

PLIN009 – Strojový překlad

Automatické hodnocení kvality SP

Drobné kapitoly o SP

Vít Baisa

jaro 2012

Motivace

- **plynulost** (fluency) – je překlad plynulý, má přirozený slovosled?
- **adekvátnost** (adequacy) – zachovává překlad význam, nebo je změněn, nekompletní?
- **srozumitelnost** (intelligibility)
- neplést s **přesností** (precision) a **pokrytím** (recall)

Stupnice hodnocení

adekvátnost	
5	veškerý význam
4	většina významu
3	dostatečně významu
2	málo z původního významu
1	žádný význam

plynulost	
5	bezchybný jazyk
4	dobrý jazyk
3	nepřirozený
2	neplynulý jazyk
1	nesrozumitelný

Anotační nástroj

Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l'ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5
Annotator: Philipp Koehn Task: WMT06 French-English	<input type="button" value="Annotate"/>	
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Nevýhody ručního hodnocení

- ruční hodnocení je pomalé, drahé, subjektivní
- mezinotátorská shoda (MAS) ukazuje, že se lidé shodnou více na plynulosti než na adekvátnosti
- jiné hodnocení: je X lepší překlad než Y?
- → ještě větší MAS

Automatické hodnocení překladu

- výhody: rychlost, cena; nevýhody: měříme opravdu kvalitu?
- gold standard: ručně připravené referenční překlady
- kandidát c se srovnává s n referenčními překlady r_i
- paradox automatického hodnocení: úkol AHKSP odpovídá situaci, kdy má student hodnotit svou vlastní písemnou práci: jak pozná, v čem udělal chybu?
- různé přístupy: n -gramová shoda mezi c a r_i , editační vzdálenost, . . .

Pokrytí a přesnost na slovech

Nejjednodušší způsob automatického hodnocení

SYSTEM A: Israeli officials responsibility of airport safety

REFERENCE: Israeli officials are responsible for airport security

- přesnost

$$\frac{\textit{correct}}{\textit{output-length}} = \frac{3}{6} = 50\%$$

- pokrytí

$$\frac{\textit{correct}}{\textit{reference-length}} = \frac{3}{7} = 43\%$$

- f-score

$$\frac{\textit{precision} \times \textit{recall}}{(\textit{precision} + \textit{recall})/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$$

Pokrytí a přesnost – nedostatky



metrika	system A	system B
přesnost	50%	100%
pokrytí	43%	100%
f-score	46%	100%

Nepostihuje se nesprávný slovosled.

BLEU

- nejznámější (standard), nejpoužívanější, nejstarší (2001)
- IBM, Papineni
- n-gramová shoda mezi referencí a kandidáty
- počítá se přesnost pro 1 až 4-gramy
- extra postih za krátkost (**brevity penalty**)

$$\text{BLEU} = \min \left(1, \frac{\text{output-length}}{\text{reference-length}} \right) \left(\prod_{i=1}^4 \text{precision}_i \right)^{\frac{1}{4}}$$

BLEU – příklad

SYSTEM A: Israeli officials responsibility of airport safety
 2-GRAM MATCH 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible
 2-GRAM MATCH 4-GRAM MATCH

metrika	system A	system B
přesnost (1gram)	3/6	6/6
přesnost (2gram)	1/5	4/5
přesnost (3gram)	0/4	2/4
přesnost (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0 %	52 %

Další metriky

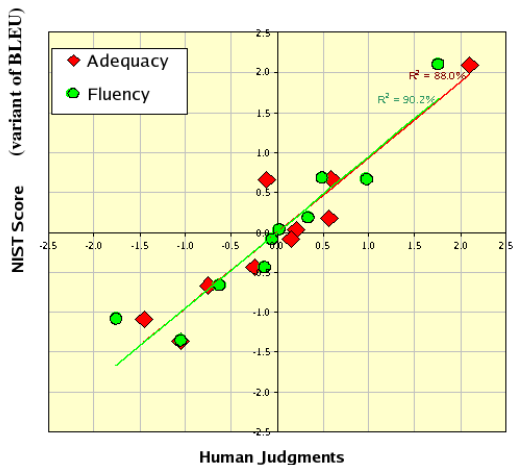
- NIST
 - NIST: National Institute of Standards and Technology
 - vážení shod n-gramů podle informační hodnoty
 - velmi podobné výsledky jako BLEU (varianta)
- NEVA
 - Ngram EVALuation
 - úprava BLEU skóre pro kratší věty
 - bere v potaz i synonyma (kladně hodnotí použití synonyma ve smyslu stylistické bohatosti)
- WAFT
 - Word Accuracy for Translation
 - editační vzdálenost mezi c a r
 - $WAFT = 1 - \frac{d+s+i}{\max(l_r, l_c)}$

Další metriky II

- TER
 - Translation Edit Rate
 - nejmenší počet kroků (smazání, přidání, prohození, změna)
 - $TER = \frac{\text{počet editací}}{\text{prům. počet ref. slov}}$
 - $r =$ dnes jsem si při fotbalu zlomil kotník
 - $c =$ při fotbalu jsem si dnes zlomil kotník
 - $TER = 4/7$
- HTER
 - Human TER
 - nejdříve ručně vytvořena r a na ni aplikováno TER
- METEOR
 - uvažuje synonyma (WordNet) a
 - morfologické varianty slov

Hodnocení hodnotících metrik

Korelace automatického hodnocení s manuálním.



Hodnocení překladu – EuroMatrix

		output language									
input language	Danish	BLEU 21.47	BLEU 18.49	BLEU 21.12	BLEU 28.57	BLEU 14.24	BLEU 28.79	BLEU 22.22	BLEU 24.32	BLEU 26.49	BLEU 28.33
	Dutch	BLEU 20.51	BLEU 18.39	BLEU 17.49	BLEU 23.01	BLEU 10.34	BLEU 24.67	BLEU 20.07	BLEU 20.71	BLEU 22.95	BLEU 19.03
	German	BLEU 22.95	BLEU 23.40	BLEU 20.75	BLEU 25.36	BLEU 11.88	BLEU 27.75	BLEU 21.36	BLEU 23.28	BLEU 25.49	BLEU 20.51
	Greek	BLEU 22.79	BLEU 20.02	BLEU 17.42	BLEU 27.28	BLEU 11.44	BLEU 32.15	BLEU 26.84	BLEU 27.67	BLEU 31.26	BLEU 21.23
	English	BLEU 25.24	BLEU 21.02	BLEU 17.64	BLEU 23.23	BLEU 13.00	BLEU 31.16	BLEU 25.39	BLEU 27.10	BLEU 30.16	BLEU 24.83
	Finnish	BLEU 20.02	BLEU 17.09	BLEU 14.57	BLEU 18.20	BLEU 21.86	BLEU 22.49	BLEU 18.39	BLEU 19.14	BLEU 21.16	BLEU 18.65
	French	BLEU 23.73	BLEU 21.13	BLEU 18.54	BLEU 26.13	BLEU 30.00	BLEU 12.63	BLEU 32.48	BLEU 35.37	BLEU 38.47	BLEU 22.68
	Italian	BLEU 21.47	BLEU 20.07	BLEU 16.92	BLEU 24.83	BLEU 27.89	BLEU 11.08	BLEU 36.09	BLEU 31.20	BLEU 34.04	BLEU 20.26
	Portuguese	BLEU 23.27	BLEU 20.23	BLEU 18.27	BLEU 26.46	BLEU 30.11	BLEU 11.99	BLEU 39.04	BLEU 32.07	BLEU 37.95	BLEU 21.96
	Spanish	BLEU 24.10	BLEU 21.42	BLEU 18.29	BLEU 28.38	BLEU 30.51	BLEU 12.57	BLEU 40.27	BLEU 32.31	BLEU 35.92	BLEU 23.90
	Swedish	BLEU 30.35	BLEU 21.94	BLEU 18.97	BLEU 22.86	BLEU 30.20	BLEU 15.37	BLEU 29.77	BLEU 23.94	BLEU 25.95	BLEU 28.66

Hodnocení překladu podle jazykových párů – II

	Target Language																					
	EN	BG	DE	CS	DA	EL	ES	ET	FI	FR	HU	IT	LT	LV	MT	NL	PL	PT	RO	SK	SL	SV
EN	↻	40.5	46.8	52.6	50.0	41.0	55.2	34.8	38.6	50.1	37.2	50.4	39.6	43.4	39.8	52.3	48.2	53.0	49.0	44.7	50.7	52.0
BG	61.3	↻	38.7	39.4	39.6	34.5	46.9	25.5	26.7	42.4	22.0	43.5	29.3	29.1	25.9	44.9	35.1	45.9	36.8	34.1	34.1	39.9
DE	53.6	26.3	↻	35.4	43.1	32.8	47.1	26.7	29.5	39.4	27.6	42.7	27.6	30.3	19.8	50.2	30.2	44.1	30.7	29.4	31.4	41.2
CS	58.4	32.0	42.6	↻	43.6	34.6	48.9	30.7	30.5	41.6	27.4	44.3	34.5	35.8	26.3	46.5	39.2	45.7	36.5	43.6	41.3	42.9
DA	57.6	28.7	44.1	35.7	↻	34.3	47.5	27.8	31.6	41.3	24.2	43.8	29.7	32.9	21.1	48.5	34.3	45.4	33.9	33.0	36.2	47.2
EL	59.5	32.4	43.1	37.7	44.5	↻	54.0	26.5	29.0	48.3	23.7	48.6	29.0	32.6	23.8	48.9	34.2	52.5	37.2	33.1	36.3	43.3
ES	60.0	31.1	42.7	37.5	44.4	39.4	↻	25.4	28.5	51.3	24.0	51.7	26.8	30.5	24.6	48.8	33.9	57.3	38.1	31.7	33.9	43.7
ET	52.0	24.6	37.3	35.2	37.8	28.2	40.4	↻	37.7	33.4	30.9	37.0	35.0	36.9	20.5	41.3	32.0	37.8	28.0	30.6	32.9	37.3
FI	49.3	23.2	36.0	32.0	37.9	27.2	39.7	34.9	↻	29.5	27.2	36.6	30.5	32.5	19.4	40.6	28.8	37.5	26.5	27.3	28.2	37.6
FR	64.0	34.5	45.1	39.5	47.4	42.8	60.9	26.7	30.0	↻	25.5	56.1	28.3	31.9	25.3	51.6	35.7	61.0	43.8	33.1	35.6	45.8
HU	48.0	24.7	34.3	30.0	33.0	25.5	34.1	29.6	29.4	30.7	↻	33.5	29.6	31.9	18.1	36.1	29.8	34.2	25.7	25.6	28.2	30.5
IT	61.0	32.1	44.3	38.9	45.8	40.6	26.9	25.0	29.7	52.7	24.2	↻	29.4	32.6	24.6	50.5	35.2	56.5	39.3	32.5	34.7	44.3
LT	51.8	27.6	33.9	37.0	36.8	26.5	21.1	34.2	32.0	34.4	28.5	36.8	↻	40.1	22.2	38.1	31.6	31.6	29.3	31.8	35.3	35.3
LV	54.0	29.1	35.0	37.8	38.5	29.7	25.3	34.2	32.4	35.6	29.3	38.9	38.4	↻	23.3	41.5	34.4	39.6	31.0	33.3	37.1	38.0
MT	72.1	32.2	37.2	37.9	38.9	33.7	48.7	26.9	25.8	42.4	22.4	43.7	30.2	33.2	↻	44.0	37.1	45.9	38.9	35.8	40.0	41.6
NL	56.9	29.3	46.9	37.0	45.4	35.3	49.7	27.5	29.8	43.4	25.3	44.5	28.6	31.7	22.0	↻	32.0	47.7	33.0	30.1	34.6	43.6
PL	60.8	31.5	40.2	44.2	42.1	34.2	46.2	29.2	29.0	40.0	24.5	43.2	33.2	35.6	27.9	44.8	↻	44.1	38.2	38.2	39.8	42.1
PT	60.7	31.4	42.9	38.4	42.8	40.2	60.7	26.4	29.2	53.2	23.8	52.8	28.0	31.5	24.8	48.3	34.5	↻	39.4	32.1	34.4	43.9
RO	60.8	33.1	38.5	37.8	40.3	35.6	50.4	24.8	26.2	46.5	25.0	44.8	28.4	29.9	28.7	43.0	35.8	48.5	↻	31.5	35.1	39.4
SK	60.8	32.6	39.4	48.1	41.0	33.3	46.2	29.8	28.4	39.4	27.4	41.8	33.8	36.7	28.5	44.4	39.0	43.3	35.3	↻	42.6	41.8
SL	61.0	33.1	37.9	43.5	42.6	34.0	47.0	31.1	28.8	38.2	25.7	42.3	34.6	37.3	30.0	45.9	38.2	44.1	35.8	38.9	↻	42.7
SV	58.5	26.9	41.0	35.6	46.6	33.3	46.6	27.4	30.9	38.9	22.7	42.0	28.2	31.0	23.7	45.6	32.2	44.2	32.7	31.3	33.5	↻