The University of Helsinki submissions to the WMT19 news task

Aarne Talman, Umut Sulubacak, Raúl Vázquez, Yves Scherrer, Sami Virpioja,
Alessandro Raganato, Arvi Hurskainen and Jörg Tiedemann

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
HUMANISTINEN TIEDEKUNTA
HUMANISTISKA FAKULTETEN
FACULTY OF ARTS

ACADEMY OF FINLAND

Main submissions

Language pairs:
- English<->Finnish
- English ->German

Transformer architecture (Vaswani et al., 2017):
- big version, 6 layers of hidden size 4096, 16 attention heads, and a dropout of 0.1
- OpenNMT-py framework (Klein et al., 2017)
- MarianNMT (Junczys-Dowmunt et al., 2018)

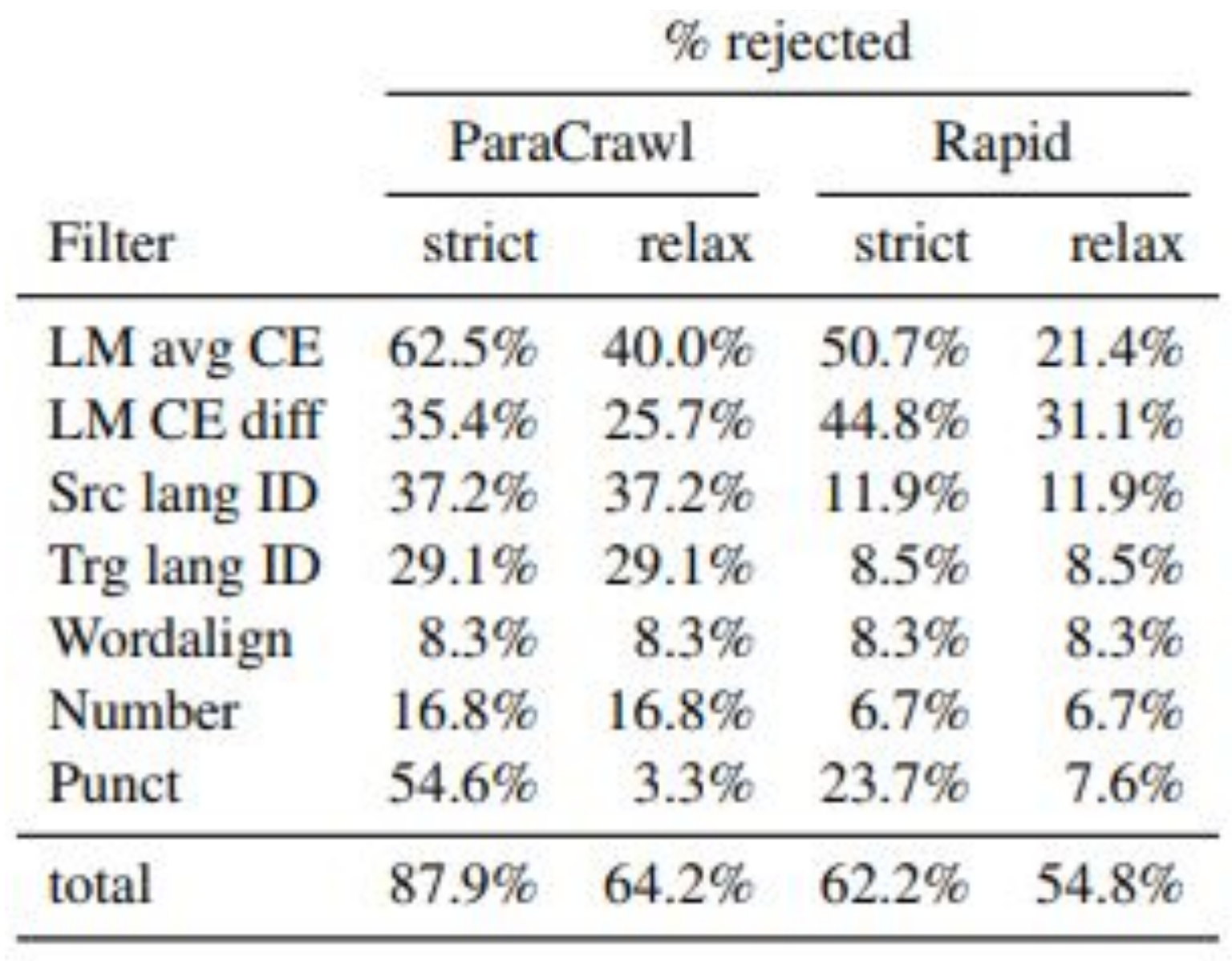
Subword units (Sennrich et al., 2016):
- joint BPE vocabulary of 37 000 units for each language pair for English-Finnish
- vocabulary of 35 000 units for English-German

Back-translation (Sennrich et al., 2016):
- We translated the monolingual English news data from the years 2007–2018, from which we used a filtered and sampled subset of 7M sentences for our Finnish–English systems, and the Finnish data from years 2014–2018 using our WMT 2018 submissions
- We also used the back-translations we generated for the WMT 2017 news translation task with an SMT model to create 5.5M sentences of from the Finnish news2014 and news2016 corpora (Östling et al., 2017).
- For English-German we created backtranslations with a standard Transformer model resulting in 10.3M sentence pairs

- For English–>Finnish, our experiments also include a rule-based system (Raganato et al., 2018).

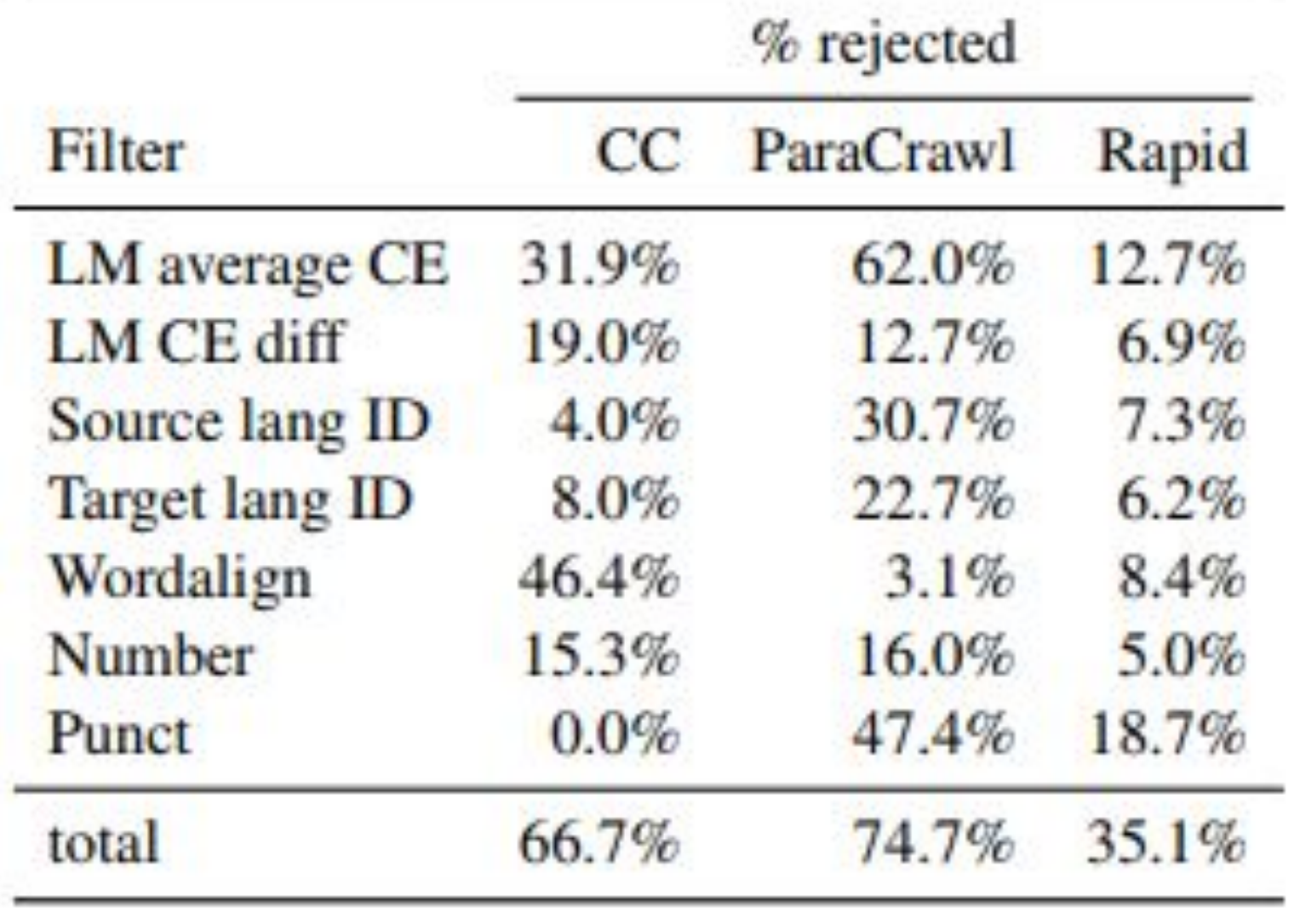
Data Filtering

We used multiple filtering methods:
- ○ Heuristic filters
- ○ Language Identifiers
- ○ Language model filters
- ○ Word-alignment filter



	% rejected			
	ParaCrawl		Rapid	
Filter	strict	relax	strict	relax
LM avg CE	62.5%	40.0%	50.7%	21.4%
LM CE diff	35.4%	25.7%	44.8%	31.1%
Src lang ID	37.2%	37.2%	11.9%	11.9%
Trg lang ID	29.1%	29.1%	8.5%	8.5%
Wordalign	8.3%	8.3%	8.3%	8.3%
Number	16.8%	16.8%	6.7%	6.7%
Punct	54.6%	3.3%	23.7%	7.6%
total	87.9%	64.2%	62.2%	54.8%

Table 3: Percentage of lines rejected by each filter for English–Finnish data sets. The strict version is the same as for English–German, and the relax version applies relaxed thresholds.



	% rejected		
Filter	CC	ParaCrawl	Rapid
LM average CE	31.9%	62.0%	12.7%
LM CE diff	19.0%	12.7%	6.9%
Source lang ID	4.0%	30.7%	7.3%
Target lang ID	8.0%	22.7%	6.2%
Wordalign	46.4%	3.1%	8.4%
Number	15.3%	16.0%	5.0%
Punct	0.0%	47.4%	18.7%
total	66.7%	74.7%	35.1%

Table 2: Percentage of lines rejected by each filter for English–German data sets. Each line can be rejected by several filters. The total of rejected lines is the last row of the table.

English <-> Finnish

English -> Finnish

Training:
- Filtered versions of Europarl, ParaCrawl, Rapid, Wikititles, newsdev2015 and newstest2015 as well as backtranslations (8.5M sentence pairs)

Validation:
- Newstest 2016

Finnish -> English

Training:
- Filtered versions of Europarl, ParaCrawl, Rapid, Wikititles, newsdev2015 and newstest2015 as well as backtranslations (12.3M–26.7M sentence pairs, different samplings of back-translations)

Validation:
- Newstest 2016

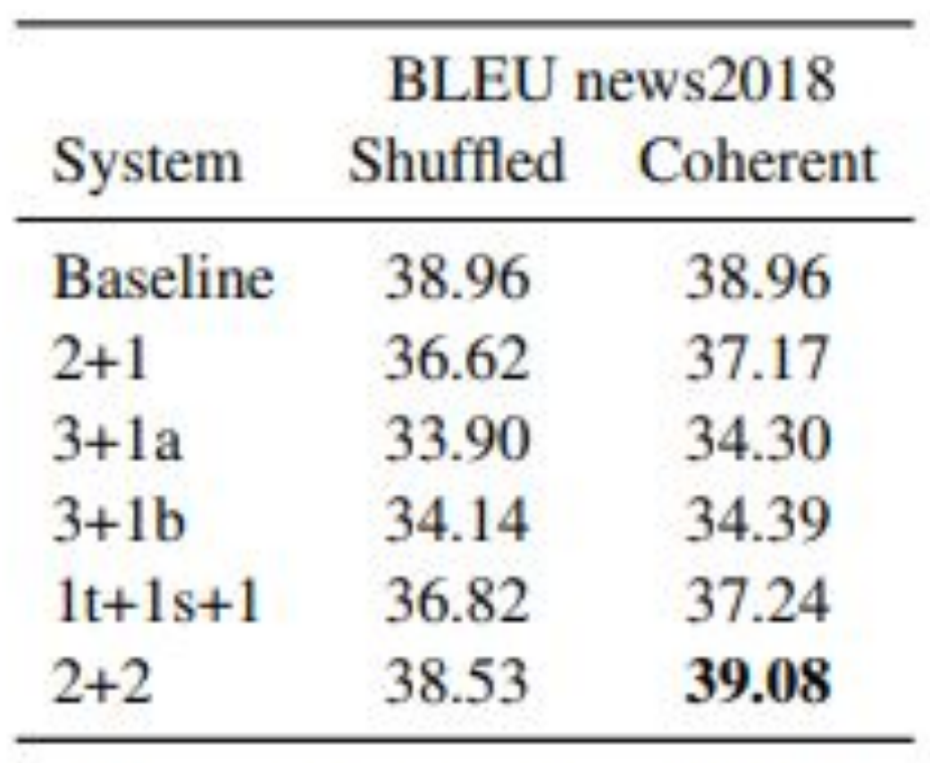
English -> German

Training:
- Filtered versions of Europarl, NewsCommentary, Rapid, CommonCrawl, ParaCrawl, Wikititles, and backtranslations
- 15.7M sentence pairs

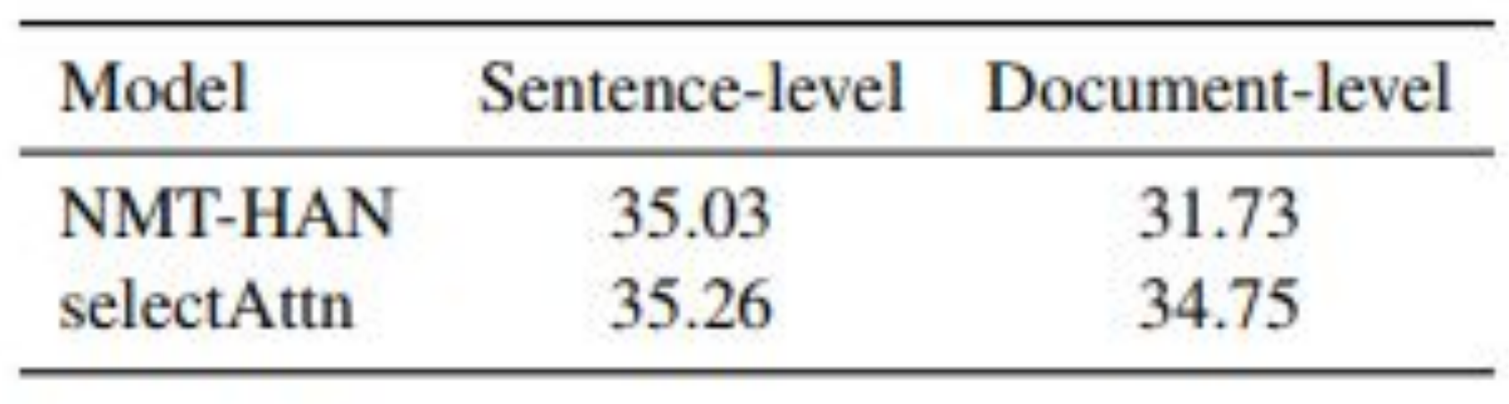
Validation:
- Newstest 2011-2016

Document-level systems (English -> German)

We did experiments with two types of document-level models:
- Concatenation models (Tiedemann and Scherrer, 2017)
- Hierarchical attention models: NMT-HAN (Miculicich et al., 2018) and selectAttn (Maruf et al., 2019).



	BLEU news2018	
System	Shuffled	Coherent
Baseline	38.96	38.96
2+1	36.62	37.17
3+1a	33.90	34.30
3+1b	34.14	34.39
1t+1s+1	36.82	37.24
2+2	38.53	**39.08**



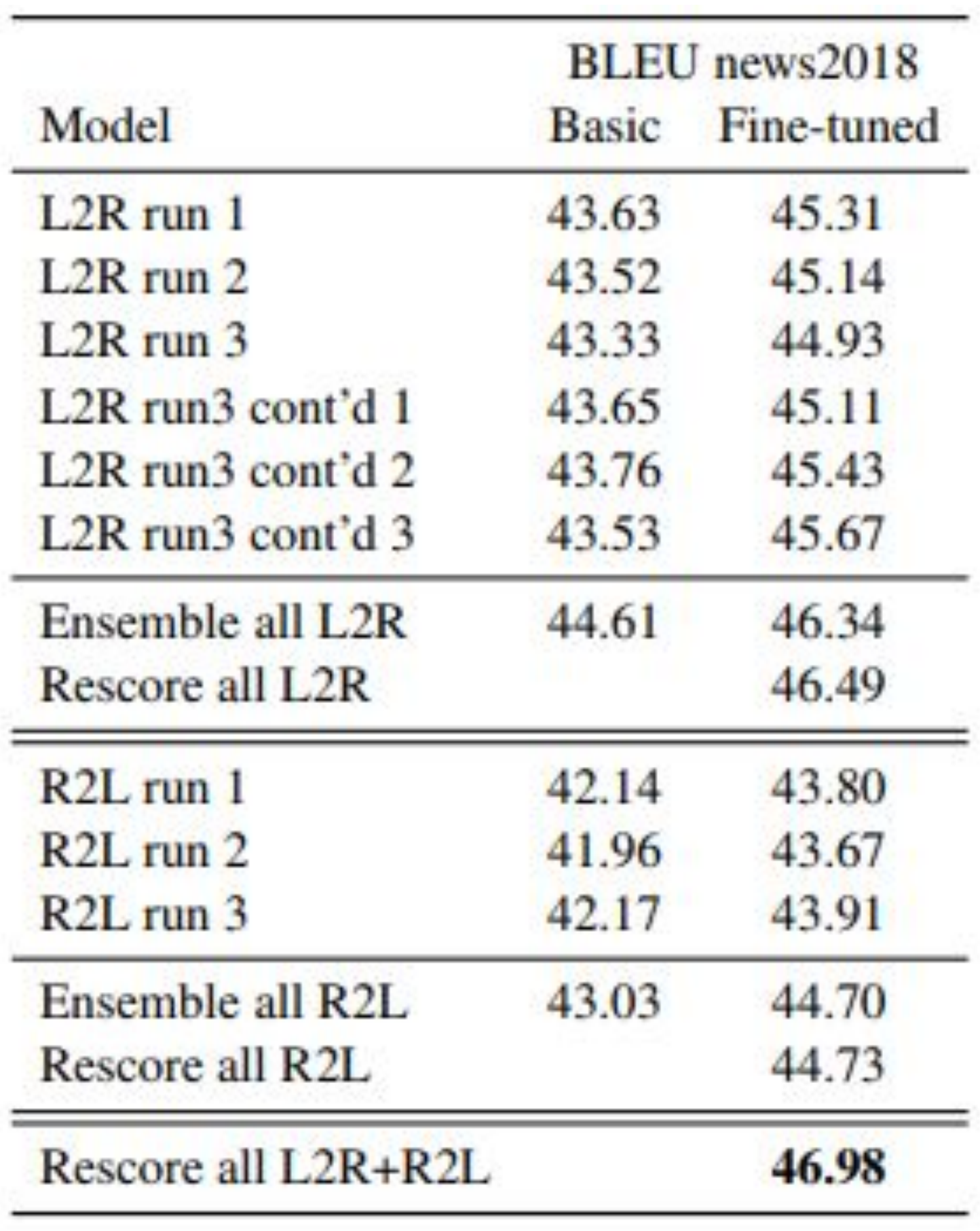
Model	Sentence-level	Document-level
NMT-HAN	35.03	31.73
selectAttn	35.26	34.75

Table 11: Results (case-sensitive BLEU) of the hierarchical attention models on the coherent newstest 2018 dataset.

Rule-based system (English - Finnish)

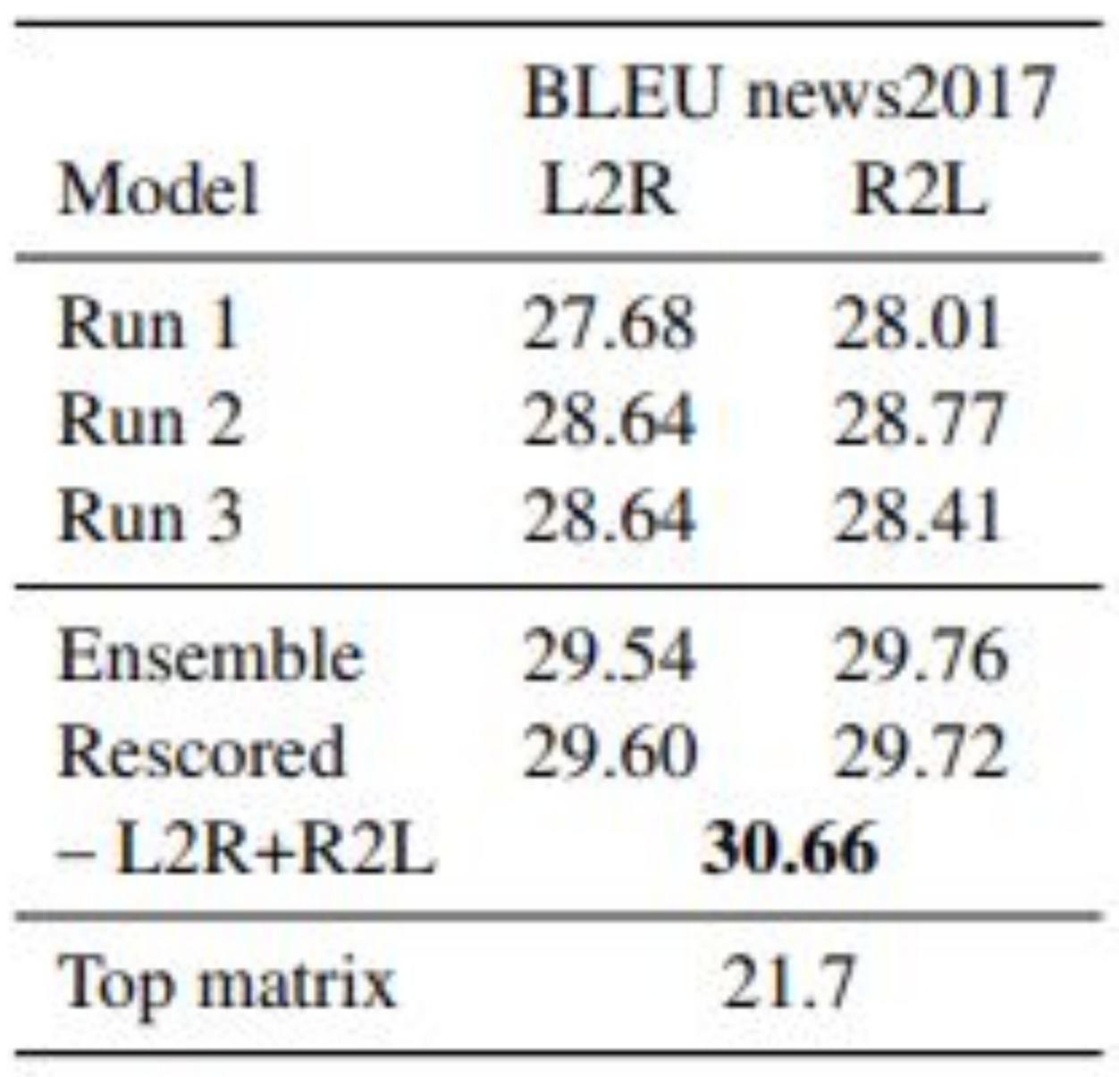
Rule-based MT:
- Updated version of the Hurskainen and Tiedemann (2017) and Raganato et al. (2018) system, improving mainly:
  - Translation of English noun compounds
  - Translation of questions
  - Translation of Temporal subordinate clauses
  - Rule optimization (30% of rules were removed)

Experiments



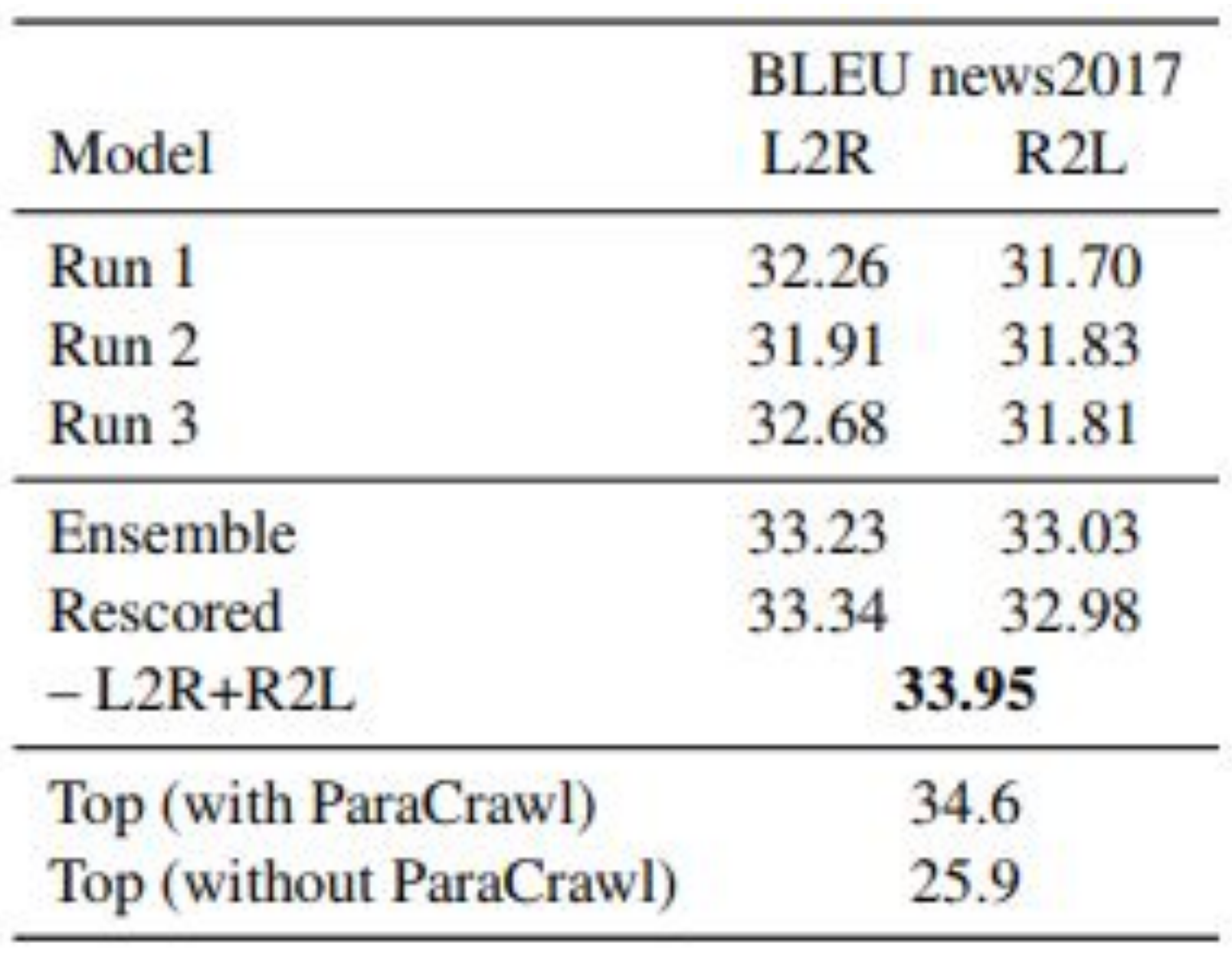
	BLEU news2018	
Model	Basic	Fine-tuned
L2R run 1	43.63	45.31
L2R run 2	43.52	45.14
L2R run 3	43.33	44.93
L2R run3 cont'd 1	43.65	45.11
L2R run3 cont'd 2	43.76	45.43
L2R run3 cont'd 3	43.53	45.67
Ensemble all L2R	44.61	46.34
Rescore all L2R		46.49
R2L run 1	42.14	43.80
R2L run 2	41.96	43.67
R2L run 3	42.17	43.91
Ensemble all R2L	43.03	44.70
Rescore all R2L		44.73
Rescore all L2R+R2L		**46.98**

Table 5: English–German results from individual MarianNMT transformer models and their combinations (cased BLEU).



	BLEU news2017	
Model	L2R	R2L
Run 1	27.68	28.01
Run 2	28.64	28.77
Run 3	28.64	28.41
Ensemble	29.54	29.76
Rescored	29.60	29.72
– L2R+R2L	**30.66**	
Top matrix	21.7	

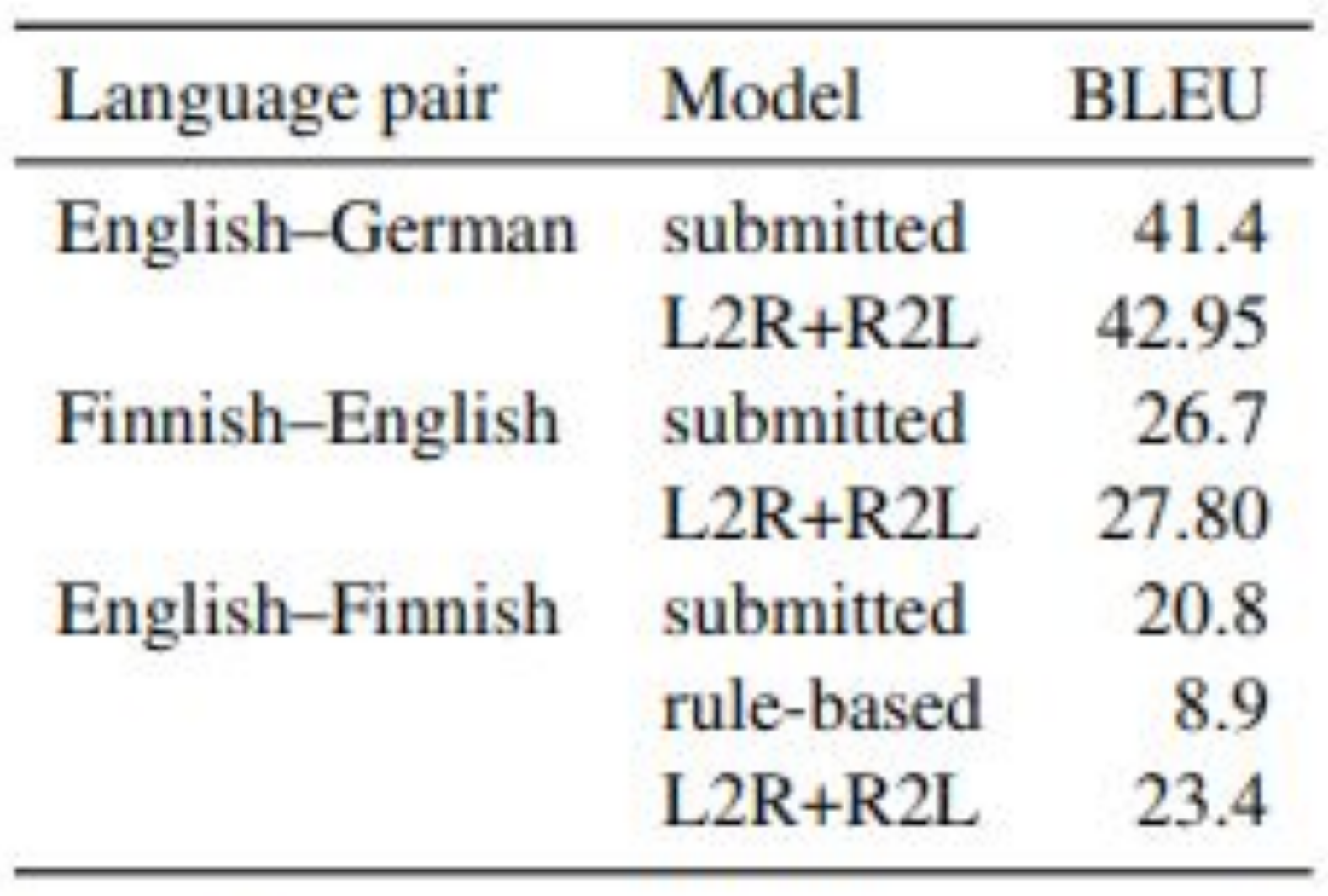
Table 8: Results from individual MarianNMT transformer models and their combinations for English to Finnish (cased BLEU). The *top matrix* result refers to the best system reported in the on-line evaluation matrix (accessed on May 16, 2019).



	BLEU news2017	
Model	L2R	R2L
Run 1	32.26	31.70
Run 2	31.91	31.83
Run 3	32.68	31.81
Ensemble	33.23	33.03
Rescored	33.34	32.98
– L2R+R2L	**33.95**	
Top (with ParaCrawl)	34.6	
Top (without ParaCrawl)	25.9	

Table 9: Results from individual MarianNMT transformer models and their combinations for Finnish to English (cased BLEU). Results denoted as top refer to the top systems reported at the on-line evaluation matrix (accessed on May 16, 2019), one trained with the 2019 data sets and one with 2017 data.

Final submission



Language pair	Model	BLEU
English–German	submitted	41.4
	L2R+R2L	42.95
Finnish–English	submitted	26.7
	L2R+R2L	27.80
English–Finnish	submitted	20.8
	rule-based	8.9
	L2R+R2L	23.4

Table 12: Final results (case-sensitive BLEU scores) on the 2019 news test set; partially obtained after the deadline.

Shared second place in the manual evaluation for
English -> German
Shared third place for
English -> Finnish
Shared fifth place for
Finnish -> English