explanation for this is that we collected considerably more judgements this year, and when there are more judgements, it is likely that more statistically significant differences are found, which seems to be the case here. Most of the standard scores are lower in 2020 than they were in 2019, and this is likely due to the fact that the best scoring systems score higher (e.g. 4 systems at the level or above the best 2019 system on English_ewt) which tends to push down the lower systems in terms of z score.

Looking at the 2020 results only, for both Meaning Similarity (Table 13) and Readability (Table 14), SR'20 systems are generally ranked higher than SR'19 systems, T1 systems are generally ranked higher than T2 systems, and where a system has both 'a' (restricted subtrack) and 'b' (open subtrack) variants, the 'b' system is ranked higher in most (though not all) cases. This is all as expected.

In terms of Meaning Similarity, the best score is higher than last year on all three comparable datasets: the best system obtains an average z score of 0.54 on English_ewt (0.507 in 2019), 0.375 on Russian_syntagrus (0.238), and 0.518 on Spanish_ancora (0.378). For Readability, the best z scores are 0.426, 0.546 and 0.446 respectively, compared to 0.507, 0.238 and 0.519 in 2019, that is, only for

English (ewt)

Rank	Ave.	Ave. z	n	N	System
1	75.7	0.426	797	913	ADAPT20BT1
–	75.7	0.417	669	1,402	HUMAN
	73.9	0.374	807	917	IMS20AT1
	73.9	0.370	810	927	IMS20BT1
	73.4	0.346	811	926	IMS19T1
	71.8	0.321	806	908	CONCORDIA20AT2
	72.5	0.320	830	953	ADAPT20AT1
	70.2	0.270	860	969	CONCORDIA20AT1
	68.6	0.185	823	947	NILC20AT2
	67.3	0.159	807	936	IMS20BT2
	65.8	0.109	753	866	IMS20AT2
11	63.6	0.027	808	923	IMS19T2
12	58.2	−0.152	839	946	BME20AT1
	56.7	−0.208	822	935	BME19T1

English (Wiki)

Rank	Ave.	Ave. z	n	N	System
–	87.4	0.592	955	2,605	HUMAN
1	86.3	0.551	910	1,274	ADAPT20BT1
2	83.3	0.444	938	1,302	IMS20BT1
	79.6	0.401	942	1,290	CONCORDIA20AT1
	82.1	0.383	949	1,314	IMS20AT1
	81.5	0.373	948	1,346	ADAPT20AT1
	80.6	0.370	952	1,283	CONCORDIA20AT2
	81.3	0.361	937	1,321	IMS19T1
8	75.4	0.213	930	1,273	NILC20AT2
9	70.2	0.055	932	1,256	IMS20BT2
	69.0	−0.030	963	1,284	IMS20AT2
11	67.3	−0.095	932	1,233	IMS19T2
	67.8	−0.128	932	1,306	TILBURG19T1
13	64.4	−0.181	897	1,239	BME19T1
14	60.8	−0.299	933	1,297	BME20AT1
	62.3	−0.303	912	1,242	RALI19T1
16	56.1	−0.562	940	1,329	RALI19T2

Russian (syntagrus)

Rank	Ave.	Ave. z	n	N	System
–	93.5	0.635	499	689	HUMAN
1	90.8	0.546	943	1,068	IMS20AT1
2	89.1	0.489	972	1,106	IMS20BT1
3	87.3	0.424	949	1,085	IMS19T1
4	69.7	−0.166	966	1,106	BME20AT1
	67.3	−0.230	943	1,078	BME19T1

Russian (Wiki)

Rank	Ave.	Ave. z	n	N	System
–	92.6	0.924	414	658	HUMAN
1	85.5	0.720	695	897	IMS20AT1
2	83.1	0.643	677	868	IMS19T1
	83.1	0.635	705	916	IMS20BT1
4	63.2	0.050	703	900	BME20AT1
5	47.9	−0.415	710	911	TILBURG19T1
6	37.7	−0.781	679	883	BME19T1

Spanish (ancora)

Rank	Ave.	Ave. z	n	N	System
–	90.7	0.595	676	1,110	HUMAN
1	87.6	0.446	944	1,164	IMS20AT1
	86.7	0.433	929	1,144	IMS20BT1
	85.8	0.391	920	1,112	IMS19T1
4	80.2	0.173	923	1,127	IMS20BT2
	79.5	0.158	933	1,149	IMS20AT2
6	77.0	0.066	923	1,157	IMS19T2
7	67.1	−0.378	923	1,155	BME19T1
	66.4	−0.401	906	1,125	BME20AT1

Spanish (Wiki)

Rank	Ave.	Ave. z	n	N	System
–	87.7	0.371	524	798	HUMAN
1	86.5	0.289	643	733	IMS20AT1
	85.6	0.249	623	721	IMS20BT1
	84.9	0.237	620	715	IMS19T1
4	81.5	0.142	620	711	IMS20BT2
	79.3	0.046	619	713	IMS20AT2
	79.0	0.044	637	725	IMS19T2
	77.2	0.015	574	663	BME20AT1
	74.8	−0.053	614	699	TILBURG19T1
9	62.2	−0.628	592	679	BME19T1

Table 14: Human evaluation results for **Readability**. Ave. = average score for system; Ave. z = corresponding average standardized score; systems ranked by Ave. z score; horizontal lines indicate groups, such that systems in a group all significantly outperform all systems in lower ranked groups; n = distinct test sentences assessed; N = total number of judgments; HUMAN = original reference texts.

the Russian dataset the absolute z score is higher. However, both in English and Spanish, the human references are also scored lower,

On the English_ewt dataset, for both criteria , **ADAPT** and **IMS** get all their T1 submissions in the first rank (including the 2019 one for **IMS**); for Readability, **Concordia**'s T1 and T2 submissions, **NILC**'s T2 and **IMS**'s T2 also make it to the first cluster, which contains 10 systems. It is the first time that some T2 submissions make it to the same cluster as the human-written references. In terms of Meaning Similarity, three of the four T2 top-ranking submissions are found at the seventh rank, in the third cluster (**IMS**'s and **Concordia**'s 2020 submissions); **NILC** ranks last in for this criterion, indicating that although the generated texts read well, they do not contain the expected information.

IMS's restricted track T1 submissions ranks in the first cluster for all four non-English datasets according to both criteria. Note that this is consistent with the results of the automatic evaluations, in which even though **IMS**'s open track systems gets overall better results, it is not the case on every individual dataset. On the Spanish Wikipedia dataset, **IMS** gets their three T1 submissions in the same cluster as the human-written texts (Readability), while also scoring high in terms of Meaning Similarity. Here again it is the first time that a system gets at a level where it is statistically not significantly ranked lower than a human reference in a non-English language.

On the English_wiki dataset, **ADAPT** ranks first for both criteria, and is the only one in the same clus-

	BLEU	NIST	DIST	BERT	Read.
English (ewt)	0.94**	0.85**	0.97**	0.94**	0.78**
English (Wiki)	0.95**	0.92**	0.97**	0.97**	0.74**
Russian (syntagrus)	1.00**	0.98**	1.00**	0.97**	1.00**
Russian (Wiki)	0.99**	0.88**	0.98**	0.99**	0.99**
Spanish (ancora)	0.87**	0.72*	0.98**	0.97**	0.95**
Spanish (Wiki)	0.94**	0.80**	0.99**	0.99**	0.95**

Table 15: Pearson correlation of BLEU, NIST, DIST, BERT and Readability scores with human assessment of Meaning Similarity Ave z. ** = significant at p< 0.01; * = significant at p< 0.05.

	BLEU	NIST	DIST	BERT	Mean. Sim.
English (ewt)	0.69**	0.39	0.82**	0.70**	0.78**
English (Wiki)	0.79**	0.62**	0.81**	0.81**	0.74**
Russian (syntagrus)	1.00**	0.99**	1.00**	0.96**	1.00**
Russian (Wiki)	1.00**	0.91**	0.97**	0.99**	0.99**
Spanish (ancora)	0.67*	0.47	0.98**	0.87**	0.95**
Spanish (Wiki)	0.92**	0.77**	0.96**	0.97**	0.95**

Table 16: Pearson correlation of BLEU, NIST, DIST, BERT and Meaning Similarity scores with human assessment of Readability Ave z. ** = significant at p< 0.01; * = significant at p< 0.05.

ter as the human references for Readability. The very high scores of **ADAPT** reflect the also very strong automatic evaluation (94 BLEU, see previous subsection). On this dataset, the gap between **ADAPT** and the other systems was a little bit surprising, and we realised that it used models trained on WikiText-103 (Merity et al., 2016), which seems to contain a small proportion of the sentences of the test set. This was of course not something done on purpose by the team, since the participants did not know that the new test sets would contain exclusively Wikipedia material. Note that **IMS**'s 2019 system remains extremely competitive on this dataset, outperforming all other submissions except **ADAPT**. The other 2019 systems (**Tilburg**, **RALI**, **BME-UW**) occupy the lower ranks as expected, with **NILC**'s T2 submission.

6.3 Comparisons between human and automatic evaluations

In this section, we compare the metrics results presented in Section 6.1 with human evaluation results presented in Section 6.2 first informally, then more systematically, in two ways. First in terms of Pearson correlation of metrics with Meaning Similarity z scores (Table 15) and Readability z scores (Table 16), both at the test set level; second by plotting all system scores for the T2_en_ewt test set in groups by system, once in decreasing order of Meaning Similarity scores (Figure 4), and once in decreasing order of Readability scores (Figure 5).

In terms of human and automatic metrics for individual systems, the **IMS** 2020 results are very high both in terms of automatic evaluations and human assessments for many of the languages offered, and represent a substantial improvement over their 2019 system, both through method improvement (higher results despite using the same data resources for training), and through using better data (open track results). **ADAPT** got very high automatic and human scores in the open subtrack for the English T1 test sets compared to other English T1 submissions, to the point of drawing level with the human en_ewt texts for Readability (Table 14, top left). **Concordia** obtained the highest Readability scores on the English T2 submissions, while matching the **IMS** Meaning Similarity scores; however, the corresponding metric scores were lower. Overall, **IMS**, **ADAPT**, **BME-TUW** and **Concordia** all achieved major improvements in their systems. The **NILC** results appear to show the limitations of using an off-the-shelf generic model.

There are indications that we do capture different aspects with the two quality criteria. For instance, **NILC** gets good Readability scores but low Meaning Similarity, which is typical of generative models like GPT-2, which are able to generate very good texts but without ensuring that they correspond to the input fed to the system. **Concordia**'s T2 submissions seems to have similar features, even though it scores consistently higher than **NILC**'s submission.

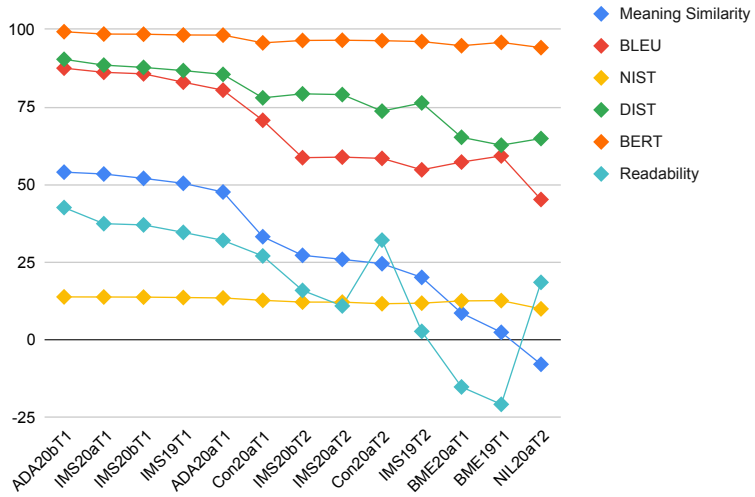

Figure 4: Visualisation of the scores of the 13 submissions on the English_ewt dataset according to the 6 metrics. On the X axis, the systems as ranked by Meaning Similarity Ave. z in Table 13.[12]

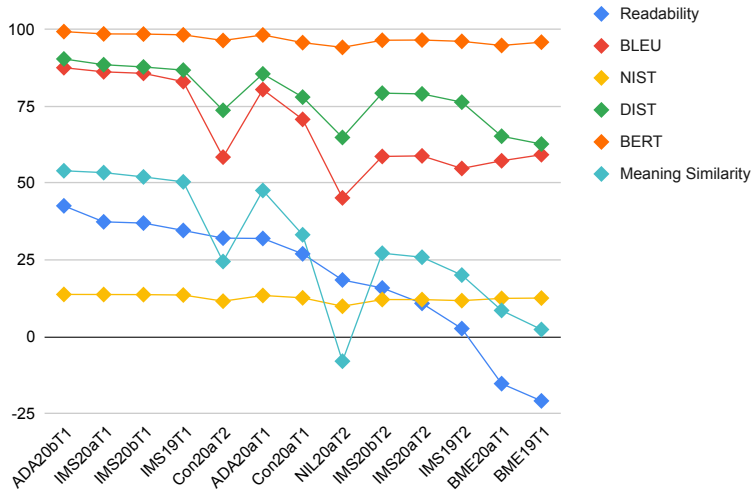

Figure 5: Visualisation of the scores of the 13 submissions on the English_ewt dataset according to the 6 metrics. On the X axis, the systems as ranked by Readability Ave. z in Table 14.[12]

Regarding correlations of metrics with Meaning Similarity, Table 15 shows these to be generally in the mid to high nineties, with isolated lower scores, e.g. for BLEU on Spanish_ancora. One notable exception is the lower correlation of Meaning Similarity with Readability on the English test sets. This may be due to English outputs being of higher quality generally than those for other languages, and evaluators finding it easier to judge Meaning Similarity separately when sentences are more readable.

Regarding correlations of metrics with Readability (Table 16), the quality criterion generally supposed to correlate better with metrics, there are more cases of lower correlation including all scores for both English test sets, as well as BLEU, NIST and BERT for Spanish_ancora and NIST for Spanish_wiki. For the Russian test sets, all correlations are exceptionally high.

For Readability, the reported correlations are noticeably lower than the ones reported in both 2018 and 2019. One possible explanation is that in the previous years, the differences between the systems were very clear, but now we have many more good systems, so it is more difficult for the automatic metrics

[12]Lines have been added to the plot to show up clearly which scores are for the same evaluation method, even in black and white version.

to capture the differences between them. This year, in contrast to SR'19, the DIST metric has the most consistently high correlation scores with the human evaluation methods.

Figure 4 plots results for all systems on English_ewt, with score points for each evaluation method connected by a line to indicate that they are for the same method, systems ordered across the x dimension in order of decreasing Meaning Similarity. From the plot we can see clearly that, as Meaning Similarity goes down, generally so do the metrics, and also Readability except for scores for **Concordia20aT2** and **NILC20aT2** which buck the trend.

Figure 5 shows a similar plot, this time ordered by decreasing Readability. This shows very clearly that all metrics and Meaning Similarity scores dip at the same two points, **Concordia20aT2** and **NILC20aT2**, although this is less evident for BERT and NIST scores (in the latter case because NIST has a very different range from the other evaluation methods and because it is not bounded at the top end cannot simply be mapped to a 0–1 range). So here too, it is clear that for those two systems, Readability is lower than would be expected on the basis of all other evaluation methods including human assessment of Meaning Similarity.

7 Concluding Remarks

The 2020 edition of the SR Shared Task (SR'20) saw 5 teams submitting new systems and 4 teams submitting outputs for the new test sets using their 2019 systems. Datasets, evaluation scripts, system outputs and more information about the shared task can be found on the GenChal repository.[13]

Among the notable trends we can observe in evaluations this year are the following: (i) the best Shallow Track English systems appear to have closed the gap to human-written texts in terms of all evaluation measures; (ii) for the first time we have seen outputs for a non-English language (Spanish) approach the quality of human-written reference texts; and (iii) allowing additional resources to be used in system building can make a very big difference to performance. Further progress has also been made in SR'20 for deep track systems: the best Deep Track system performed equally well or better than most Shallow Track systems for both Readability and Meaning similarity.

Overall, the SR'20 results provide further evidence that generation from structured meaning representations can be done with impressive success by current neural methods. Our aim for next year's edition of the shared task is to add linked tasks corresponding to an earlier stage in the generation process, asking for submissions which use the intermediate UD representations as well as submissions that by-pass them in order to compare which gives better results overall.

Acknowledgements

SR'20 is endorsed by SIGGEN. The work on the organisation, realisation, and evaluation of this shared task was supported by (1) the European Commission in the context of its H2020 Program under the grant numbers 870930-RIA, 779962-RIA, 825079-RIA, 786731-RIA at Universitat Pompeu Fabra; (2) the Applied Data Analytics Research & Enterprise Group, University of Brighton, UK, (3) Science Foundation Ireland (sfi.ie) under the SFI Research Centres Programme co-funded under the European Regional Development Fund, grant number 13/RC/2106 (ADAPT Centre for Digital Content Technology, www.adaptcentre.ie) at Trinity College Dublin; and (4) the Coordination for the Improvement of Higher Education Personnel in Brazil (CAPES) under the grant 88887.508597/2020-00.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August. Association for Computational Linguistics.

G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, A. F. Smeaton, Y. Graham, and W. Kraaij. 2019. Trecvid 2019: An evaluation campaign to benchmark video activity

[13] https://sites.google.com/site/genchalrepository/surface-realisation/sr-20-multilingual

detection, video captioning and matching, and video search & retrieval. In *Proceedings of TRECVID*, volume 2019.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors. 2020. *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics, Online, November.

Anya 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 (ACL-HLT'11)*.

Anya Belz, Michael 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*, ENLG '11, pages 217–226, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wanxiang Che, Jiang Guo, Yuxuan Wang, Bo Zheng, Huaipeng Zhao, Yang Liu, Dechuan Teng, and Ting Liu. 2017. The hit-scir system for end-to-end parsing of universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 52–62.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ondřej Dušek, David M Howcroft, and Verena Rieser. 2019. Semantic noise matters for neural natural language generation. *arXiv preprint arXiv:1911.03905*.

Thiago Castro Ferreira and Emiel Krahmer. 2019. Surface Realization Shared Task 2019 (SR'19): The Tilburg University Approach. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation*, Hong Kong, China.

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.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, FirstView:1–28, 1.

Yvette Graham, George Awad, and Alan Smeaton. 2017. Evaluation of Automatic Video Captioning Using Direct Assessment. *ArXiv e-prints*, October.

Yvette Graham, George Awad, and Alan Smeaton. 2018. Evaluation of automatic video captioning using direct assessment. *PLOS ONE*, 13(9):1–20, 09.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. *CoRR*, abs/1906.09833.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Xinyu Hua and Lu Wang. 2019. Sentence-level content planning and style specification for neural text generation.

Adám Kovács, Evelin Ács, Judit Ács, András Kornai, and Gábor Recski. 2019. BME-UW at SR'19: Surface Realization with Interpreted Regular Tree Grammars. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation*, Hong Kong, China.

Guy Lapalme. 2019. Realizing Universal Dependencies Structures Using a Symbolic Approach. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation*, Hong Kong, China.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

KyungTae Lim, Cheoneum Park, Changki Lee, and Thierry Poibeau. 2018. Sex bist: A multi-source trainable parser with deep contextualized lexical representations. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 143–152.

Qingsong Ma, Yvette Graham, Timothy Baldwin, and Qun Liu. 2017. Further investigation into reference bias in monolingual evaluation of machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2466–2475, Copenhagen, Denmark, September. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Simon Mille, Bernd Bohnet, Leo Wanner, and Anja Belz. 2017. Shared task proposal: Multilingual surface realization using universal dependency trees. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 120–123.

Simon Mille, Anya 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 1st Workshop on Multilingual Surface Realisation (MSR), 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–12, Melbourne, Australia.

Simon Mille, Anya 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)*, pages 1–17.

Paul Molins and Guy Lapalme. 2015. Jsrealb: A bilingual text realizer for web programming. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 109–111.

K. Papineni, S. Roukos, T. Ward, and W. j. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, July 7–12,.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D Manning. 2019. Universal dependency parsing from scratch. *arXiv preprint arXiv:1901.10457*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal dependency parsing with cross-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Asso. for Comp. Linguistics.

Xiang Yu, Agnieszka Falenska, Marina Haid, Ngoc Thang Vu, and Jonas Kuhn. 2019. IMSurReal: IMS at the Surface Realization Shared Task 2019. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation*, Hong Kong, China.

Xiang Yu, Simon Tannert, Ngoc Thang Vu, and Jonas Kuhn. 2020. Fast and accurate non-projective dependency tree linearization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1451–1462, Online, July. Association for Computational Linguistics.

Fangzhou Zhai, Vera Demberg, Pavel Shkadzko, Wei Shi, and Asad Sayeed. 2019. A hybrid model for globally coherent story generation. In *Proceedings of the Second Workshop on Storytelling*, pages 34–45.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Liang Zhao, Jingjing Xu, Junyang Lin, Yichang Zhang, Hongxia Yang, and Xu Sun. 2020. Graph-based multi-hop reasoning for long text generation. *arXiv preprint arXiv:2009.13282*.