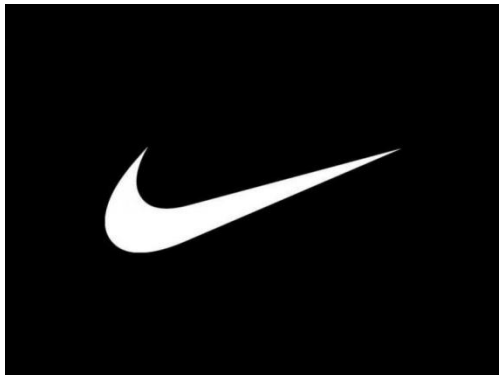
ROMIP MTeval 2013: Report from the organisers

Pavel Braslavski
Alexander Beloborodov
Maxim Khalilov
Serge Sharoff

WE DID IT!



Rationale

- Evaluation is important for development of systems and methods of MT
- No independent evaluation
- Problems with Russian:
  long-distance dependencies (phrase-based SMT)
  ratio of forms to lemmas (Moses operates with forms)

Corpus

- English source texts (En → Ru)
- 8,356 texts = 148,864 sentences
- News (~100,000), en.wikinews.org
- «Regulations» (~48,000), formal texts, extracted from the web using (Sharoff, 2010)
- ~1,000 sentences released as an example

Additionally:

- 1M sentences = Yandex corpus
- 119K sentences = TAUS corpus

The Hearst statement came shortly after White House press secretary Robert Gibbs called her remarks "offensive and reprehensible.“

Ambassadors from the United States of America, Australia and Britain have all met with Fijian military officers to seek insurances that there wasn't going to be a coup.

Acts Not Considered as Farriery Activities such as trimming feet or removing old shoes when there is no intention of fitting shoes afterwards do not fall within the definition of farriery, and so it is not an offence for others to carry them out.

If you are given a discount for booking more than one person onto the same date and you later wish to transfer some of the delegates to another event, the fees will be recalculated and you will be asked to pay additional fees due as well as any administrative charge.

Evaluation method

- Automatic evaluation (comparison against reference translations) – 945 sentences
- Manual evaluation (ranking systems using pairwise comparisons with averaged ranks) – 330 sentences

+ unambiguous ranking

-- greater amount of evaluations

(8 systems → 28 pairs/sentence)

Participants

ID	Short description
P1	Compreno (ABBYY)
P2	Pharaon (anonymous participant) Moses, Yandex and TAUS corpora.
P3,4	Balagur (Data Analysis School) Moses, Yandex (1M) and news (200K)
P5	ETAP-3 (IITP RAS) Rule-based, dictionary with about 100,000 entries
P6,7	Pereved (MIPT) Moses, parallel corpus from the Web

+ 4 online systems (OS1..OS2)

Automatic evaluation

Метрика / ID	OS1	OS2	OS3	OS4	P1	P2	P3	P4	P5	P6	P7
All (947 sentences)											
BLEU	0.150	0.141	0.133	0.124	**0.157**	0.112	0.105	0.073	0.094	0.071	0.073
METEOR	**0.258**	0.240	0.231	0.240	0.251	0.207	0.169	0.133	0.178	0.136	0.149
TER	**0.755**	0.766	0.764	0.758	0.758	0.796	0.901	0.931	0.826	0.934	0.830
GTM	**0.351**	0.338	0.332	0.336	0.349	0.303	0.246	0.207	0.275	0.208	0.230
News (759 sentences)											
BLEU	0.137	0.131	0.123	0.114	**0.153**	0.103	0.096	0.070	0.083	0.066	0.067
METEOR	0.241	0.224	0.214	0.222	**0.242**	0.192	0.156	0.127	0.161	0.126	0.136
TER	0.772	0.776	0.784	0.777	**0.768**	0.809	0.908	0.936	0.844	0.938	0.839
GTM	0.335	0.324	0.317	0.320	**0.339**	0.290	0.233	0.201	0.257	0.199	0.217

TAUS LABS DQF TOOLS

Home

COMPARISON TASK

SOURCE (ENGLISH)

За свою карьеру он сыграл 1,033 сезонов 157 плей-офф кубка Стенли, а также три Игры всех звезд и три Кубка Канады.

TARGETS (ENGLISH)

#	Translation	Select the best translation
1.	его карьера включает 1 033 регулярного сезона и 157 кубок стэнли игр плей-офф , три all-star игры , канады и три чашки .	☐
2.	Его карьера включены 1033 регулярном сезоне и 157 Stanley Cup играх плей-офф, три All-Star игр, и три Кубка Канады.	☐

COMMENTS

Segment: 3 of 33
Filename: translation_45365-85132.txt

Previous Next

Human evaluation

- 14 evaluators (volunteers and participants)
- Total evaluation size – 10,920 pairs
- Time for each pair ~30-90 sec
- ~1 person-month
- Percentage agreement on 1,680 pairs – 56%
- Kappa = 0.34

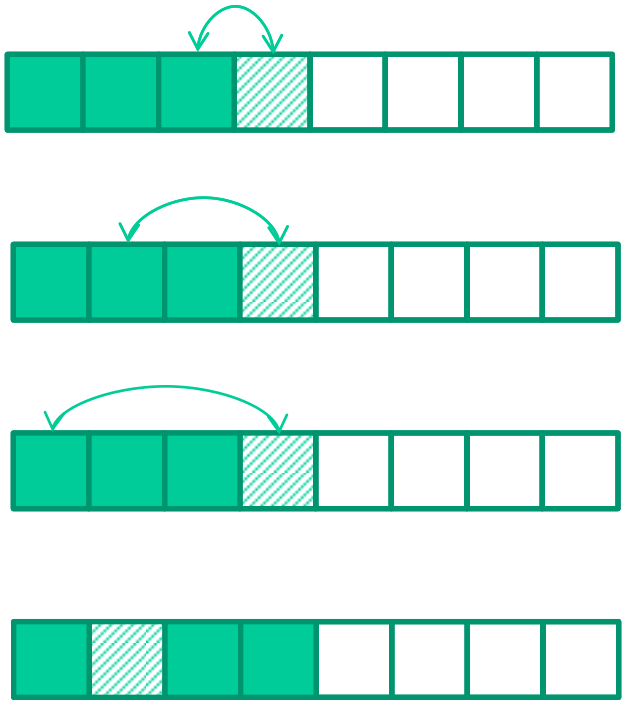
Human evaluation results

All (330 sentences)							
OS3	P1	OS1	OS2	OS4	P5	P2	P4
3.159	3.350	3.530	3.961	4.082	5.447	5.998	6.473
News (190 sentences)							
OS3	P1	OS1	OS2	OS4	P5	P2	P4
2.947	3.450	3.482	4.084	4.242	5.474	5.968	6.353
Regulations (140 sentences)							
P1	OS3	OS1	OS2	OS4	P5	P2	P4
3.214	3.446	3.596	3.793	3.864	5.411	6.039	6.636

Groups with statistically significant difference:
(OS1, OS3, P1) > (OS2, OS4) > P5 > (P2, P4)

How to reduce evaluation?

- Human-assisted sorting
- Insertion sort:
  best case: *n-1*
  worst case: *n(n-1)/2*
- Binary insertion sort
  ~*n*log*n*



Results

- Insertion sort: 5,131 comparison operations (15.5 per sentence; 56% of the exhaustive comparison);
- binary insertion sort: 4,327 comparison operations (13.1 per sentence; 47% of the exhaustive comparison).

Exhaustive vs reduced evaluations

Insertion sort							
P1	OS1	OS3	OS2	OS4	P5	P4	P2
3.318	3.327	3.588	4.221	4.300	5.227	5.900	6.118
Binary insertion sort							
OS1	P1	OS3	OS2	OS4	P5	P2	P4
2.924	3.045	3.303	3.812	4.267	5.833	5.903	6.882

Groups with statistical significant difference:
(OS1, OS3, P1) > (OS2, OS4) > P5 > (P2, P4)

Plans for 2014

- More genres
- Russian → English
- Tuning automatic similarity measures for translation into Russian
- Inviting more participants (+ «light track»)

Data

http://romip.ru/mteval/data/