IA161 Pokročilé techniky zpracování přirozeného jazyka

Strojový překlad

Vít Baisa

Překlad angličtina→čeština

Moses is an implementation of the statistical (or data-driven) approach to machine translation (MT). This is the dominant approach in the field at the moment, and is employed by the online translation systems deployed by the likes of Google and Microsoft.

Mojžíš je implementace statistické (nebo řízené daty) přístupu k strojového překladu (MT). To je převládajícím přístupem v oblasti v současné době, a je zaměstnán pro on-line překladatelských systémů nasazených likes Google a Microsoft.

Moses je implementace statistického (nebo daty řízeného) přístupu k strojovému překladu (MT). V současné době jde o převažující přístup v rámci strojového překladu, který je použit online překladovými systémy nasazenými Googlem a Microsoftem.

Mojžíš je provádění statistické (nebo aktivovaný) přístup na strojový překlad (mt). To je dominantní přístup v oblasti v tuto chvíli, a zaměstnává on - line překlad systémů uskutečněné takové, Google a Microsoft.

Statistický strojový překlad

- pravidlové systémy motivovány lingvistikou
- SMT inspirován teorií informace a statistikou
- 50 miliónů stránek denně přeložených pomocí SMT
- Google Translate, Bing Translator, Moses
- **gisting**: nejčastější užití MT na internetu
- ve skutečnosti je výstup z MT vždy revidován

Strojový překlad: co se překládá

- webové stránky
- technické manuály, návody
- vědecké dokumenty
- prospekty, katalogy
- obecně texty z omezených domén
- stránky na Wikipedii (mezi jaz. mutacemi)

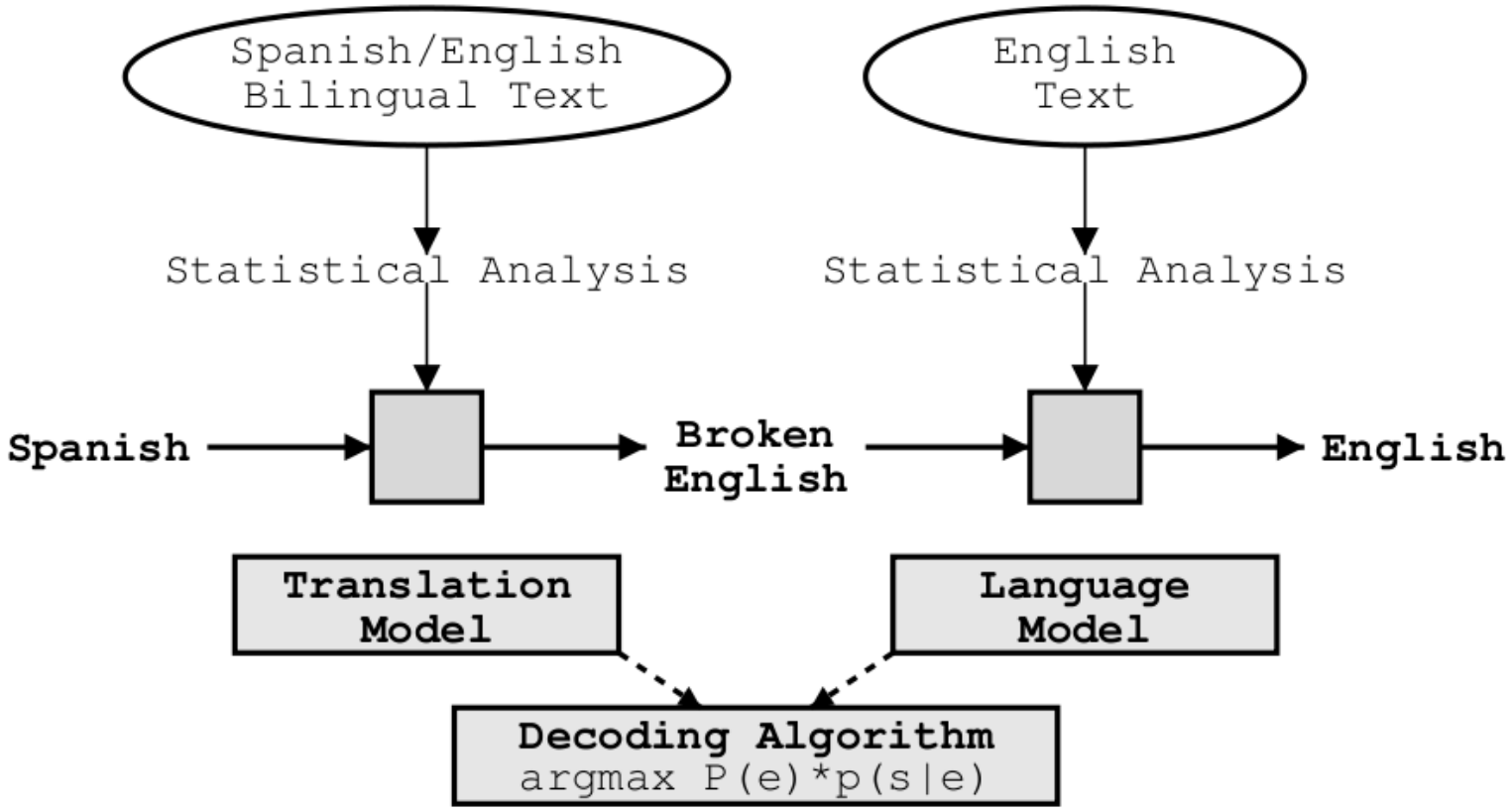
Strojový překlad v současnosti

- intenzivní sběr paralelních dat
- vývoj systémů vzhledem k hodnoticím metrikám
- západ: zájem o angličtinu jako cílový jazyk
- EU: překlad mezi 24 úředními jazyky EU (EuroMatrix)
- softwarové firmy zaměřeny na En jako zdrojový jazyk
- velké páry (En$\leftrightarrow$Sp, En$\leftrightarrow$Fr): velmi dobrý překlad
- Google Translate jako gold standard
- morfologicky bohaté jazyky: horší výsledky
- En-* a *-En páry převažují
- Moses: volně dostupný statistický strojový překlad

Data: paralelní korpusy

- Europarl: kolekce textů Evropského parlamentu
- OPUS: paralelní texty různého původu, největší
- Acquis Communautaire: zákony EU (demo: EUR-Lex)
- DGT překladová paměť, MyMemory
- InterCorp – ručně zarovnané beletr. texty (ČNK, FFUK)
- volně dostupné jsou řádově 10 až 100 miliónů slov veliké
- vícejazyčné stránky (Wikipedie)
- srovnatelné korpusy (comparable corpora): texty ze stejné domény

Schéma SMT


Spanish/English
Bilingual Text
English
Text
Statistical Analysis
Statistical Analysis
Spanish
Broken
English
English
Translation
Model
Language
Model
Decoding Algorithm
argmax P(e)*p(s|e)

SMT – princip noisy channel

Vyvinut Shannonem (1948) pro potřeby samoopravujících se kódů, pro korekce kódovaných signálů přenášených po zašuměných linkách na základě informace o původní zprávě a typu chyb vznikajících na linkách.

OCR. Rozpoznávání textu z obrázků je chybové, ale dokážeme odhadnout, co by mohlo být v textu (jazykový model) a jaké chyby často vznikají: záměna l-1-I, rn-m apod.

$$\begin{aligned} e^* &= \arg\max_e p(e|f) \\ &= \arg\max_e \frac{p(e)p(f|e)}{p(f)} \\ &= \arg\max_e p(e)p(f|e). \end{aligned}$$

Jazykovým modelům se budeme věnovat na pozdější přednášce.

Lexikální překlad

Standardní slovník neobsahuje překladové pravděpodobnosti jednotlivých významů slov.

key $\rightarrow$ *klíč, tónina, klávesa*

Jak často jsou jednotlivé překladové ekvivalenty (významy) v překladech používány?

key $\rightarrow$ *klíč* (0.7), *tónina* (0.18), *klávesa* (0.08), ...

Potřebujeme lexikální pravděpodobnostní rozložení p_f s vlastností:

$$\sum_e p_f(e) = 1$$

$$\forall e : 0 \leq p_f(e) \leq 1$$

$$p_{\text{klíč}}(\text{key}) \; ? \; p_{\text{mrkev}}(\text{carrot})$$

Výpočet překladové pravděpodobnosti

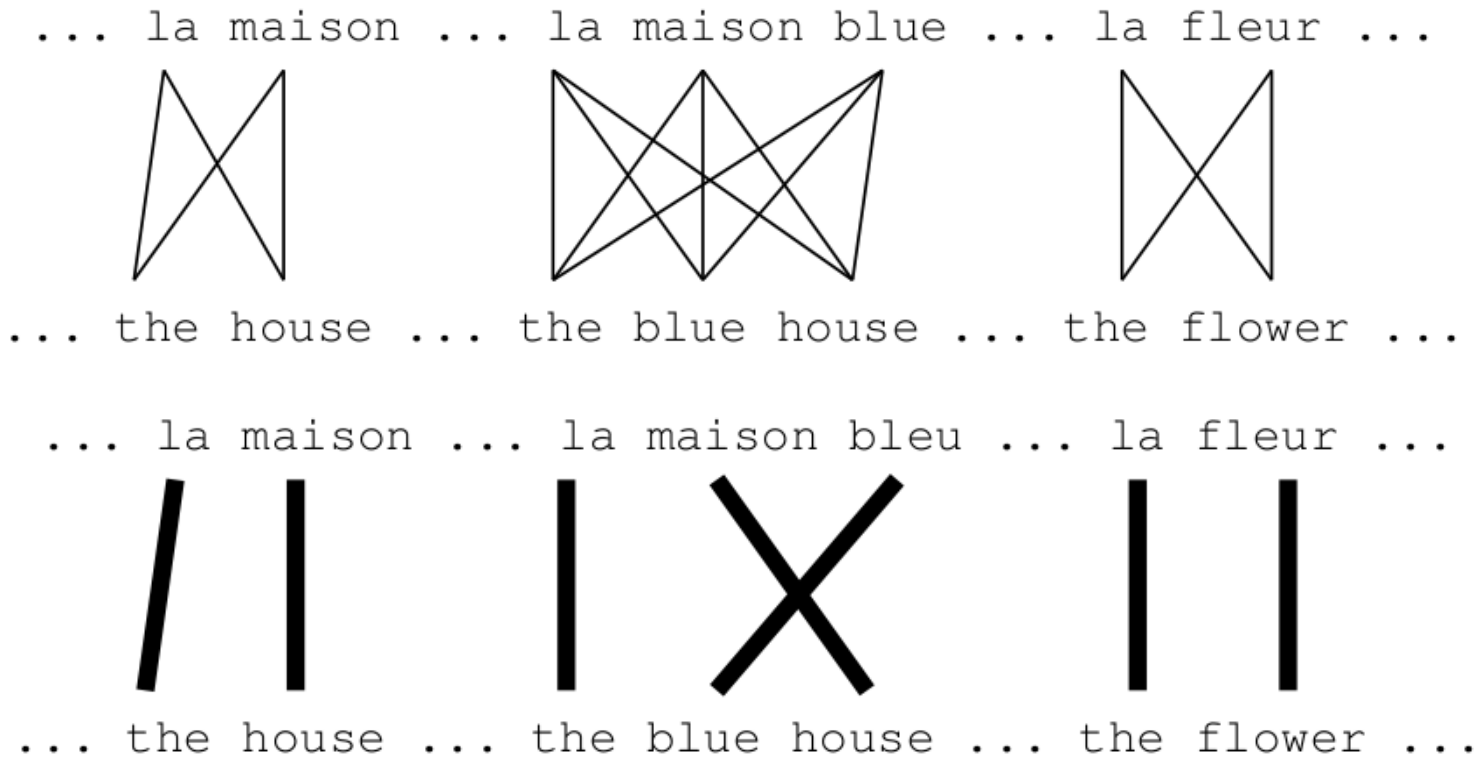
Potřebujeme znát hodnotu funkce t pro všechna slova (věty).

K tomu se využívá **paralelní korpus** se zarovnanými větami. Ty většinou nejsou k dispozici. To je úkol tzv. **sentence-alignment**.

Většinou se využívá:

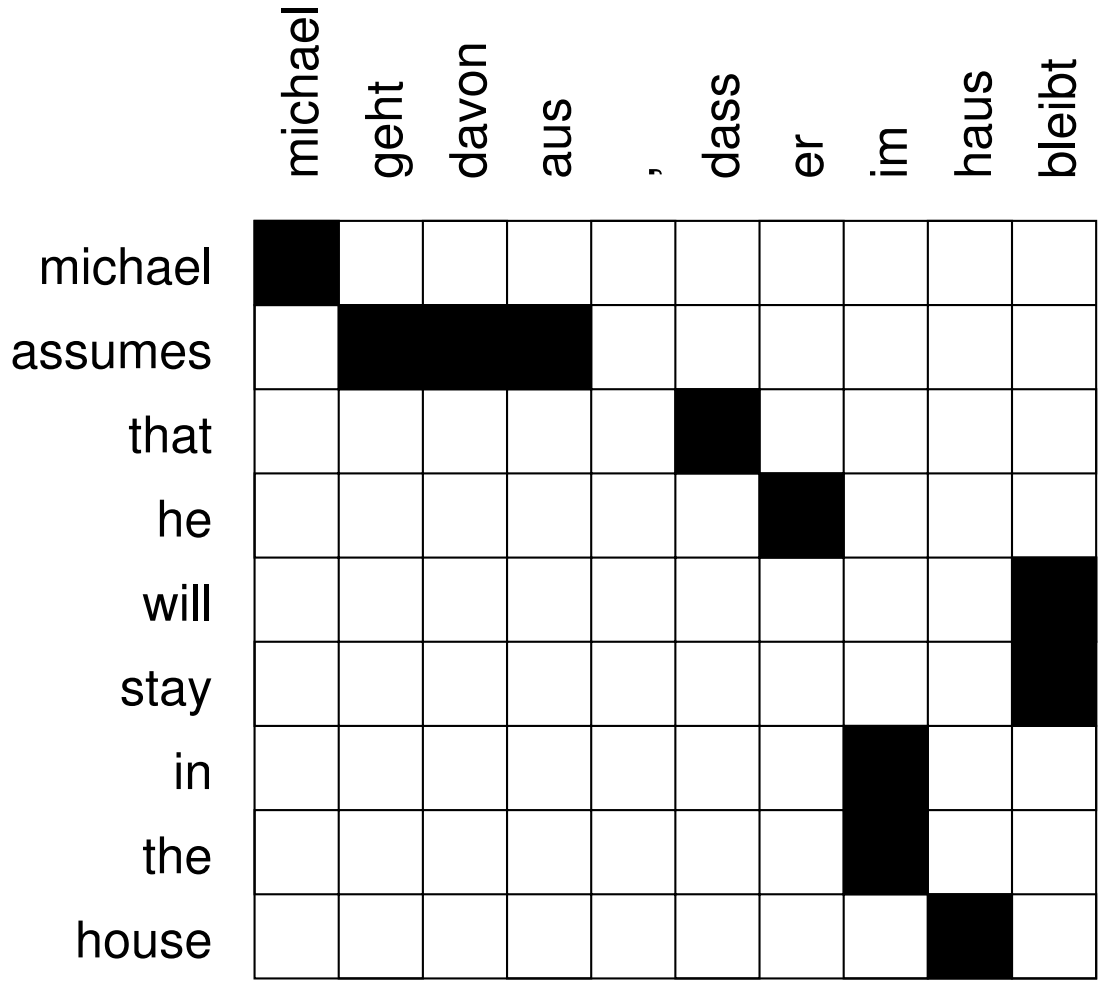
- srovnání délky vět,
- překladový slovník či
- spoluvýskyt jmen, čísel, znaků, málo častých slov.

Word alignment

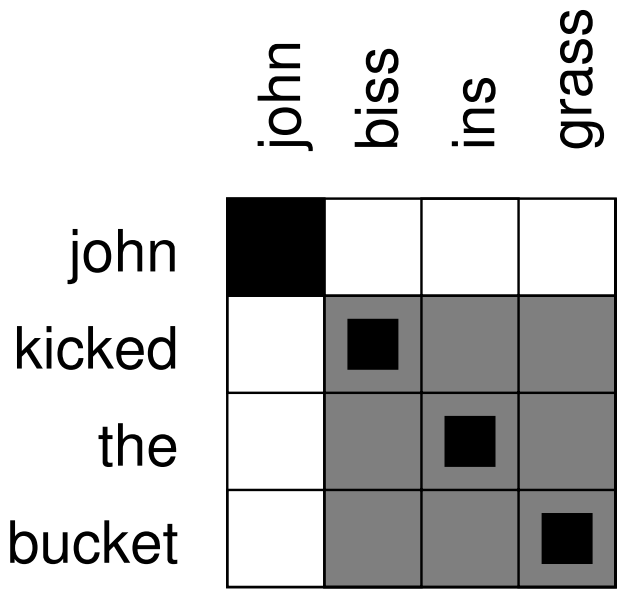


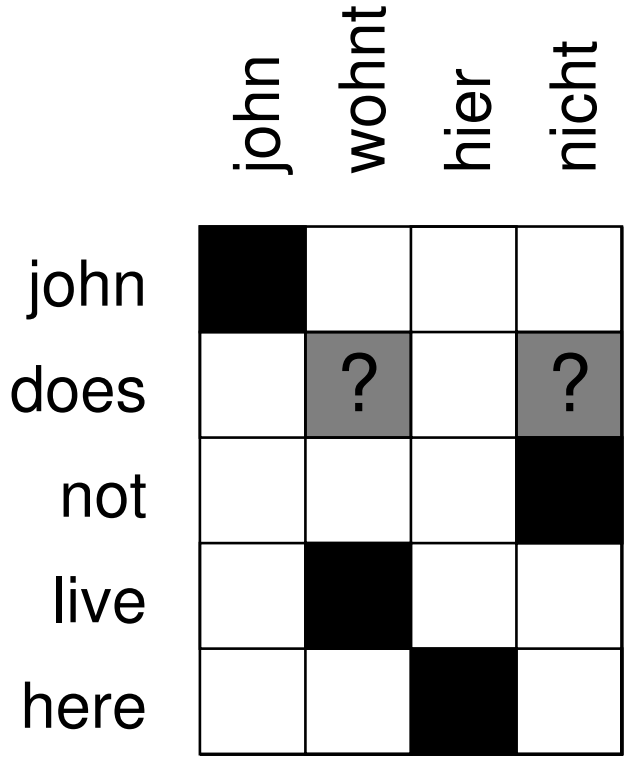
$$p(la|the) = 0.453$$
$$p(le|the) = 0.334$$
$$p(maison|house) = 0.876$$
$$p(bleu|blue) = 0.563$$
$$\ldots$$

Matice zarovnání slov



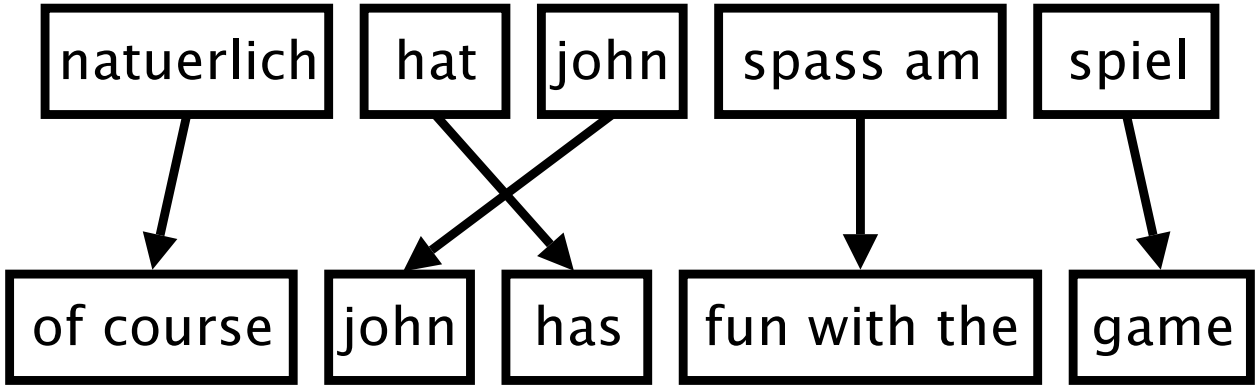
Problémy s lexikálním překladem





Frázový překladový model

State-of-the-art statistického strojového překladu. Nepřekládají se pouze samostatná slova. Když to jde, tak i celé sekvence slov.

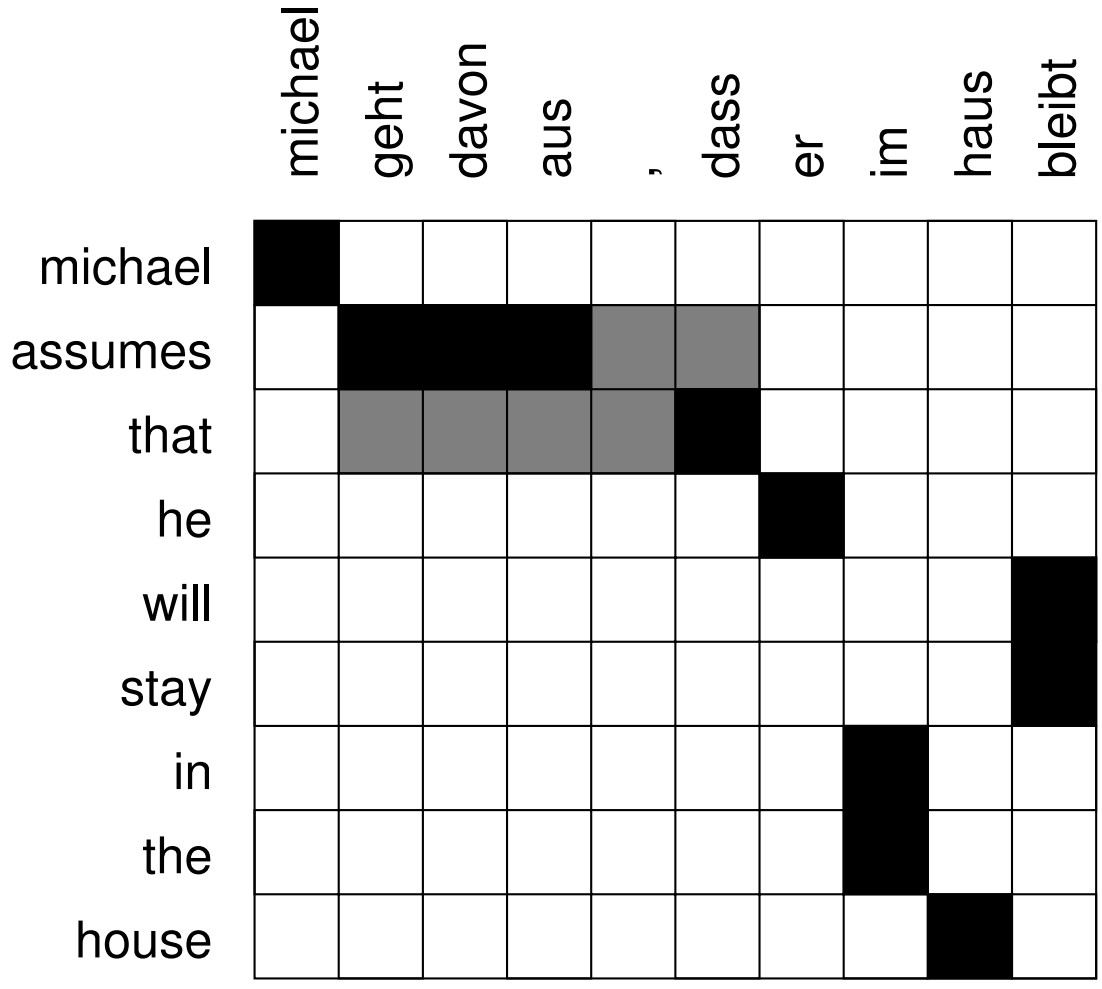


Fráze nejsou lingvisticky motivované, pouze statisticky. Německé *am* se zřídka překládá jedním slovem *with*. Statisticky významný kontext *spass am* pomáhá správnému překladu. Klasické fráze by se dělily jinak: *(fun (with (the game)))*.

Výhody frázového překladu

- často překládáme $n : m$ slov, slovo je tedy nevhodný atomický prvek
- překlad skupin slov pomáhá řešit překladové víceznačnosti
- a také plynulost
- můžeme se učit překládat delší a delší fráze
- jednodušší model: neuvažujeme fertilitu, NULL token atd.

Extrahování frází



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
- různé přístupy: n-gramová shoda mezi c a r_i, editační vzdálenost, . . .
- BLEU: nejpoužívanější
- METEOR: nejlépe koreluje s lidským hodnocením

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**)

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

BLEU – příklad

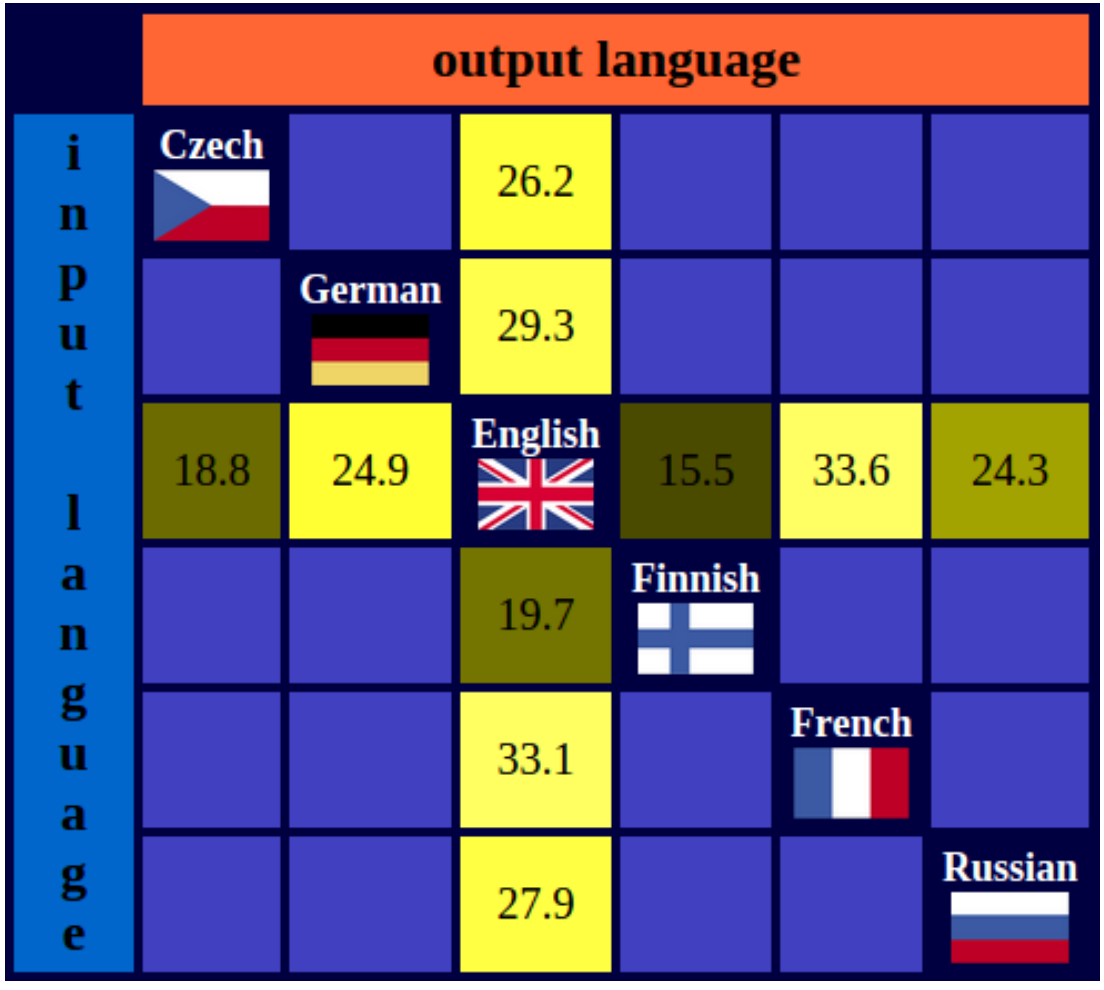
SYSTEM A: Israeli officials responsibility of airport safety
2-GRAM MATCH 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

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

metrika	systém A	systém 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 %

Hodnocení překladu podle jazykových párů



http://matrix.statmt.org/