Machine Translation 1: Introduction, Approaches, Evaluation, Word Alignment

Ondřej Bojar
bojar@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

Outline of Lectures on MT

1. Introduction.
   - Why is MT difficult.
   - MT evaluation.
   - Approaches to MT.
   - First peek into phrase-based MT
   - Document, sentence and word alignment.
2. Statistical Machine Translation.
   - Phrase-based: Assumptions, beam search, key issues.
   - Neural MT: Sequence-to-sequence, attention, self-attentive.
3. Advanced Topics.
   - Linguistic Features in SMT and NMT.
   - Multilinguality, Multi-Task, Learned Representations.

Supplementary Materials

Videolectures & Wiki:

`http://mttalks.ufal.ms.mff.cuni.cz/`

Slides and Lectures from MT Marathon (see Programme):

`http://www.statmt.org/mtm15` and the neural `/mtm16`

Books:

- Ondřej Bojar: Čeština a strojový překlad. ÚFAL, 2012.
- Philipp Koehn: Statistical Machine Translation. Cambridge University Press, 2009.
  With some slides: `http://statmt.org/book/`
  NMT: `https://arxiv.org/pdf/1709.07809.pdf`

Why is MT Difficult?

- Ambiguity and word senses.
- Target word forms.
- Negation.
- Pronouns.
- Co-ordination and apposition; word order.
- Space of possible translations.

. . . aside from the well-known hard things like idioms:

John kicked the bucket.

Ambiguity and Word Senses

The plant is next to the bank.

Ambiguity and Word Senses

The plant is next to the bank.

He is a big data scientist: (big data) scientist or big (data scientist)?

Put it on the rusty/velvety coat rack.

Ambiguity and Word Senses

The plant is next to the bank.

He is a big data scientist: (big data) scientist or big (data scientist)?

Put it on the rusty/velvety coat rack.

Spal celou Petkevičovu přednášku.

Ženu holí stroj.

Ambiguity and Word Senses

The plant is next to the bank.

He is a big data scientist: (big data) scientist or big (data scientist)?

Put it on the rusty/velvety coat rack.

Spal celou Petkevičovu přednášku.

Ženu holí stroj.

Dictionary entries are not much better:

kniha účetní, napětí dovolené, plán prací, tři prdele

Ambiguity and Word Senses

The plant is next to the bank.

He is a big data scientist: (big data) scientist or big (data scientist)?

Put it on the rusty/velvety coat rack.

Spal celou Petkevičovu přednášku.

Ženu holí stroj.

Dictionary entries are not much better:

kniha účetní, napětí dovolené, plán prací, tři prdele

A real-world example:

SRC	One tap and the machine issues a slip with a number.
REF	Jedno ťuknutí a ze stroje vyjede papírek s číslem.
Moses 1	Z jednoho kohoutku a stroj vydá složenky s číslem.
Moses 2	Jeden úder a stroj vydá složenky s číslem.
Google	Jedním klepnutím a stroj problémy skluzu s číslem.

Target Word Form

Tense:

- English present perfect for recent past events.
- Spanish has two types of past tense: a specific and indetermined time in the past.

Cases, genders, . . . :

- Czech has 7 cases, 3 numbers and 4 genders:

The cat is on the mat. → kočka

He saw a cat. → kočku

He saw a dog with a cat. → kočkou

He talked about a cat. → kočce

⇒ Need to choose the right form when producing Czech.

Context Needed to Choose Right

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	kočky	.
	pily	dvě	zelená	pruhovaná	koček	
	. . .	dvou	zelené	pruhované	kočkám	
	viděl	dvěma	zelení	pruhovaní	kočkách	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	. . .		zelených	pruhovaných		
	uviděl		zelenému	pruhovanému		
	uviděla		zeleným	pruhovaným		
	. . .		zelenou	pruhovanou		
viděl jsem			zelenými	pruhovanými		
viděla jsem			. . .	. . .		

Context Needed to Choose Right

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	**kočky**	.
	pily	**dvě**	zelená	pruhovaná	koček	
	. . .	dvou	**zelené**	**pruhované**	kočkám	
	viděl	dvěma	zelení	pruhovaní	kočkách	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	. . .		zelených	pruhovaných		
	uviděl		zelenému	pruhovanému		
	uviděla		zeleným	pruhovaným		
	. . .		zelenou	pruhovanou		
viděl jsem			zelenými	pruhovanými		
viděla jsem			. . .	. . .		

Context Needed to Choose Right

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	**kočky**	.
	pily	**dvě**	zelená	pruhovaná	koček	
	. . .	**dvou**	**zelené**	**pruhované**	kočkám	
	viděl	dvěma	zelení	pruhovaní	**kočkách**	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	. . .		**zelených**	**pruhovaných**		
zrak mi utkvěl na			zelenému	pruhovanému		
	uviděla		zeleným	pruhovaným		
	. . .		zelenou	pruhovanou		
viděl jsem			zelenými	pruhovanými		
viděla jsem			. . .	. . .		

Negation

- French negation is around the verb:

  Je ne parle pas français.

- Czech negation is doubled:

  Nemám žádné námitky.

- Northern and southern Italy supposedly differ in the semantics of what you're doing with your public transport ticket upon entering the bus:

  make valid or invalid (in/validare).

- Some sentences even ambiguous with respect to negation:

  Baterky už došly. (No batteries left. Batteries just arrived.)

  Z práce odcházím dobita. (I leave the work exhausted/recharged.)

Pronouns

- English requires the subject explicit ⇒ guess from the verb:

  Četl knihu. = He read a book.

  Spal jsem. = I slept.

- The gender must match the referent:

  He saw a book. It was red.

  Viděl knihu. Byla černá.

  He saw a pen. It was red.

  Viděl pero. Bylo černé.

- Czech agreement with subject:

Source	Could I use your cell phone?
Google	Mohl bych používat svůj mobilní telefon?
Moses	Mohl jsem použít svůj mobil?

Co-ordination, Apposition; Order

Co-ordination and apposition:

- How many people were there? The comma tells us:

  Předseda vlády, Petr Nečas , a Martin Lhota přednesli příspěvky o...

- Which scope ("brackets") is the outer one?

Input	We have both countries inside and outside the Eurozone.
Reference	Máme tu země eurozóny a země stojící mimo eurozónu.
MT Output	Máme obě země uvnitř a vně eurozóny.

Word order:

- $n!$ word permutations in principle.
- More on this next week.

Space of Possible Translations

How many good translations has the following sentence?

And even though he is a political veteran, the Councilor Karel Brezina responded similarly.

Space of Possible Translations

Examples of 71 thousand correct translations of the English:

And even though he is a political veteran, the Councilor Karel Brezina responded similarly.

A ačkoli ho lze považovat za politického veterána, radní Březina reagoval obdobně.
Ač ho můžeme prohlásit za politického veterána, reakce radního Karla Březiny byla velmi obdo
A i přestože je politický matador, radní Karel Březina odpověděl podobně.
A přestože je to politický veterán, velmi obdobná byla i reakce radního K. Březiny.
A radní K. Březina odpověděl obdobně, jakkoli je politický veterán.
A třebaže ho můžeme považovat za politického veterána, reakce Karla Březiny byla velmi podo
Byť ho lze označit za politického veterána, Karel Březina reagoval podobně.
Byť ho můžeme prohlásit za politického veterána, byla i odpověď K. Březiny velmi podobná.
K. Březina, i když ho lze prohlásit za politického veterána, odpověděl velmi obdobně.
Odpověď Karla Březiny byla podobná, navzdory tomu, že je politickým veteránem.
Radní Březina odpověděl velmi obdobně, navzdory tomu, že ho lze prohlásit za politického vet
Reakce K. Březiny, třebaže je politický veterán, byla velmi obdobná.
Velmi obdobná byla i odpověď Karla Březiny, ačkoli ho lze prohlásit za politického veterána.

MT Evaluation

You need a goal to be able to check your progress.

An example from the history:

- Manual judgement at Euratom (Ispra) of a Systran system (Russian→English) in 1972 revealed huge differences in judging; (Blanchon et al., 2004):
  - 1/5 (D–) for output quality (evaluated by teachers of language),
  - 4.5/5 (A+) for usability (evaluated by nuclear physicists).

- Metrics can drive the research for the topics they evaluate.
  - Some measured improvement required by sponsors: NIST MT Eval, DARPA, TC-STAR, EuroMatrix+.
  - BLEU has lead to a focus on phrase-based MT.
- Other metrics may similarly change the community's focus.

Our MT Task

We restrict the task of MT to the following conditions.

- Translate individual sentences, ignore larger context.
- No writers' ambitions, we prefer literal translation.
- No attempt at handling cultural differences.

Expected output quality:

1. Worth reading. (Not speaking the src. lang. I can sort of understand.)
2. Worth editing. (I can edit the MT output to obtain publishable text.)
3. Worth publishing, no editing needed.

- Neural MT and large data in 2018: Between 2 and 3.
- Cross-sentence relations are still a big problem.

Manual Evaluation

Black-box: Judging hypotheses produced by MT systems:

- Adequacy and fluency of whole sentences.
- Ranking of full sentences from several MT systems:

  Longer sentences hard to rank. Candidates incomparably poor.
- Ranking of constituents, i.e. parts of sentences:

  Tackles the issue of long sentences. Does not evaluate overall coherence.
- Comprehension test: Blind editing+correctness check.
- Task-based: Does MT output help as much as the original?

  Do I dress appropriately given a translated weather forecast?

Gray-box: Analyzing errors in systems' output.

Glass-box: System-dependent: Does this component work?

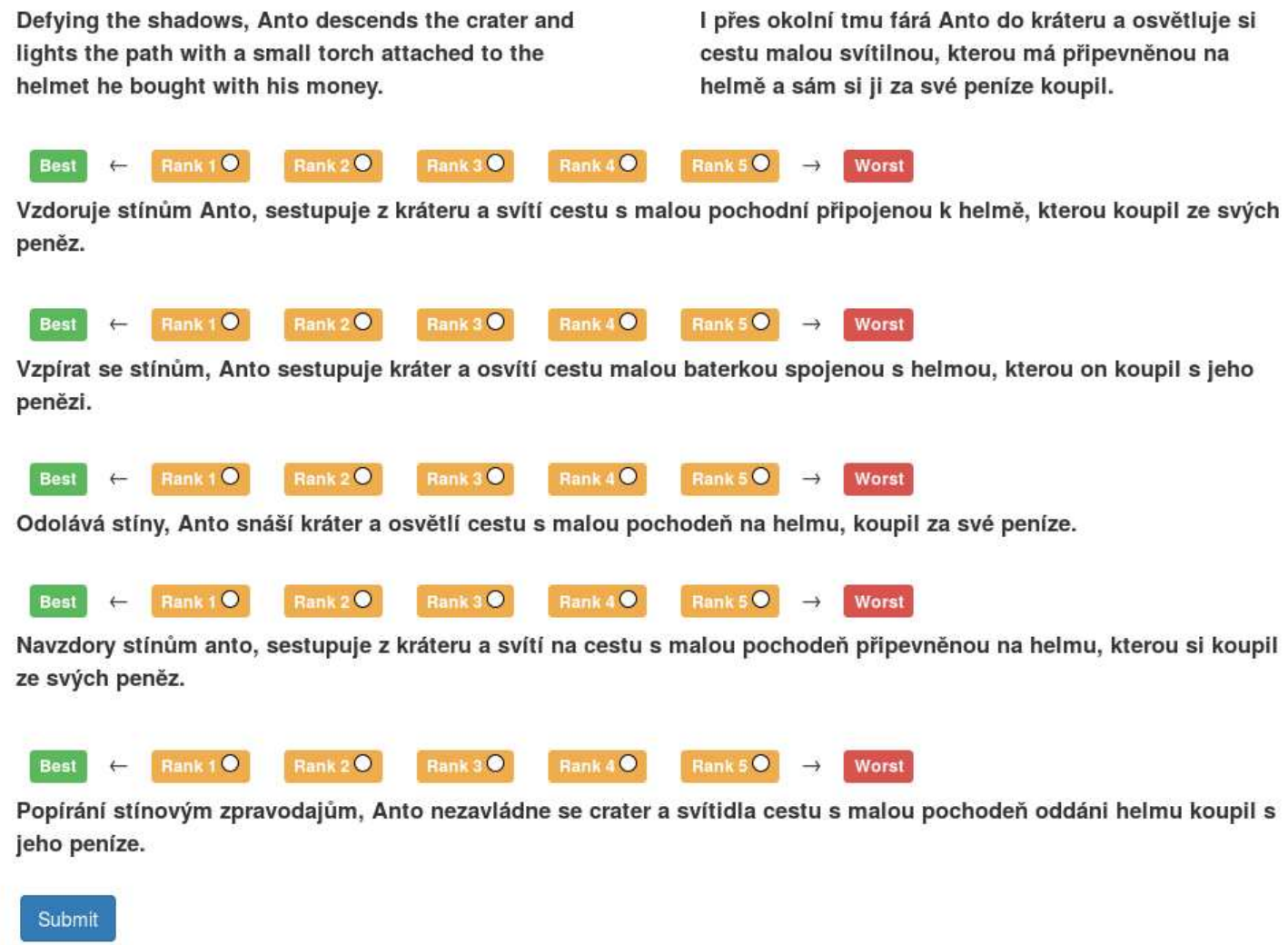
Ranking (of Constituents)

Source: Können die USA **ihre Besetzung aufrechterhalten**, wenn sie dem irakischen Volk nicht Nahrung, Gesundheitsfürsorge und andere grundlegende Dienstleistungen anbieten können?

Reference: Can the US **sustain its occupation** if it cannot provide food, health care, and other basic services to Iraq's people?

Translation	Rank
The United States can **maintain its employment** when it the Iraqi people not food, health care and other basic services on offer?.	○ 1 ◉ 2 ○ 3 ○ 4 ○ 5 Worst Best
The US can **maintain its occupation**, if they cannot offer the Iraqi people food, health care and other basic services?	○ 1 ○ 2 ○ 3 ○ 4 ◉ 5 Worst Best
Can the US **their occupation sustained** if it to the Iraqi people not food, health care and other basic services can offer?	○ 1 ○ 2 ◉ 3 ○ 4 ○ 5 Worst Best
Can the United States **maintain their occupation**, if the Iraqi people do not food, health care and other basic services can offer?	○ 1 ○ 2 ○ 3 ◉ 4 ○ 5 Worst Best
The United States **is maintained**, if the Iraqi people, not food, health care and other basic services can offer?	○ 1 ◉ 2 ○ 3 ○ 4 ○ 5 Worst Best
Annotator: ccb **Task:** WMT07 German-English News Corpus	
Instructions: Rank each constituent translation from Best to Worst relative to the other choices (ties are allowed). Grade **only the highlighted part** of each translation. *Please note that segments are selected automatically, and they should be taken as an approximate guide. They might include extra words on either end that are not in the actual alignment, or miss words.*	

Ranking Sentences (since 2013)



Ranking Sentences (Eye-Tracked)

Project suggestion: Analyze the recorded data: path patterns / errors in words.

Comprehension 1/2 (Blind Editing)

Original: They are often linked to other alterations sleep as nightmares, night terrors, the nocturnal enuresis (pee in bed) or the sleepwalking, but it is not always the case.

Edit:

They are often linked to other sleep disorders, such as nightmares, night terrors, the nocturnal enuresis (bedwetting) or sleepwalking, but this is not always the case.

Reset Edit

- (•) Edited.
- ( ) No corrections needed.
- ( ) Unable to correct.

Annotator: ccb **Task:** WMT09 Multisource-English News Editing

Instructions:
Correct the translation displayed, making it as fluent as possble. If no corrections are needed, select "No corrections needed." If you cannot understand the sentence well enough to correct it, select "Unable to correct."

Comprehension 2/2 (Judging)

Source: Au même moment, les gouvernements belges, hollandais et luxembourgeois ont en parti nationalisé le conglomérat européen financier, Fortis.
Les analystes de Barclays Capital ont déclaré que les négociations frénétiques de ce week end, conclues avec l'accord de sauvetage" semblent ne pas avoir réussi à faire revivre le marché".
Alors que la situation économique se détériorasse, la demande en matières premières, pétrole inclus, devrait se ralentir.
"la prospective d'équité globale, de taux d'intérêt et d'échange des marchés, est devenue incertaine" ont écrit les analystes de Deutsche Bank dans une lettre à leurs investisseurs."
"nous pensons que les matières premières ne pourront échapper à cette contagion.

Reference: Meanwhile, the Belgian, Dutch and Luxembourg governments partially nationalized the European financial conglomerate Fortis.
Analysts at Barclays Capital said the frantic weekend negotiations that led to the bailout agreement "appear to have failed to revive market sentiment."
As the economic situation deteriorates, the demand for commodities, including oil, is expected to slow down.
"The outlook for global equity, interest rate and exchange rate markets has become increasingly uncertain," analysts at Deutsche Bank wrote in a note to investors.
"We believe commodities will be unable to escape the contagion.

Translation	Verdict
While the economic situation is deteriorating, demand for commodities, including oil, should decrease.	◉ Yes ○ No
While the economic situation is deteriorating, the demand for raw materials, including oil, should slow down.	◉ Yes ○ No
Alors que the economic situation deteriorated, the request in rawmaterial enclosed, oil, would have to slow down.	○ Yes ◉ No
While the financial situation damaged itself, the first matters affected, oil included, should slow down themselves.	○ Yes ◉ No
While the economic situation is depressed, demand for raw materials, including oil, will be slow.	○ Yes ◉ No
Annotator: ccb **Task:** WMT09 French-English News Edit Acceptance	
Instructions: Indicate whether the edited translations represent fully fluent and meaning-equivalent alternatives to the reference sentence. The reference is shown with context, the actual sentence is **bold**.	

Task/Quiz-Based Evaluation

Moses 2007	Google 16.2.2010
Na provoz světla na roundabout, obrátit levice a projet ballymun. Otočit vlevo na křižovatce. ballymun / Collins Avenue Road Dcu je umístěna na Collins 500m na pravém boku Avenue.	Na semaforech na kruhový objezd, odbočit doleva a jet přes Ballymun. Odbočit vlevo na Collins Avenue / Ballymun silniční křižovatky. DCU se nachází na Collins Avenue 500 m na pravé straně.

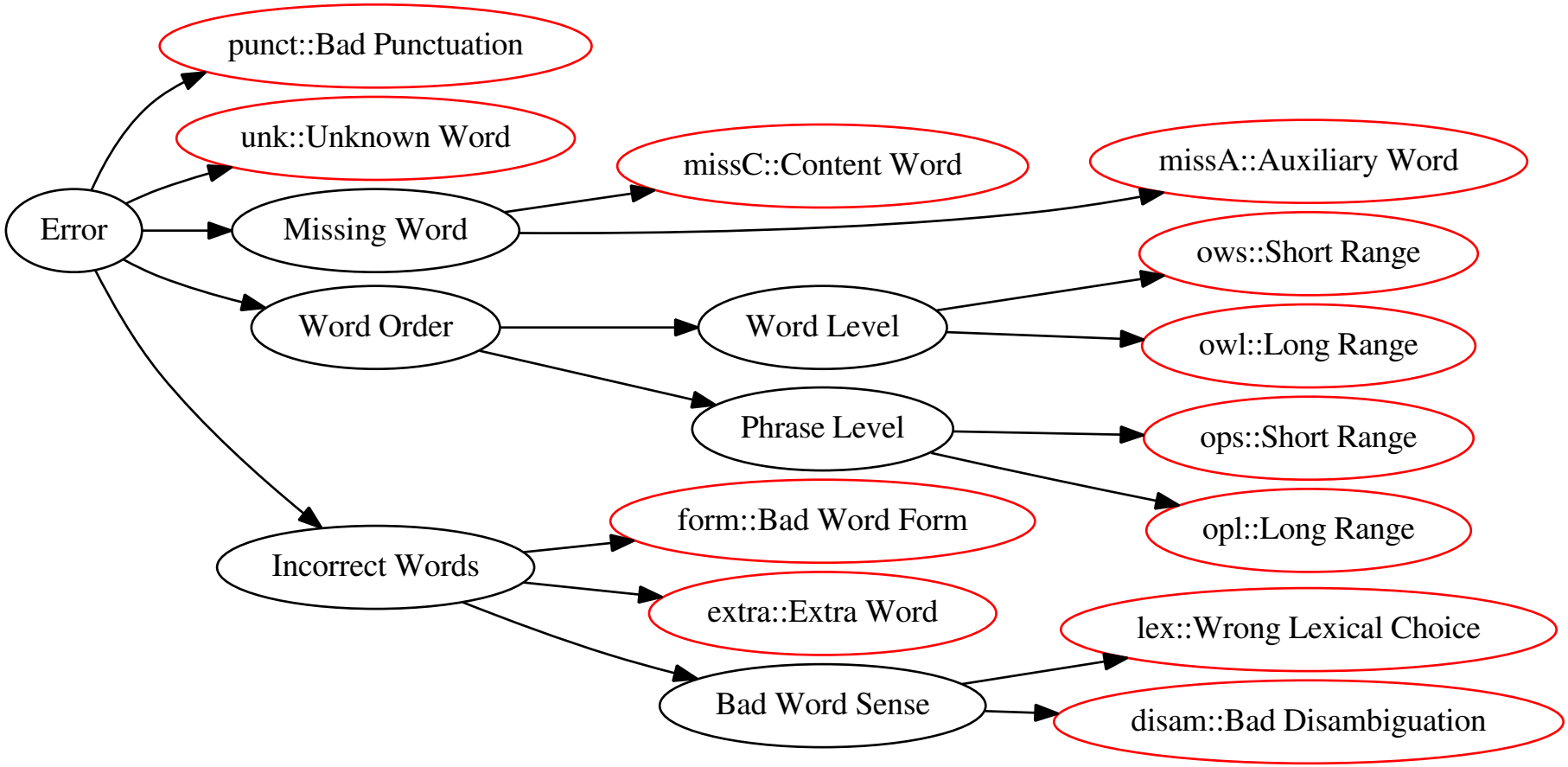
Zaškrtněte pravdivá tvrzení:

1. DCU leží na Collins Avenue.
2. V daném městě mají na kruhových objezdech zřejmě semafory.
3. Při příjezdu budete mít DCU po levé straně.

Original: At the traffic lights on the roundabout, turn left and drive through Ballymun. Turn left at the Collins Avenue/Ballymun Road crossroads. DCU is located on Collins Avenue 500m on the right hand side. Correct answer: yyn

Evaluation by Flagging Errors

Classification of MT errors, following Vilar et al. (2006).



Error Flagging Example

Src	Perhaps there are better times ahead.
Ref	Možná se tedy blýská na lepší časy.
	Možná, že extra::tam jsou lepší disam::krát lex::dopředu.
	Možná extra::tam jsou příhodnější časy vpředu.
missC::v_budoucnu	Možná form::je lepší časy.
	Možná jsou lepší časy lex::vpřed.

Results on WMT09 Dataset

	google	cu-bojar	pctrans	cu-tectomt	Total
Automatic: BLEU	13.59	14.24	9.42	7.29	–
Manual: Rank	0.66	0.61	0.67	0.48	–
disam	406	379	569	659	2013
lex	211	208	231	340	990
Total bad word sense	617	587	800	999	3003
missA	84	111	96	138	429
missC	72	199	42	108	421
Total missed words	156	310	138	246	850
form	783	735	762	713	2993
extra	381	313	353	394	1441
unk	51	53	56	97	257
Total serious errors	1988	1998	2109	2449	8544
ows	117	100	157	155	529
punct	115	117	150	192	574
...	...	...	...	...	...
tokenization	7	12	10	6	35
Total errors	2319	2354	2536	2895	10104

Contradictions in (Manual) Eval

Results for WMT10 Systems:

Evaluation Method	Google	CU-Bojar	PC Translator	TectoMT
$\geq$ others (WMT10 official)	**70.4**	65.6	62.1	60.1
$>$ others	49.1	45.0	**49.4**	44.1
Edits deemed acceptable [%]	**55**	40	43	34
Quiz-based evaluation [%]	80.3	75.9	80.0	**81.5**
Automatic: BLEU	**0.16**	0.15	0.10	0.12
Automatic: NIST	**5.46**	5.30	4.44	5.10

. . . each technique provides a different picture.

Problems of Manual Evaluation

- Expensive in terms of time/money.
- Subjective (some judges are more careful/better at guessing).
- Not quite consistent judgments from different people.
- Not quite consistent judgments from a single person!
- Not reproducible (too easy to solve a task for the second time).
- Experiment design is critical!

- Black-box evaluation important for users/sponsors.
- Gray/Glass-box evaluation important for the developers.

Automatic Evaluation

- Comparing MT output to reference translation.

  There are **hundreds of thousands** equally correct translations.

  See Bojar et al. (2013a) and Dreyer and Marcu (2012)

- Fast and cheap.
- Deterministic, replicable.
- Allows automatic model optimization.

- Usually good for checking progress.
- Usually bad for comparing systems of different types.

BLEU (Papineni et al., 2002)

- Based on **geometric** mean of n-gram **precision**.

$\approx$ ratio of 1- to 4-grams of hypothesis confirmed by a ref. translation

Src	The legislators hope that it will be approved in the next few days .	Confirmed
Ref	Zákonodárci doufají , že bude schválen v příštích několika dnech .	1 2 3 4
Moses	**Zákonodárci doufají , že bude schválen v** nejbližších dnech .	9 7 5 4
TectoMT	**Zákonodárci doufají , že bude** schváleno další páru volna .	6 4 3 2
Google	Zákonodárci naději **, že bude schválen v** několika příštích dnů .	9 4 3 2
PC Tr.	Zákonodárci doufají že to bude schválený v nejbližších dnech .	7 2 0 0

n-grams confirmed: none, unigram, bigram, *trigram*, **fourgram**

E.g. Moses produced 10 unigrams (9 confirmed), 9 bigrams (7 confirmed), . . .

$$\text{BLEU} = \text{BP} \cdot \exp\left(\frac{1}{4}\log\left(\frac{9}{10}\right) + \frac{1}{4}\log\left(\frac{7}{9}\right) + \frac{1}{4}\log\left(\frac{5}{8}\right) + \frac{1}{4}\log\left(\frac{4}{7}\right)\right)$$

BP is "brevity penalty"; $\frac{1}{4}$ are uniform weights, the "denominator" equivalent for $\sqrt[4]{\cdot}$ in geometric mean in the log domain.

BLEU: Avoiding Cheating

- Confirmed counts "clipped" to avoid overgeneration.
- "Brevity penalty" applied to avoid too short output:

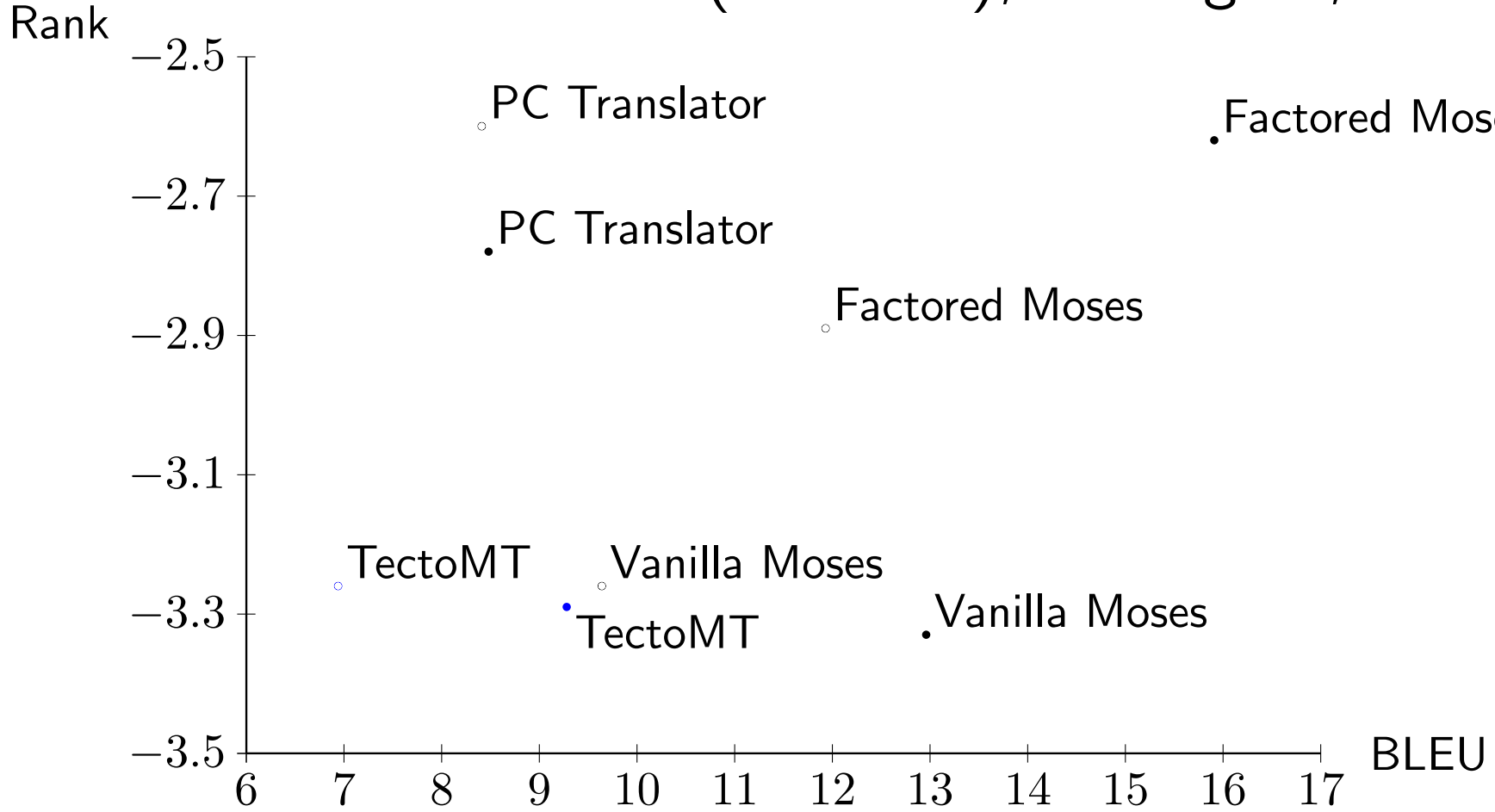
$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases}$$

Ref 1:	The cat is on the mat .
Ref 2:	There is a cat on the mat .
Candidate:	The the the the the the the .
	$\Rightarrow$ Clipping: only $\frac{3}{8}$ unigrams confirmed.
Candidate:	The the .
	$\Rightarrow$ $\frac{3}{3}$ unigrams confirmed but the output is too short.
	$\Rightarrow$ BP $= e^{1-7/3} = 0.26$ strikes.

The candidate length c and "effective" ref. length r calculated over the whole test set.

Correlation with Human Judgments

BLEU scores vs. human rank (WMT08), the higher, the better:



⇒ PC Translator nearly won Rank but nearly lost in BLEU.

Technical Problems of BLEU

BLEU scores are not comparable:

- across languages.
- on different test sets.
- with different number of reference translations.
- with different implementations of the evaluation tool.

- There are different definitions of "reference length":
  Papineni et al. (2002) not specific. One can choose the shortest, longest, average, closest (the smaller or the larger!).
- Very sensitive to tokenization:
  Beware esp. of malformed tokenization of Czech by foreign tools.

Fundamenal Problems of BLEU

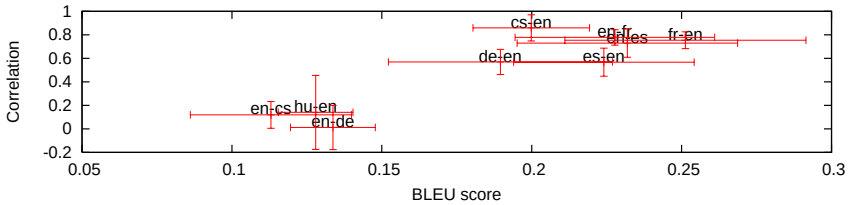
- BLEU overly sensitive to word forms and sequences of tokens.

Confirmed by Ref	Contains Error Flags	1-grams	2-grams	3-grams	4-grams
Yes	Yes	6.34%	1.58%	0.55%	0.29%
Yes	No	36.93%	13.68%	5.87%	2.69%
No	Yes	22.33%	41.83%	54.64%	63.88%
No	No	**34.40%**	**42.91%**	**38.94%**	**33.14%**
Total n-grams		35 531	33 891	32 251	30 611

30–40% of tokens not confirmed by reference but without errors.

⇒ Enough space for MT systems to differ unnoticed.

⇒ Low BLEU scores correlate even less:



Fix 1: Coarser Metric (SemPOS)

Instead of giving credit for 1, 3, 5 and 8 four-, three-, bi- and unigrams, overestimating cu-bojar:

SRC	Congress yields: US government can pump 700 billion dollars into banks
REF	kongres ustoupil : vláda usa může do bank napumpovat 700 miliard dolarů
cu-bojar	kongres výnosy : vláda usa může čerpadlo 700 miliard dolarů v bankách
pctrans	kongres vynáší : us vláda může čerpat 700 miliardu dolarů do bank

E.g. SemPOS (Kos and Bojar, 2009) gives credit for 8 lemmas:

REF	kongres ustoupit : vláda usa banka napumpovat 700 miliarda dolar
cu-bojar	kongres výnos : vláda usa moci čerpadlo 700 miliarda dolar banka
pctrans	kongres vynášet : us vláda čerpat 700 miliarda dolar banka

And correlates better with human judgments:

Metric	Sentence-level Correlation	System-level Correlation
SemPOS	0.21±0.57	0.81±0.18
BLEU	**0.03±0.63** !!	0.40±0.23

Fix 2: More References

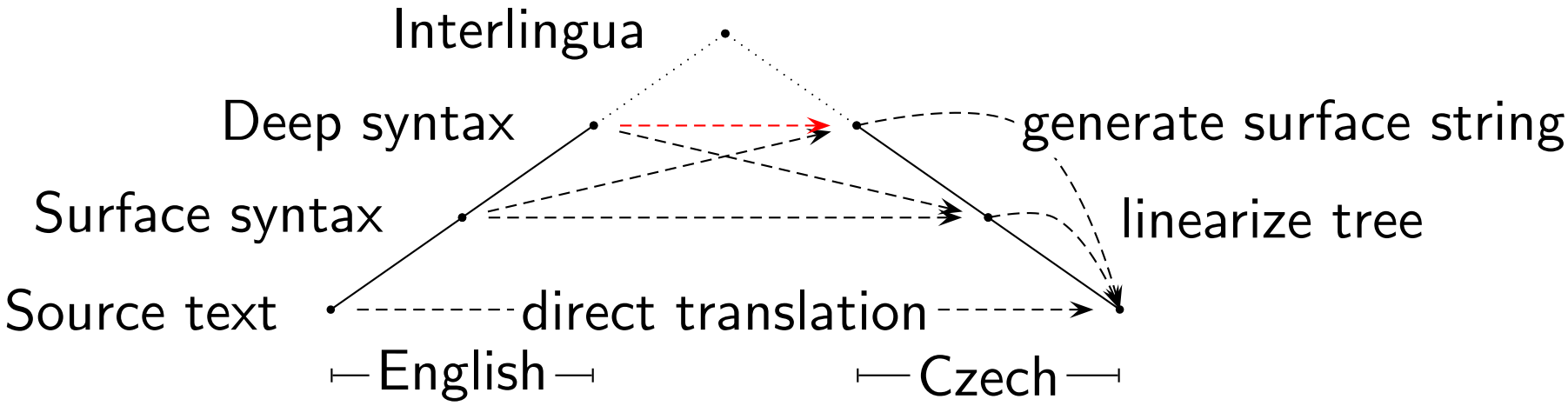
Bojar et al. (2013b) use post-editing to obtain more refs.

- 100 sents with 6-7 post-edited refs as good as 3000 independent refs.

Bojar et al. (2013a) create many reference translations.

- Avg. 123k references per one input sentence.
- Annotation was restricted to 2 hours/sentence.

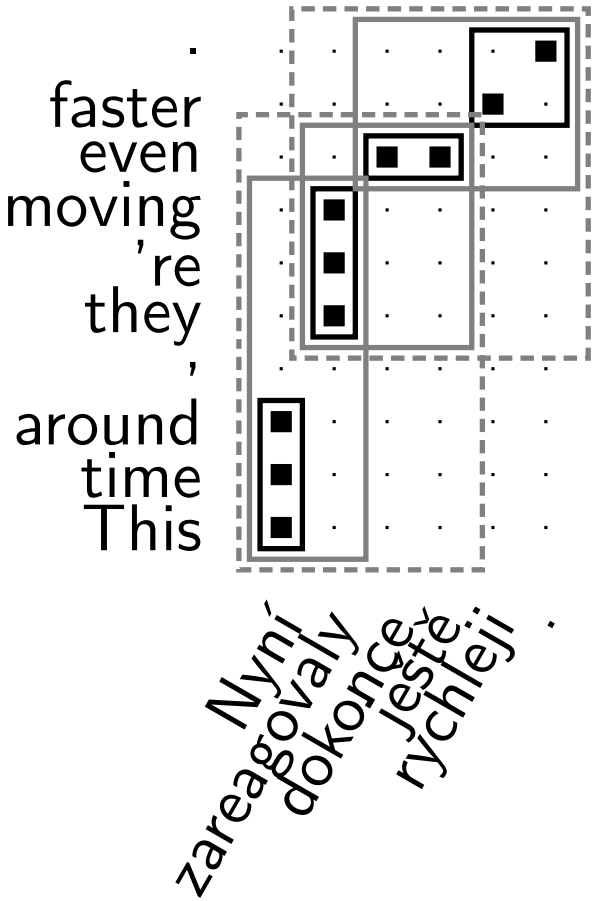
Approaches to Machine Translation



- The deeper analysis, the easier the transfer should be.
- A hypothetical interlingua captures pure meaning.

- Rule-based systems implemented by linguists-programmers.
- Statistical systems learn automatically from data.
  - "Classical SMT" works with translation units, e.g. "phrases".
  - Neural systems use deep learning, more end-to-end.

Phrase-Based MT Overview



This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
...	=	...
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychleji
...	=	...

Phrase-based MT: choose such segmentation of input string and such phrase "replacements" to make the output sequence "coherent" (3-grams most probable).

More next week.

Mining the Web

Goal: Given two languages, find parallel texts.

- Hervé Saint-Amand's master's thesis (Saarbrücken).
  - Search for pages in English containing the word *česky*.
- Bitextor: Esplà-Gomis and Forcada (2010)
- PANACEA tools (`http://myexperiment.elda.org/workflows/7`)
- Students' project ParaSite: proof of concept, fixes needed.
- Recent: **ParaCrawl:** `http://paracrawl.eu/`

Quasi-comparable sources (incl. Wikipedia):

- Texts on the same topic but written independently.
- Can hope to find parallel sentences but no longer segments.
- Technique: "lightly supervised training" (Schwenk, 2008).

Document Alignment

Goal: Given bag of texts in two languages, find pairs.

- A project at FJFI (Jahoda et al., 2007)
- A project at MFF: (Klempová et al., 2009)
  - Evaluation suggested that the first part is tricky: finding source URLs.
- Václav Novák (ÚFAL): aligning subtitles.
  - Not generic enough: focus on named entities at the beg. and end only.
- ParaSite: probably good, re-evaluation would be useful.
  - Problem: Based on libraries with conflicting licenses (GPL 2.0 vs 3.0).
- Parallel paragraphs from CommonCrawl (Kúdela et al., 2017).
- WMT16 Shared Task on Bilingual Document Alignment:

  `http://www.statmt.org/wmt16/bilingual-task.html`
- WMT18 Shared Task on Corpus Filtering:

  `http://www.statmt.org/wmt18/parallel-corpus-filtering.html`

Sentence Alignment

Goal: Given a text in two languages, align sentences.

Assume: Sentences hardly ever reordered.

- Classical algorithm: Gale and Church (1993).
  - Based on similar character length of aligned sentences, no words examined.
  - Dynamic-programming search for the best alignment.
  - Allows 0 to 2 sentences in a group: 0-1, 1-0, 1-1, 2-1, 1-2, 2-2.
- Several algorithms for English-Czech evaluated by Rosen (2005).
  - Nearly perfect alignment possible by a combination of aligners.
- The "standard tool": Hunalign (Varga et al., 2005).
- Another option: Gargantua (Braune and Fraser, 2010).

Word Alignment

Goal: Given a sentence in two languages, align words (tokens).

State of the art: GIZA++ (Och and Ney, 2000):

- Unsupervised, only sentence-parallel texts needed.
- Word alignments formally restricted to a function:

$$\text{src token} \mapsto \text{tgt token or NULL}$$

- A cascade of models refining the probability distribution:
  - IBM1: only lexical probabilities: $P(kočka = cat)$
  - IBM3: adds fertility: 1 word generates several others
  - IBM4/HMM: to account for relative reordering
- Only many-to-one links created $\Rightarrow$ used twice, in both directions.

IBM Model 1

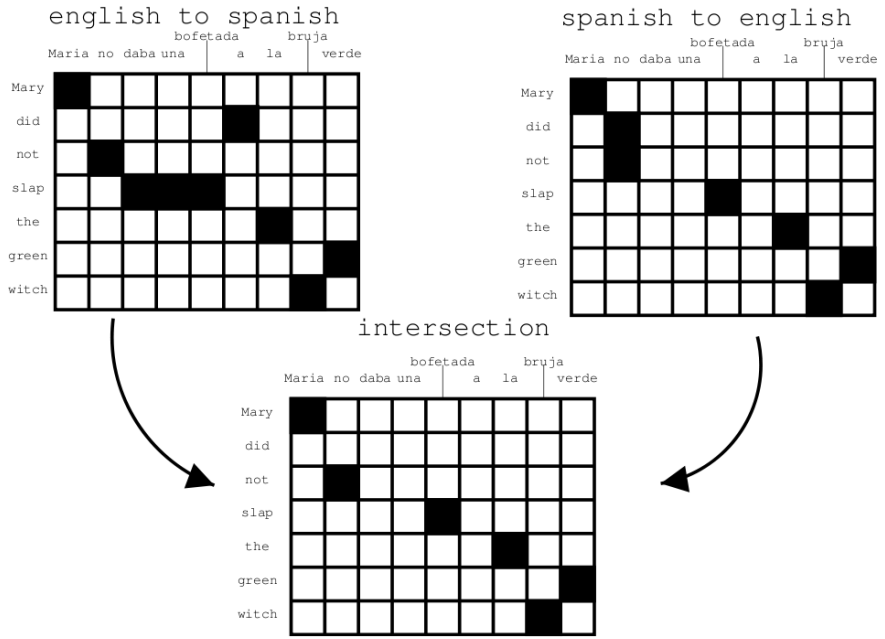
Lexical probabilities:

- Disregard the position of words in sentences.
- Estimated using Expectation-Maximization Loop.

. . . see the slides by:

- Aleš Tamchyna.
  `http://www.statmt.org/mtm15` → Programme → Tuesday Lecture
- Patrick Lambert (originally Philipp Koehn)
  `http://lium3.univ-lemans.fr/mtmarathon2010/lectures/02-wordalignment.pdf`

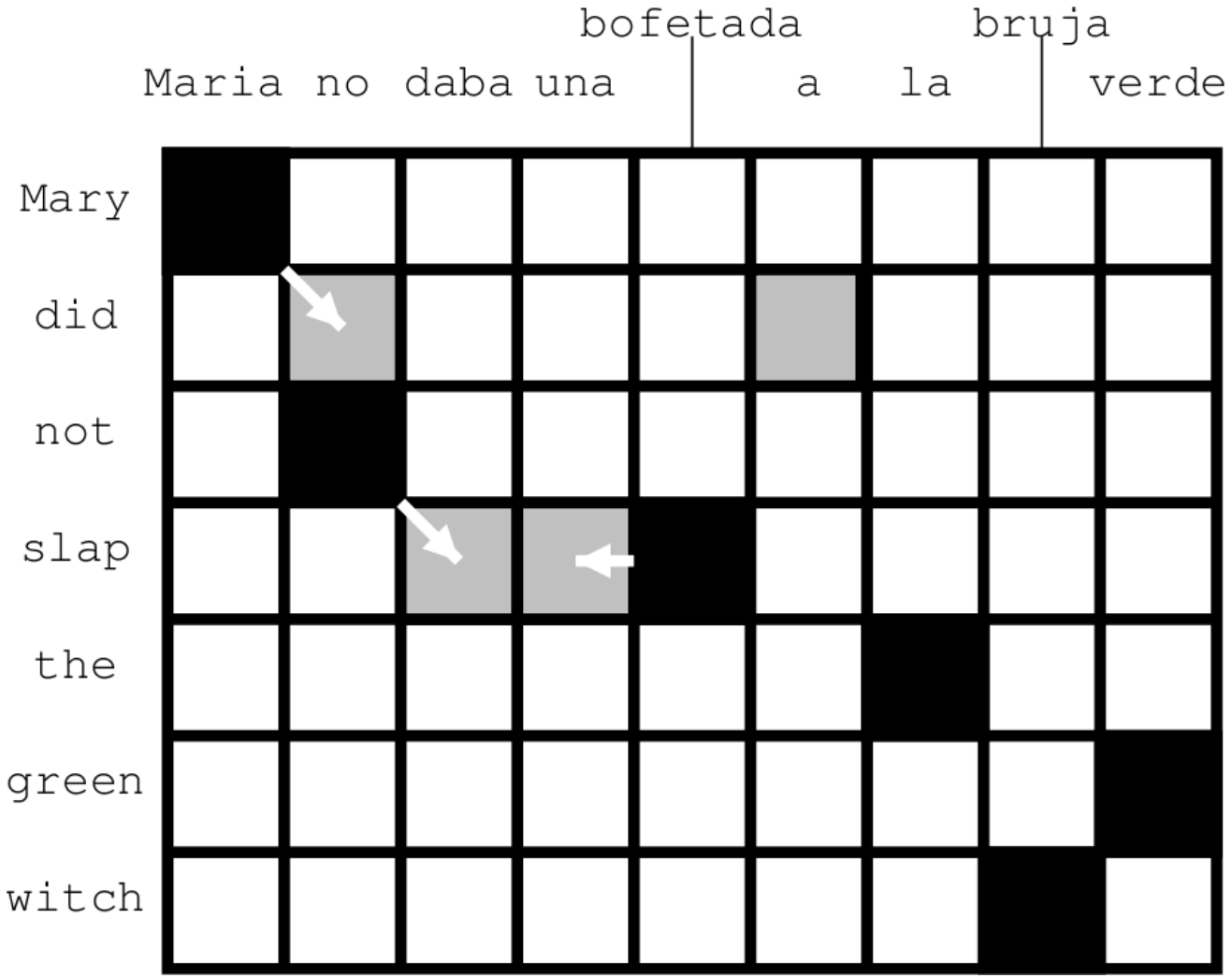
Symmetrization



"Symmetrization" of two GIZA++ runs:

- intersection: high precision, too low recall.
- popular: heuristical (something between intersection and union).
- minimum-weight edge cover (Matusov et al., 2004).

Popular Symmetrization Heuristic



Extend intersection by neighbours of the union (Och and Ney, 2003).

Troubles with Word Alignment

- Humans have troubles aligning word for word.
  - Mismatch in alignments points 9–18%. (Bojar and Prokopová, 2006)

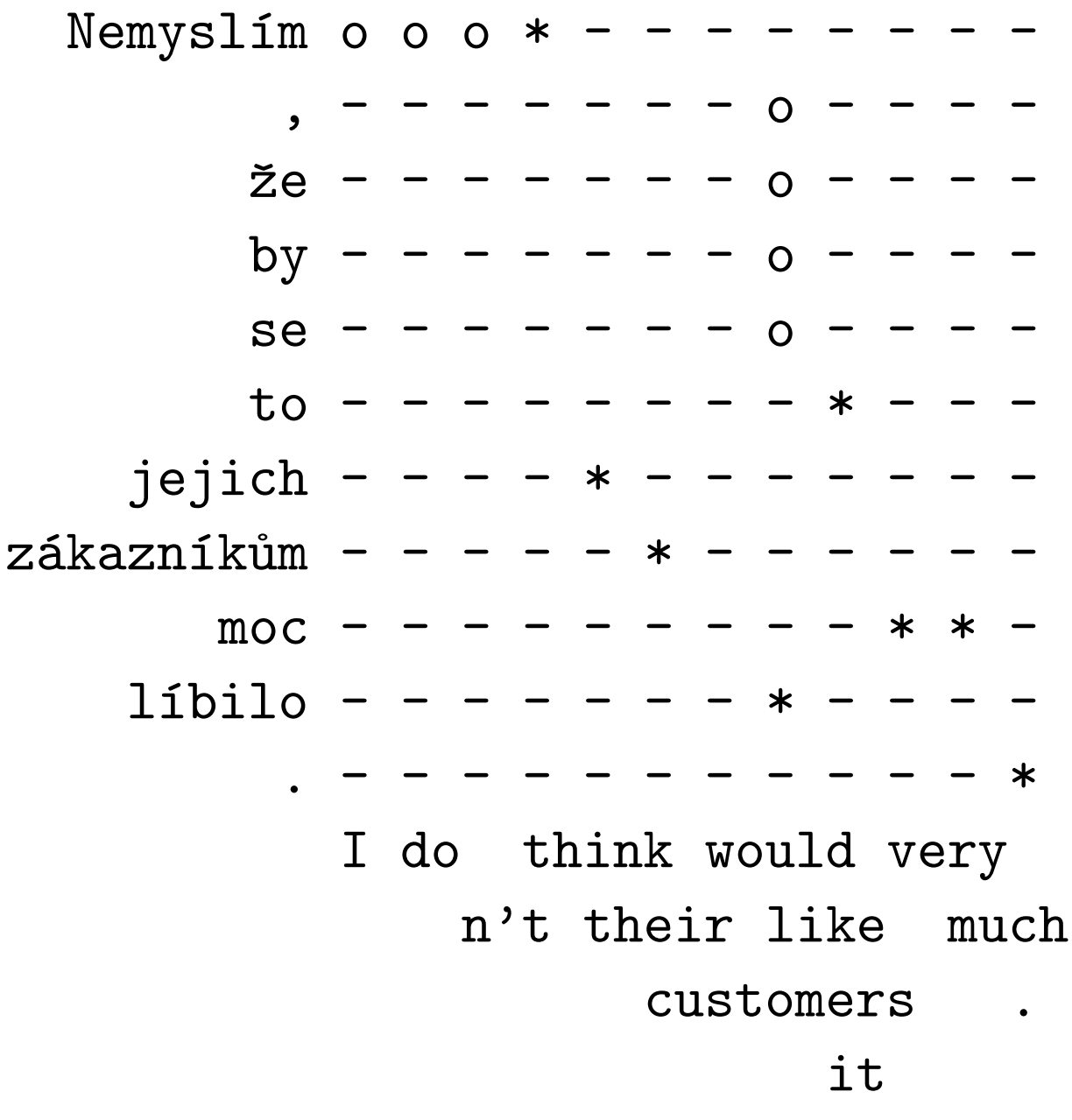
Top Problematic Words

English		Czech	
361	to	319	,
259	the	271	se
159	of	146	v
143	a	112	na
124	,	74	o
107	be	61	že
99	it	55	.
95	that	47	a

Top Problematic Parts of Speech

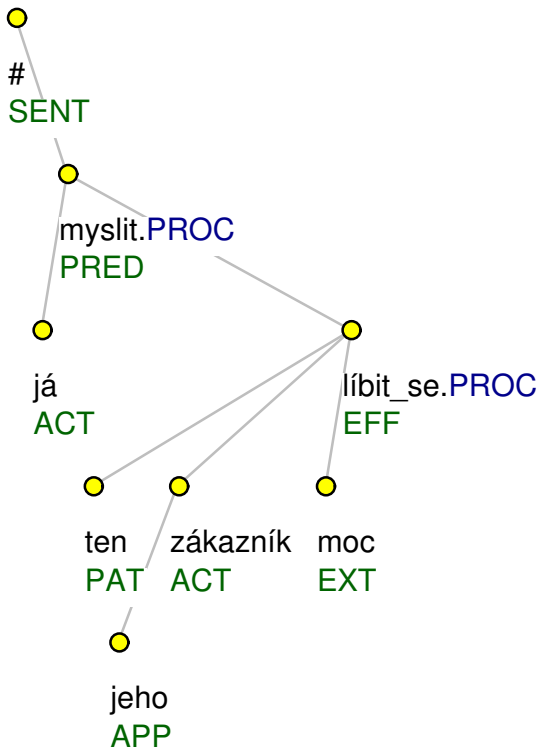
English		Czech	
679	IN	1348	N
519	DT	1283	V
510	NN	661	R
386	PRP	505	P
361	TO	448	Z
327	VB	398	A
310	JJ	280	D
245	RB	192	J

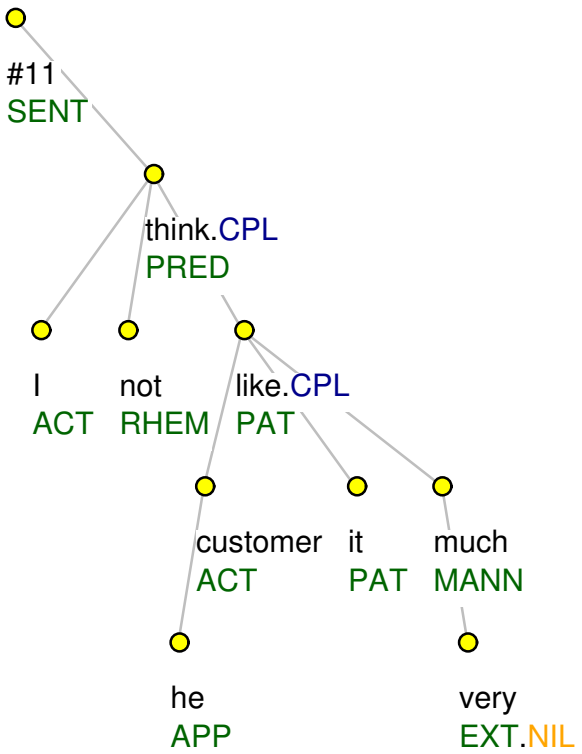
A Czech-English Example



T-Layer to the Rescue

- Only content-bearing words have a node.
- Auxiliary words hidden, dropped pronouns added.





(já) Nemyslím **, že by se** to jejich zákazníkům moc líbilo **.**

I **do** n't think their customers **would** like it very much **.**

Tectogrammatical Alignment

- Mareček et al. (2008) align t-nodes, not words.
  $\Rightarrow$ Auxiliary words do not clutter the task.
- Improves human agreement from 91% to 94.7%.
- Application to phrase-based MT: (Mareček, 2009)
  - Improved alignment error rate on content words.
  - Minor improvements in BLEU when combined with GIZA++.
- Main use: Extraction of t-lemma dictionaries for e.g. TectoMT.

Main disadvantage:

- Language-dependent.
- Heavy use of tools (tagging, parsing, deep parsing).

Ultimate Goal of Classical SMT

Find **minimum translation units** $\sim$ graph partitions:

- such that they are frequent across many sentence pairs.
- without imposing (too hard) constraints on reordering.

Translate by:

- decomposing input into these units,
- translating units independently,
- finding the best combination of the units.

Available data: Word co-occurrence statistics:

- In large monolingual data (usually up to 10^9 words).
- In smaller parallel data (up to 10^7 words per language).
- Optional automatic rich linguistic annotation.

Summary of MT Class 1

- Why is MT difficult (primarily linguistic point of view).
- MT evaluation.
  - Manual, automatic, different metrics different results.
  - Including BLEU and issues with BLEU.
- Phrase-based MT on one slide.
- Getting parallel data.
  - Including EM for word alignment.

References

Hervé Blanchon, Christian Boitet, and Laurent Besacier. 2004. Spoken Dialogue Translation Systems Evaluation: Results, New Trends, Problems and Proposals. In Proceedings of International Conference on Spoken Language Processing ICSLP 2004, Jeju Island, Korea, October.

Ondřej Bojar and Magdalena Prokopová. 2006. Czech-English Word Alignment. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), pages 1236–1239. ELRA, May.

Ondřej Bojar, Matouš Macháček, Aleš Tamchyna, and Daniel Zeman. 2013a. Scratching the Surface of Possible Translations. In Proc. of TSD 2013, Lecture Notes in Artificial Intelligence, Berlin / Heidelberg. Západočeská univerzita v Plzni, Springer Verlag.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013b. Findings of the 2013 Workshop on Statistical Machine Translation. In Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Fabienne Braune and Alexander Fraser. 2010. Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora. In Coling 2010: Posters, pages 81–89, Beijing, China, August. Coling 2010 Organizing Committee.

Markus Dreyer and Daniel Marcu. 2012. HyTER: Meaning-Equivalent Semantics for Translation Evaluation. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 162–171, Montréal, Canada, June. Association for Computational Linguistics.

Miquel Esplà-Gomis and Mikel L. Forcada. 2010. Combining Content-Based and URL-Based Heuristics to Harvest Aligned Bitexts from Multilingual Sites with Bitextor. In

References

Prague Bulletin of Mathematical Linguistics - Special Issue on Open Source Machine Translation Tools, number 93 in Prague Bulletin of Mathematical Linguistics. Charles University, January.

William A. Gale and Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics, 19(1):75–102.

František Jahoda, Vladimír Jarý, Jan Kobera, Jaromír Müller, and Václav Müller. 2007. Generování paralelních textů z webu. Student project at POPJ2 (Počítače a přirozený jazyk) seminar at FJFI, Czech Technical University.

Hana Klempová, Michal Novák, Peter Fabian, Jan Ehrenberger, and Ondřej Bojar. 2009. Získávání paralelních textů z webu. In ITAT 2009 Information Technologies – Applications and Theory, September.

Kamil Kos and Ondřej Bojar. 2009. Evaluation of Machine Translation Metrics for Czech as the Target Language. Prague Bulletin of Mathematical Linguistics, 92:135–147.

Jakub Kúdela, Irena Holubová, and Ondřej Bojar. 2017. Extracting parallel paragraphs from common crawl. The Prague Bulletin of Mathematical Linguistics, (107):36–59.

David Mareček, Zdeněk Žabokrtský, and Václav Novák. 2008. Automatic Alignment of Czech and English Deep Syntactic Dependency Trees. In Proceedings of EAMT 2008, Hamburg, Germany.

David Mareček. 2009. Using Tectogrammatical Alignment in Phrase-Based Machine Translation. In Jana Šafránková, editor, WDS'04 Proceedings of Contributed Papers, Prague. Charles University, Matfyzpress.

E. Matusov, R. Zens, and H. Ney. 2004. Symmetric Word Alignments for Statistical Machine Translation. In Proceedings of COLING 2004, pages 219–225, Geneva, Switzerland, August 23–27.

Franz Josef Och and Hermann Ney. 2000. A Comparison of Alignment Models for Statistical Machine Translation. In Proceedings of the 17th conference on Computational linguistics, pages 1086–1090. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models.

References

Computational Linguistics, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In ACL 2002, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania.

Alexandr Rosen. 2005. In Search of Best Method for Sentence Alignment in Parallel Texts. In R. GarabĂk, editor, Computer Treatment of Slavic and East European Languages, pages 174–185. Veda, Bratislava.

Holger Schwenk. 2008. Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation. In International Workshop on Spoken Language Translation, pages 182–189.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In Proceedings of the Recent Advances in Natural Language Processing RANLP 2005, pages 590–596, Borovets, Bulgaria.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error Analysis of Machine Translation Output. In International Conference on Language Resources and Evaluation, pages 697–702, Genoa, Italy, May.