C-3MA: Tartu-Riga-Zurich Translation Systems for WMT17

Matīss Rikters
Faculty of Computing
University of Latvia
matiss@lielakeda.lv

Chantal Amrhein
University of Zurich
Institute of Computational Linguistics
chantal.amrhein@uzh.ch

Maksym Del and **Mark Fishel**
Institute of Computer Science
University of Tartu
{maksym.del, fishel}@ut.ee

Confidence Scores

System	En->De		De->En	
Values	abs	rel (%)	abs	rel (%)
# recogn. NEs	4546	-	4201	-
# changed NEs	178	3.92	192	4.57
neg -> pos	116	65.17	160	83.33
pos -> neg	53	29.78	22	11.46
neg -> neg	9	5.06	10	5.21

Terminal Visualisations

System	En->De		De->En	
Dataset	Dev	Test	Dev	Test
Baseline NT	27.4	21	31.9	27.2
+ filt. synth.	30.7	22.5	36.8	28.8
+ NE forcing	30.9	22.7	36.9	29

System	En->Lv		Lv->En	
Dataset	Dev	Test	Dev	Test
Baseline NM	11.9	11.9	14.6	12.8
Baseline NT	12.2	10.8	13.2	11.6
Baseline LMT	19.8	12.9	24.3	13.4
+filt. synth. NM	16.7	13.5	15.7	14.3
+filt. synth. NT	16.9	13.6	15.0	13.8
NM+NT+LMT	-	13.6	-	14.3

GitHub Poster

Experimental Settings

Filtered Synthetic Training Data

- Translate 4 million news sentences from the monolingual data of the source language
- Train a character-level RNN from the monolingual news data of the target language
- Score each of the translated 4 million sentences with the language model; drop the worst 50%

Named Entity Forcing

- Recognise NEs in source and target corpora
- Align the NEs with Giza++; filter-out some noise; create a parallel NE dictionary
- After translating a sentence, check if the source had any NEs from the dictionary; replace the aligned word(s) in the translation

Hybrid System Combination

- Translate the same sentence with two different NMT systems and one SMT system; save attention alignment data from the NMT systems
- Choose output from the system that does not
  - Align most of its attention to a single token
  - Have only very strong one-to-one alignments
  - Otherwise - back off to the output of the SMT system

Post-processing

- Replace any
  - *<unk>* tokens in the target with the aligned source tokens
  - consecutive repeating n-grams with a single n-gram

System	BLEU	Rank Human	
		Cluster	Ave %
De->En	6 of 7	6-7 of 7	7 of 7
En->De	10 of 11	9-11 of 11	9 of 11
En->Lv	11 of 12	1-11 of 12	11 of 12
Lv->En	5 of 6	4-5 of 6	4 of 6

ej.uz/C-3MA ej.uz/nmt-poster

Acknowledgements

LATVIJAS UNIVERSITĀTE ANNO 1919 UNIVERSITY OF LATVIA
TARTU ÜLIKOOL 1632 UNIVERSITAS TARTUENSIS
UNIVERSITAS TURICENSIS