CONSTANTINE
THE PHILOSOPHER FACULTY
UNIVERSITY OF NATURAL SCIENCES
IN NITRA AND INFORMATICS

Evaluation of Various Approaches to Compute BLEU Metrics

Lucia Benková and Ľubomír Benko

Constantine the Philosopher University in Nitra

DEPARTMENT OF INFORMATICS
Faculty of Natural Sciences and Informatics CPU in Nitra

Aim of the research

- overview of automatic metrics for evaluating MT suitable for Slovak language,
- filter out the redundant metrics of automatic MT evaluation,
- reference to identify redundant metrics of various sets of similar metrics.

Dataset - corpus

- english – slovak paralel corpus,
- 66 documents of publicistic style (39 354 word tokens),
- Google Translate system (SMT, NMT),
- profesional human translators and post-editors.

Dataset composition

Feature type	Feature name	SMT	NMT	HT	PEMT	SRC
Readability	Average sentence length	17.164	17.236	17.880	17.994	19.414
	Average word length	5.571	5.664	5.764	5.706	4.951
	Number of short sentences	487	493	466	449	413
	Number of long sentences	1557	1551	1578	1595	1631
Lexico-grammatical	Frequency of noun	9314	9365	9999	9877	8713
	Frequency of adjective	4436	4407	4659	4801	3213
	Frequency of verb	4218	4400	4437	4389	5246
	Frequency of determiner	1918	1876	1973	1971	3953
	Frequency of adposition	3735	3875	4129	4155	4680
	Frequency of proper noun	2231	2198	2165	2195	3411
	Frequency of coordinating conj.	1338	1311	1396	1334	1246
	Frequency of subordinating conj.	1352	1403	1281	1377	853
	Frequency of interjection	18	8	9	10	15
	Frequency of adverb	1307	1247	1339	1382	1653
	Frequency of pronoun	1055	1260	1417	1324	2615
	Frequency of auxiliary	1626	1299	1257	1374	2432
	Frequency of numeral	1260	1311	1195	1302	1009
	Frequency of particle	573	598	777	764	1312
	Frequency of punctuation	6668	6674	6460	6646	5370
	Frequency of other	597	561	589	511	3

Methodology

1. obtaining the unstructured data and removing text formatting,
2. machine translation using various systems (SMT, NMT),
3. human translation of documents,
4. post-editing of the MT,
5. segment alignment between the texts,
6. human evaluation of examined MT,
7. automatic evaluation of examined MT (BLEU-1 by nltk, PyTorch, smoothing function and POSBLEU-1+1),
8. comparison of the translation quality,
9. evaluation of obtained results.

Results

- Human evaluation:
  - SMT: 1574 segments with error, 470 segments correct
  - NMT: 386 segments with error, 1658 segments correct
- Automatic evaluation:
  - BLEU-1 – nltk
  - BLEU-1 – PyTorch
  - BLEU-1 – PyTorch with smoothing function
  - POSBLEU-1+1

Verification of global null hypotheses

	NMT=0				NMT=1			
	H-F Epsilon	H-F Adj. df1	H-F Adj. df2	H-F Adj. p	H-F Epsilon	H-F Adj. df1	H-F Adj. df2	H-F Adj. p
BL1	0.5087	1.5260	3116.1310	0.0000	0.5558	1.6675	3404.9990	0.0000
BL1*Evaluation_Error	0.5087	1.5260	3116.1310	0.0000	0.5558	1.6675	3404.9990	0.8424

Redundancy of BLEU-1

NMT=0					NMT=1				
BLEU-1	Mean	1	2	3	BLEU-1	Mean	1	2	3
PyTorch_BLEU-1_smooth	0.504	****			PyTorch_BLEU-1_smooth	0.519	****		
PyTorch_BLEU-1	0.504	****			PyTorch_BLEU-1	0.519	****		
BLEU-1	0.626		****		BLEU-1	0.664		****	
POSBLEU-1+1	0.719			****	POSBLEU-1+1	0.743			****

Conclusion

- redundancy with smoothing technique,
- introduced methodology,
- choose the best of metrics for evaluation MT for Slovak language.

Thank you for your attention!