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

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

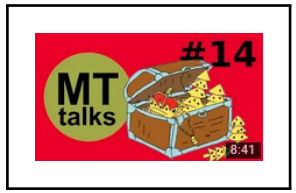
Outline of Lectures on MT

1. Introduction.
   - Why is MT difficult.
   - MT evaluation.
   - Approaches to MT.
   - Document, sentence and esp. word alignment.
   - Classical Statistical Machine Translation.
     - Phrase-Based MT.
2. Neural Machine Translation.
   - Neural MT: Sequence-to-sequence, attention, self-attentive.
   - Sentence representations.
   - Role of Linguistic Features in MT.

Supplementary Materials

Videolectures & Wiki:

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



NPFL087 Class on Machine Translation:

https://ufal.mff.cuni.cz/courses/npfl087

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 (1/2)

The plant is next to the bank.

Ambiguity and Word Senses (1/2)

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 (1/2)

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 (1/2)

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 (2/2)

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.
ÚFAL 2011a	Z jednoho kohoutku a stroj vydá složenky s číslem.
ÚFAL 2011b	Jeden úder a stroj vydá složenky s číslem.
Google 2011	Jedním klepnutím a stroj problémy skluzu s číslem.
Google 2017–8	Jeden kohoutek a zařízení vydává skluzu s číslem.
Google 2020	Jedním klepnutím a stroj vydá doklad s číslem.
ÚFAL 2018–20	Jedno klepnutí a přístroj vydá lístek 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 Correctly

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.

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 obdobná
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 podobná
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 veterán
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?

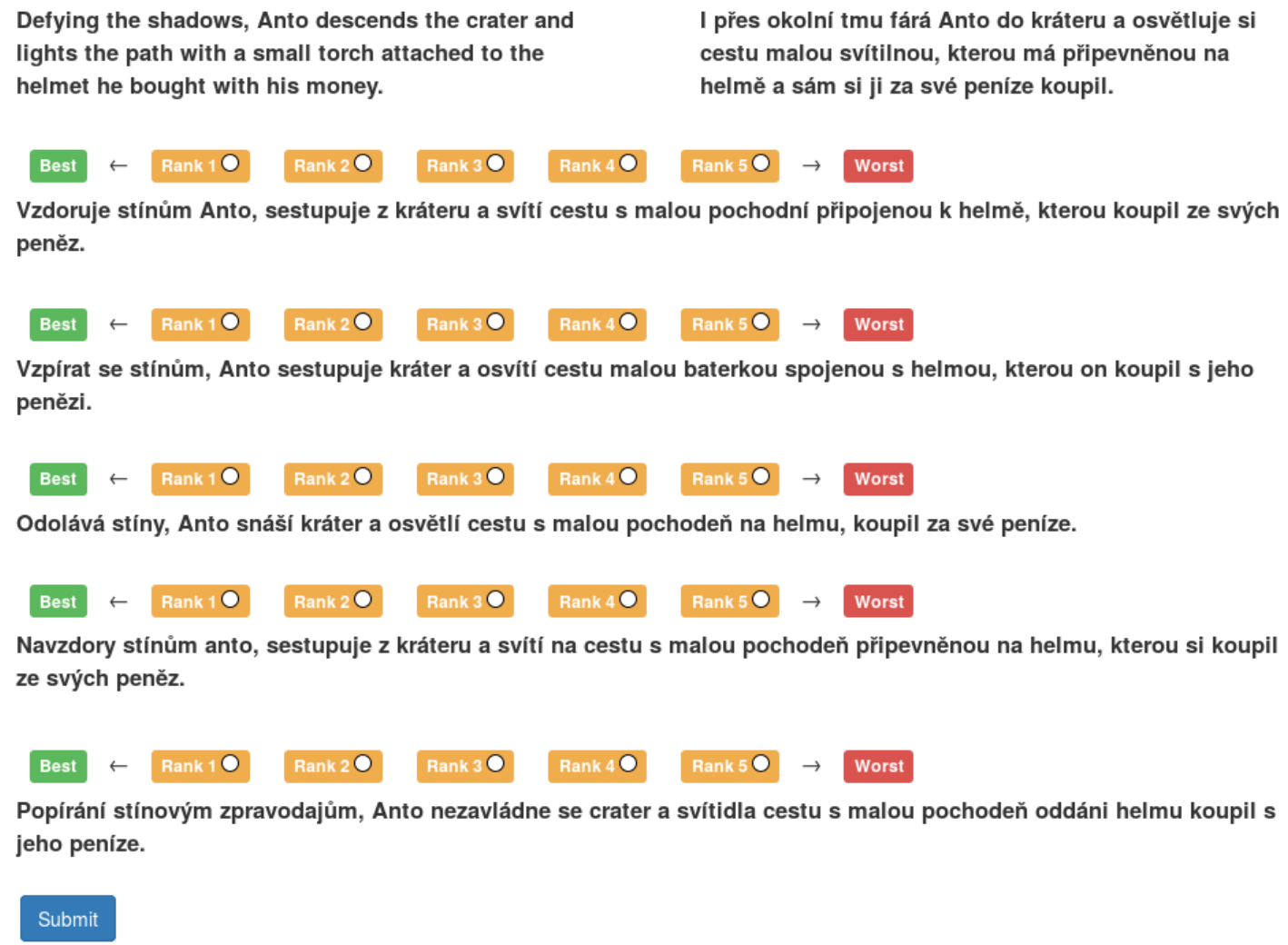
Ranking (of Constituents)

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

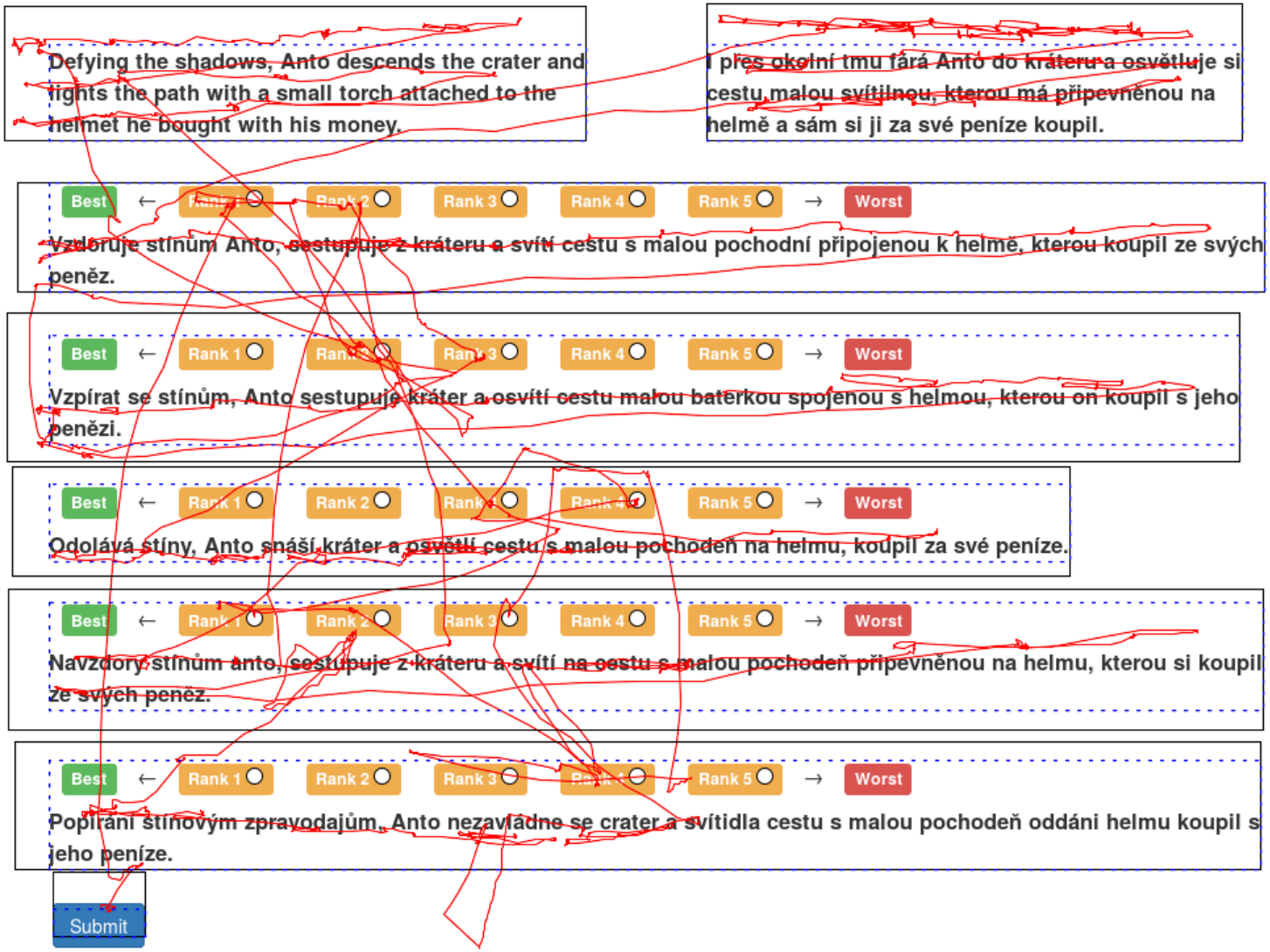
Reference: Can the US **sustain its occupation** if it cannot provide food, health care, and other basic services to Iraq's people?

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)

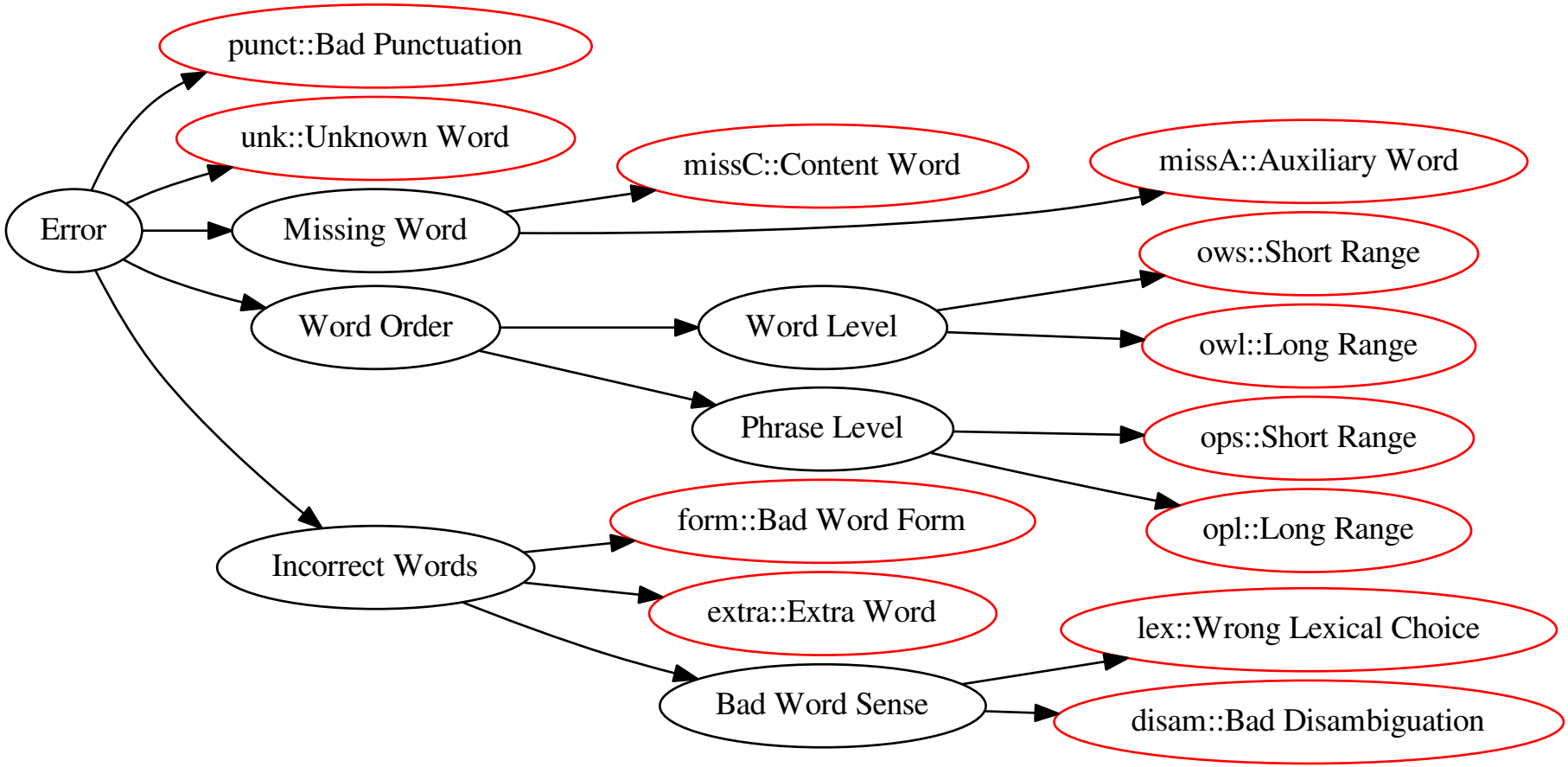
Source: Au même moment, les gouvernements belges, hollandais et luxembourgeois ont en parti nationalisé le conglomérat européen financier, Fortis.
Les analystes de Barclays Capital ont déclaré que les négociations frénétiques de ce week end, conclues avec l'accord de sauvetage" semblent ne pas avoir 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**.	

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. (2013) 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:

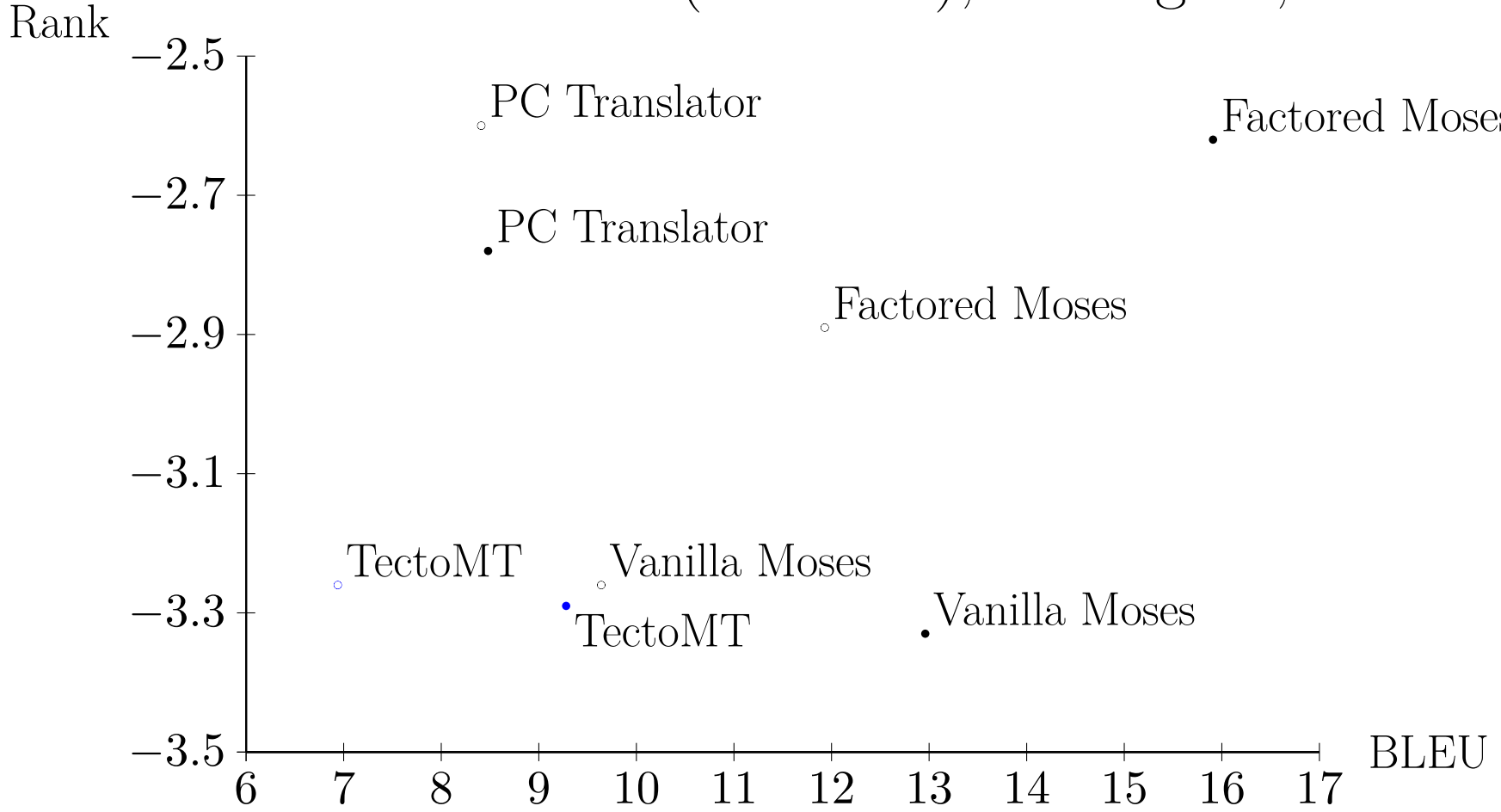
$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{1-r/c} & \text{if } c \leq r \end{cases}$$

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.
	$\Rightarrow \text{BP} = e^{1-7/3} = 0.26$ 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.

Problems of BLEU

Technical: BLEU scores are not comparable:

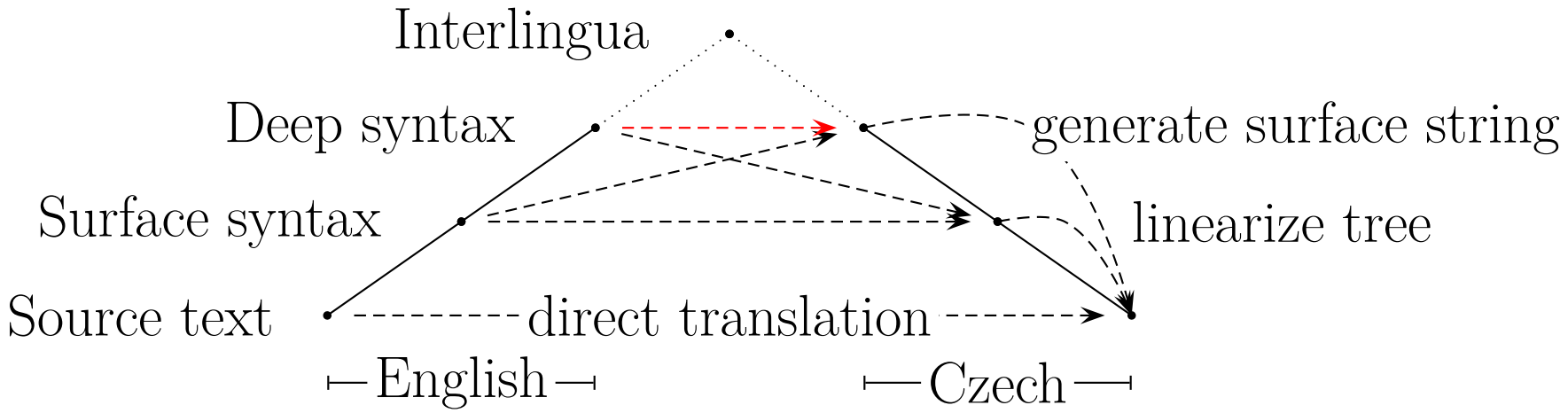
- across languages.
- on different test sets.
- with different number of reference translations.
- with different implementations of the evaluation tool.

Fundamental: BLEU overly sensitive to token forms and sequences.

⇒ Use coarser units

⇒ Use more references.

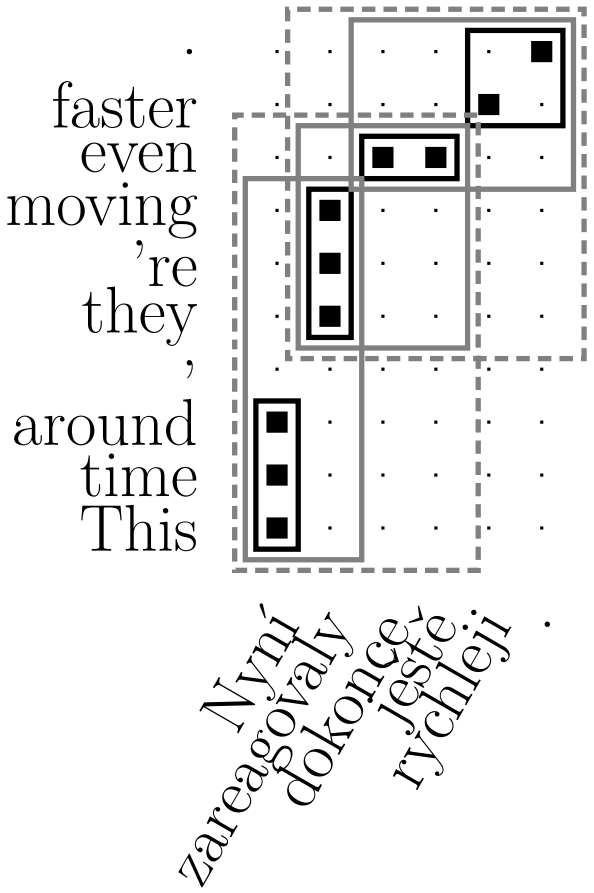
Approaches to Machine Translation



- 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).

Data Acquisition Pipeline

- Mine the Web.
  - Given two languages, find parallel texts.
  - Multiple tools, esp. Bitextor.
- Align documents.
  - Multiple tools, e.g. Paracrawl http://paracrawl.eu/
  - Parallel paragraphs from CommonCrawl (Kúdela et al., 2017).
- Align sentences.
  - Classical algorithm: Gale and Church (1993)
  - Standard tool: Hunalign (Varga et al., 2005).
  - Illustration: MT Talk #7: https://youtu.be/_4lnyoC3mtQ
- Align words.

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: $P(\text{kočka} = \text{cat})$
  - 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:

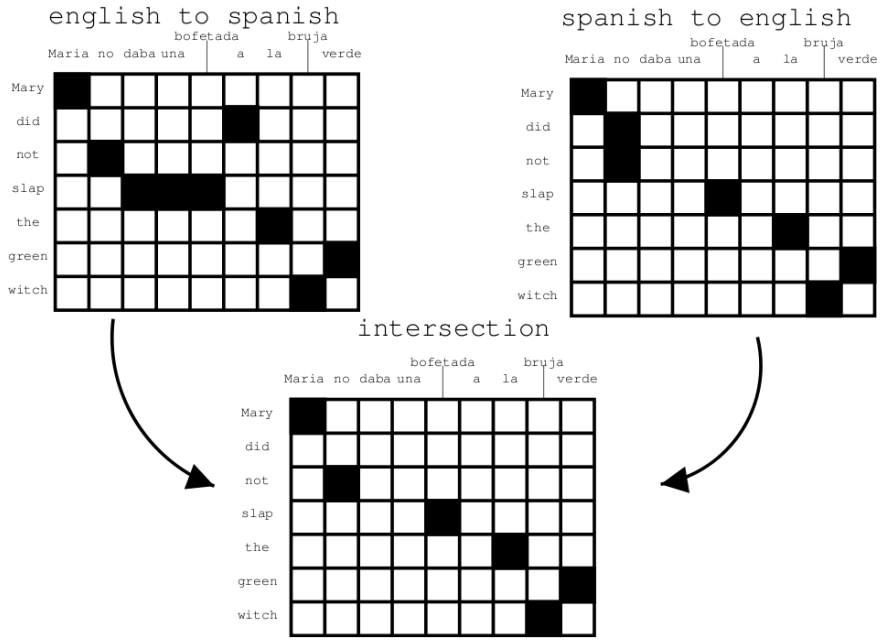
- Disregard the position of words in sentences.
- Estimated using Expectation-Maximization Loop.

See the slides by Philipp Koehn for:

- Formulas of both expectation and maximization step.
- The trick in expectation step, swapping sum and product by rearranging the sum.
- Pseudocode.

Illustration: MT Talk #8 (https://youtu.be/mqyMDLu5JPw)

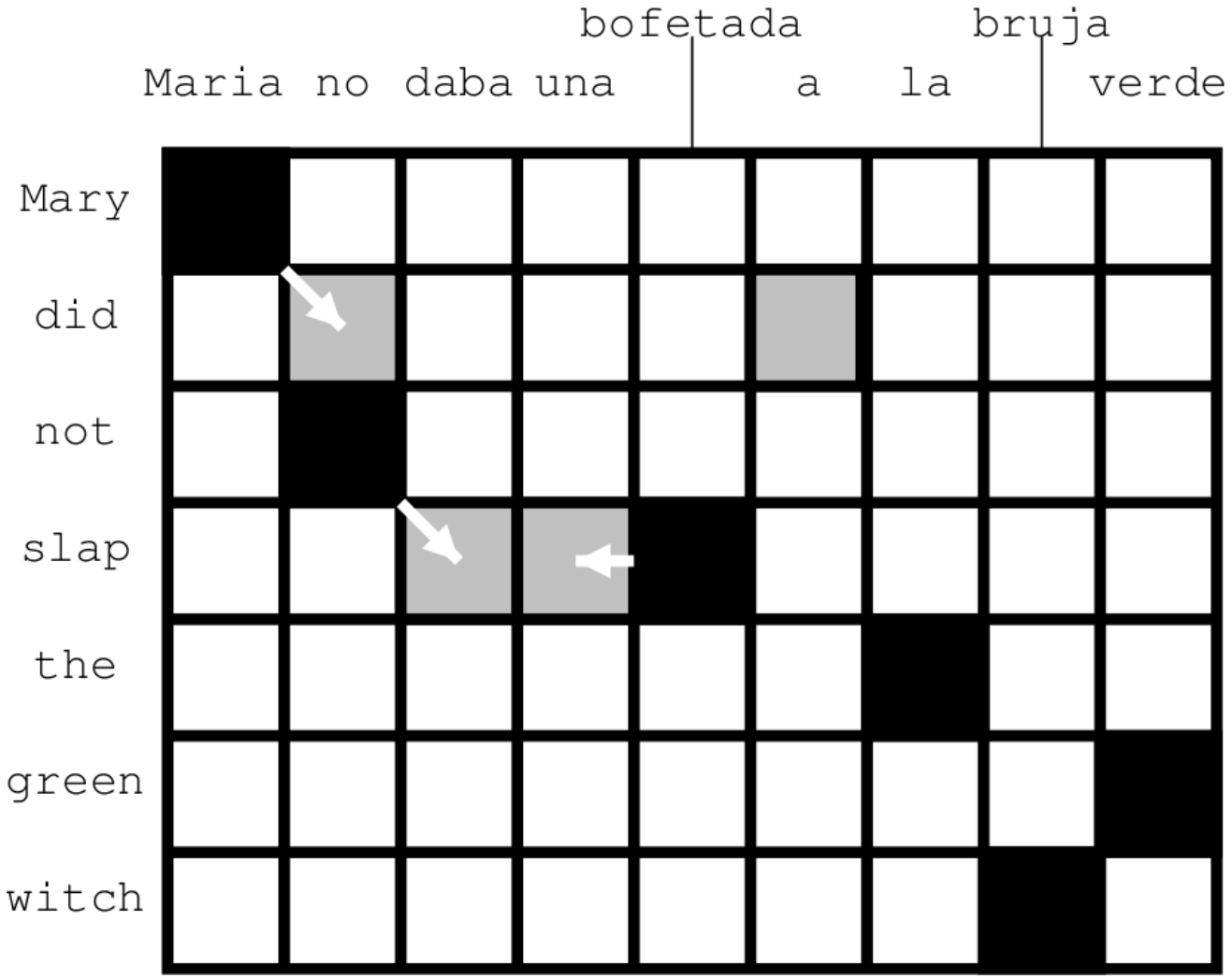
Symmetrization



"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).

Quotes on Statistical MT

Warren Weaver (1949):

> I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that is has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text.

Noam Chomsky (1969):

> ...the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

Frederick Jelinek (80's; IBM; later JHU and sometimes ÚFAL)

> Every time I fire a linguist, the accuracy goes up.

Hermann Ney (RWTH Aachen University):

> MT = Linguistic Modelling + Statistical Decision Theory

Statistical MT

Given a source (foreign) language sentence $f_1^J = f_1 \dots f_j \dots f_J$,
Produce a target language (English) sentence $e_1^I = e_1 \dots e_j \dots e_I$.

Among all possible target language sentences, choose the sentence with the highest probability:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}}\, p(e_1^I | f_1^J) \tag{1}$$

We stick to the e_1^I, f_1^J notation despite translating from English to Czech.

Brute-Force MT (1/2)

Translate only sentences listed in a "translation memory" (TM):

Good morning.	=	Dobré ráno.
How are you?	=	Jak se máš?
How are you?	=	Jak se máte?

$$p(e_1^I|f_1^J) = \begin{cases} 1 & \text{if } e_1^I = f_1^J \text{ seen in the TM} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Any problems with the definition?

Brute-Force MT (2/2)

Translate only sentences listed in a "translation memory" (TM):

Good morning.	=	Dobré ráno.
How are you?	=	Jak se máš?
How are you?	=	Jak se máte?

$$p(e_1^I|f_1^J) = \begin{cases} 1 & \text{if } e_1^I = f_1^J \text{ seen in the TM} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

- Not a probability. There may be f_1^J, s.t. $\sum_{e_1^I} p(e_1^I|f_1^J) > 1$.

  $\Rightarrow$ Have to normalize, use $\frac{\text{count}(e_1^I, f_1^J)}{\text{count}(f_1^J)}$ instead of 1.
- Not "smooth", no generalization:

Good morning.	$\Rightarrow$	Dobré ráno.
Good evening.	$\Rightarrow$	$\emptyset$

Bayes' Law

Bayes' law for conditional probabilities: $p(a|b) = \frac{p(b|a)p(a)}{p(b)}$

So in our case:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}}\, p(e_1^I | f_1^J) \qquad \text{Apply Bayes' law}$$

$$= \underset{I, e_1^I}{\operatorname{argmax}}\, \frac{p(f_1^J | e_1^I) p(e_1^I)}{p(f_1^J)} \qquad p(f_1^J) \text{ constant}$$

$$\Rightarrow \text{irrelevant in maximization}$$

$$= \underset{I, e_1^I}{\operatorname{argmax}}\, p(f_1^J | e_1^I) p(e_1^I)$$

Also called "Noisy Channel" model.

Motivation for Noisy Channel

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}}\, p(f_1^J | e_1^I) p(e_1^I) \tag{4}$$

Bayes' law divided the model into components:

$p(f_1^J | e_1^I)$ Translation model ("reversed", $e_1^I \rightarrow f_1^J$)

...is it a likely translation?

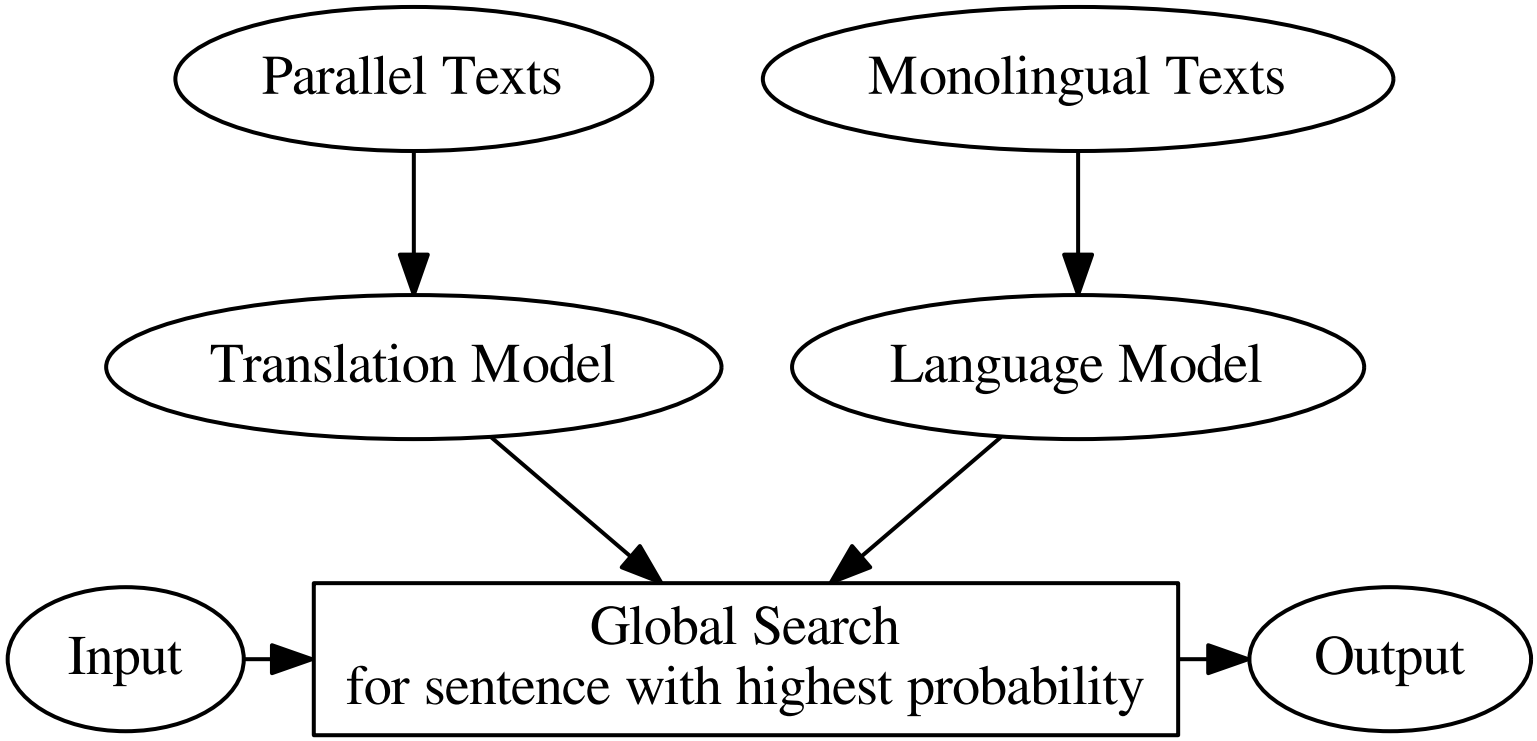
$p(e_1^I)$ Language model (LM)

...is the output a likely sentence of the target language?

- The components can be trained on different sources.

  There are far more monolingual data ⇒ language model more reliable.

Without Equations



From Bayes to Log-Linear Model

Och (2002) discusses some problems of Equation 4:

- Models estimated unreliably ⇒ maybe LM more important:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}}\, p(f_1^J | e_1^I)(p(e_1^I))^2 \tag{5}$$

- In practice, "direct" translation model equally good:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}}\, p(e_1^I | f_1^J) p(e_1^I) \tag{6}$$

- Complicated to correctly introduce other dependencies.
  ⇒ Use log-linear model instead.

Log-Linear Model (1)

- $p(e_1^I|f_1^J)$ is modelled as a weighted combination of models, called "feature functions": $h_1(\cdot,\cdot)\ldots h_M(\cdot,\cdot)$

$$p(e_1^I|f_1^J) = \frac{\exp(\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J))}{\sum_{e'^{I'}_1} \exp(\sum_{m=1}^M \lambda_m h_m(e'^{I'}_1, f_1^J))} \tag{7}$$

- Each feature function $h_m(e, f)$ relates source f to target e.
  E.g. the feature for n-gram language model:

$$h_{\mathrm{LM}}(f_1^J, e_1^I) = \log \prod_{i=1}^{I} p(e_i|e_{i-n+1}^{i-1}) \tag{8}$$

- Model weights λ_1^M specify the relative importance of features.

Log-Linear Model (2)

As before, the constant denominator not needed in maximization:

$$\begin{aligned}\hat{e}_1^{\hat{I}} &= \text{argmax}_{I,e_1^I} \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J))}{\sum_{e'^{I'}_1} \exp(\sum_{m=1}^{M} \lambda_m h_m(e'^{I'}_1, f_1^J))} \\ &= \text{argmax}_{I,e_1^I} \exp(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J))\end{aligned} \tag{9}$$

Relation to Noisy Channel

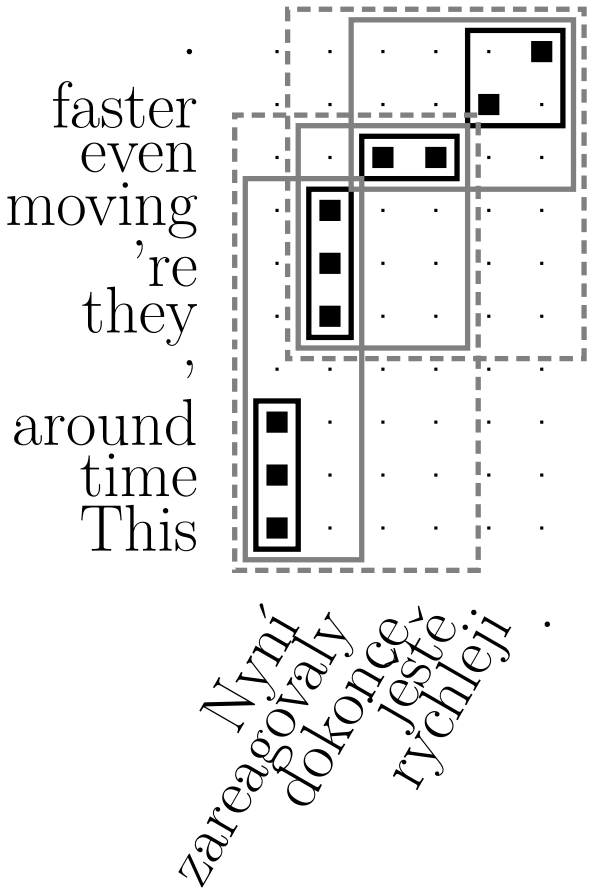
With equal weights and only two features:

- $h_{\mathrm{TM}}(e_1^I, f_1^J) = \log p(f_1^J|e_1^I)$ for the translation model,
- $h_{\mathrm{LM}}(e_1^I, f_1^J) = \log p(e_1^I)$ for the language model,

log-linear model reduces to Noisy Channel:

$$\begin{aligned}
\hat{e}_1^{\hat{I}} &= \mathrm{argmax}_{I,e_1^I} \exp\left(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\right) \\
&= \mathrm{argmax}_{I,e_1^I} \exp(h_{\mathrm{TM}}(e_1^I, f_1^J) + h_{\mathrm{LM}}(e_1^I, f_1^J)) \\
&= \mathrm{argmax}_{I,e_1^I} \exp(\log p(f_1^J|e_1^I) + \log p(e_1^I)) \\
&= \mathrm{argmax}_{I,e_1^I} p(f_1^J|e_1^I) p(e_1^I)
\end{aligned} \tag{10}$$

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).

Phrase-Based Translation Model

- Captures the basic assumption of phrase-based MT:
  1. Segment source sentence f_1^J into K phrases $\tilde{f}_1 \dots \tilde{f}_K$.
  2. Translate each phrase independently: $\tilde{f}_k \rightarrow \tilde{e}_k$.
  3. Concatenate translated phrases (with possible reordering R): $\tilde{e}_{R(1)} \dots \tilde{e}_{R(K)}$

The most important feature: phrase-to-phrase translation:

$$h_{\text{Phr}}(f_1^J, e_1^I, s_1^K) = \log \prod_{k=1}^{K} p(\tilde{f}_k | \tilde{e}_k) \tag{11}$$

Phrase-Based Features in Moses

Given parallel training corpus, phrases are extracted and scored:

in europa ||| in europe ||| 0.829007 0.207955 0.801493 0.492402
europas ||| in europe ||| 0.0251019 0.066211 0.0342506 0.0079563
in der europaeischen union ||| in europe ||| 0.018451 0.00100126 0.0319584 0.0196869

The scores are: $(\phi(\cdot) = \log p(\cdot))$

- phrase translation probabilities: $\phi_{\text{phr}}(f|e)$ and $\phi_{\text{phr}}(e|f)$
  The conditional probability of phrase $\tilde{f}_k$ given phrase $\tilde{e}_k$ is estimated from relative frequencies:

$$p(\tilde{f}_k|\tilde{e}_k) = \frac{\text{count}(\tilde{f}, \tilde{e})}{\text{count}(\tilde{e})} \tag{12}$$

- lexical weighting: $\phi_{\text{lex}}(f|e)$ and $\phi_{\text{lex}}(e|f)$ (Koehn, 2003)

Other Features Used in PBMT

- Word count/penalty: $h_{\text{wp}}(e_1^I, \cdot, \cdot) = I$

  $\Rightarrow$ Do we prefer longer or shorter output?
- Phrase count/penalty: $h_{\text{pp}}(\cdot, \cdot, s_1^K) = K$

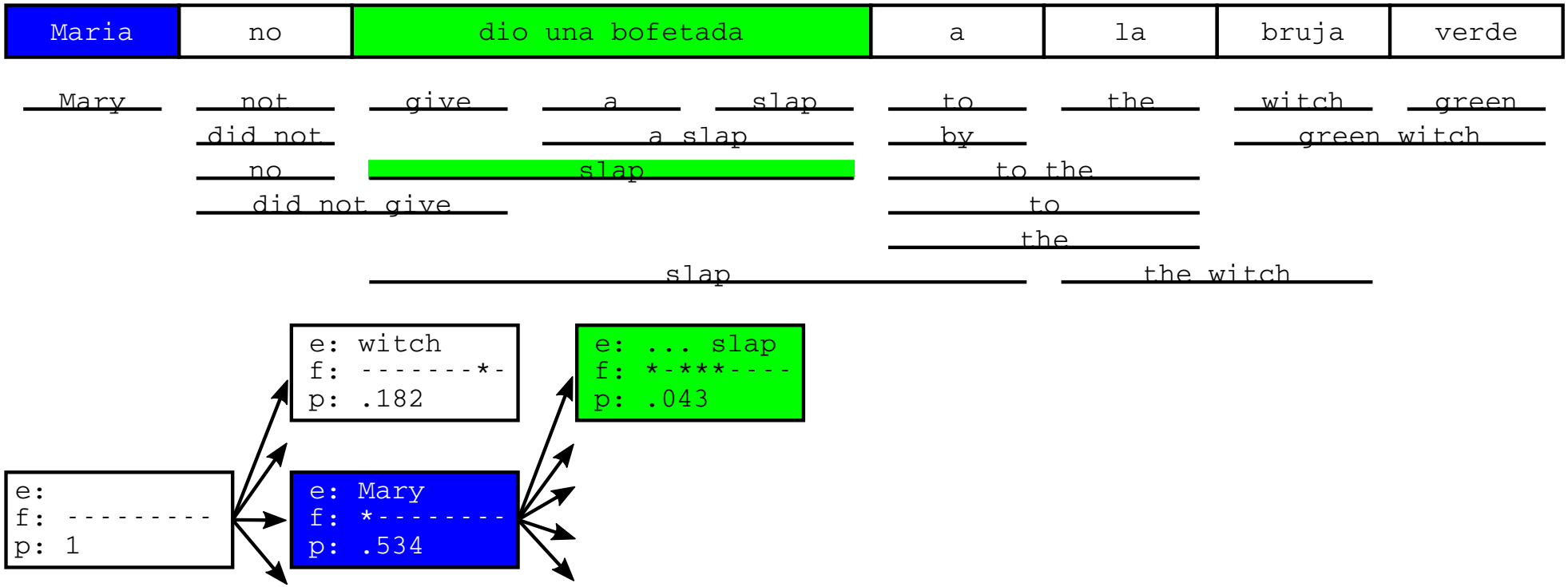
  $\Rightarrow$ Do we prefer translation in more or fewer less-dependent bits?
- Reordering model: different basic strategies (Lopez, 2009)

  $\Rightarrow$ Which source spans can provide continuation at a moment?
- n-gram LM:

$$h_{\text{LM}}(\cdot, e_1^I, \cdot) = \log \prod_{i=1}^{I} p(e_i | e_{i-n+1}^{i-1}) \tag{13}$$

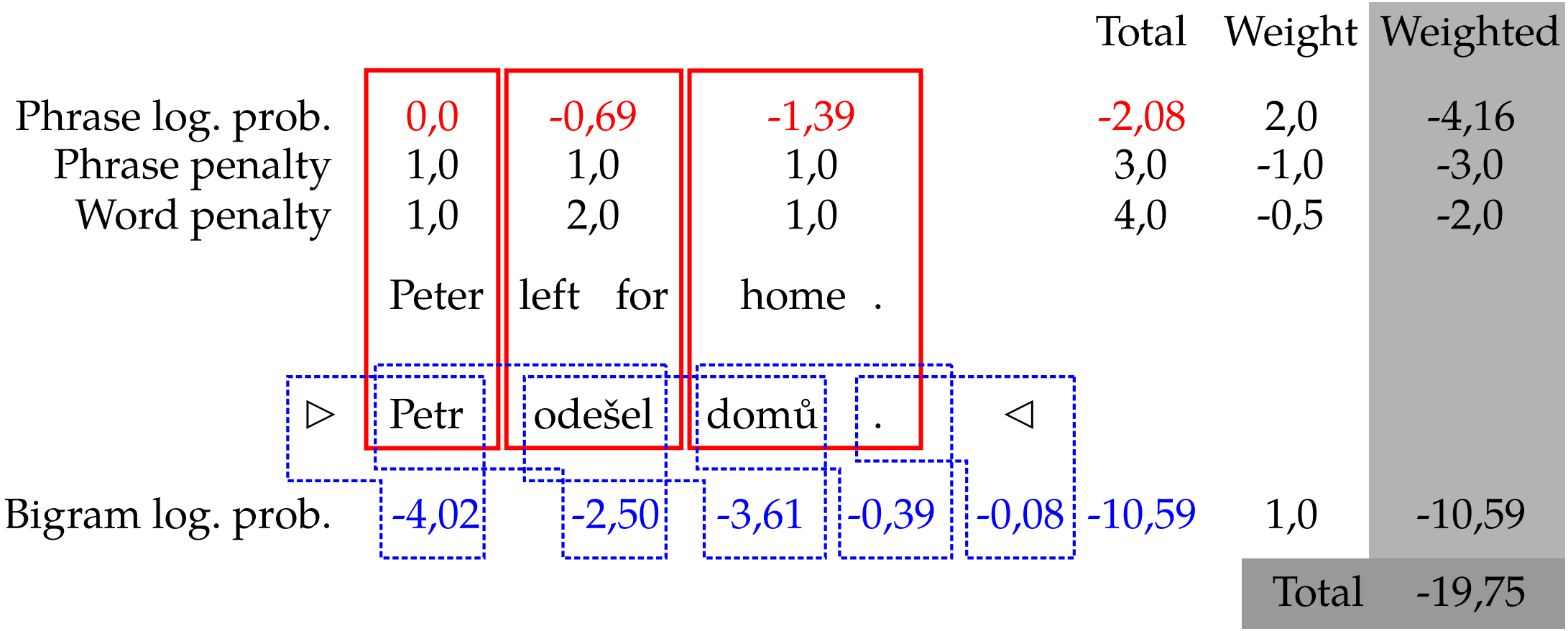
  $\Rightarrow$ Is output n-gram-wise coherent?

Decoding in Phrase-Based MT



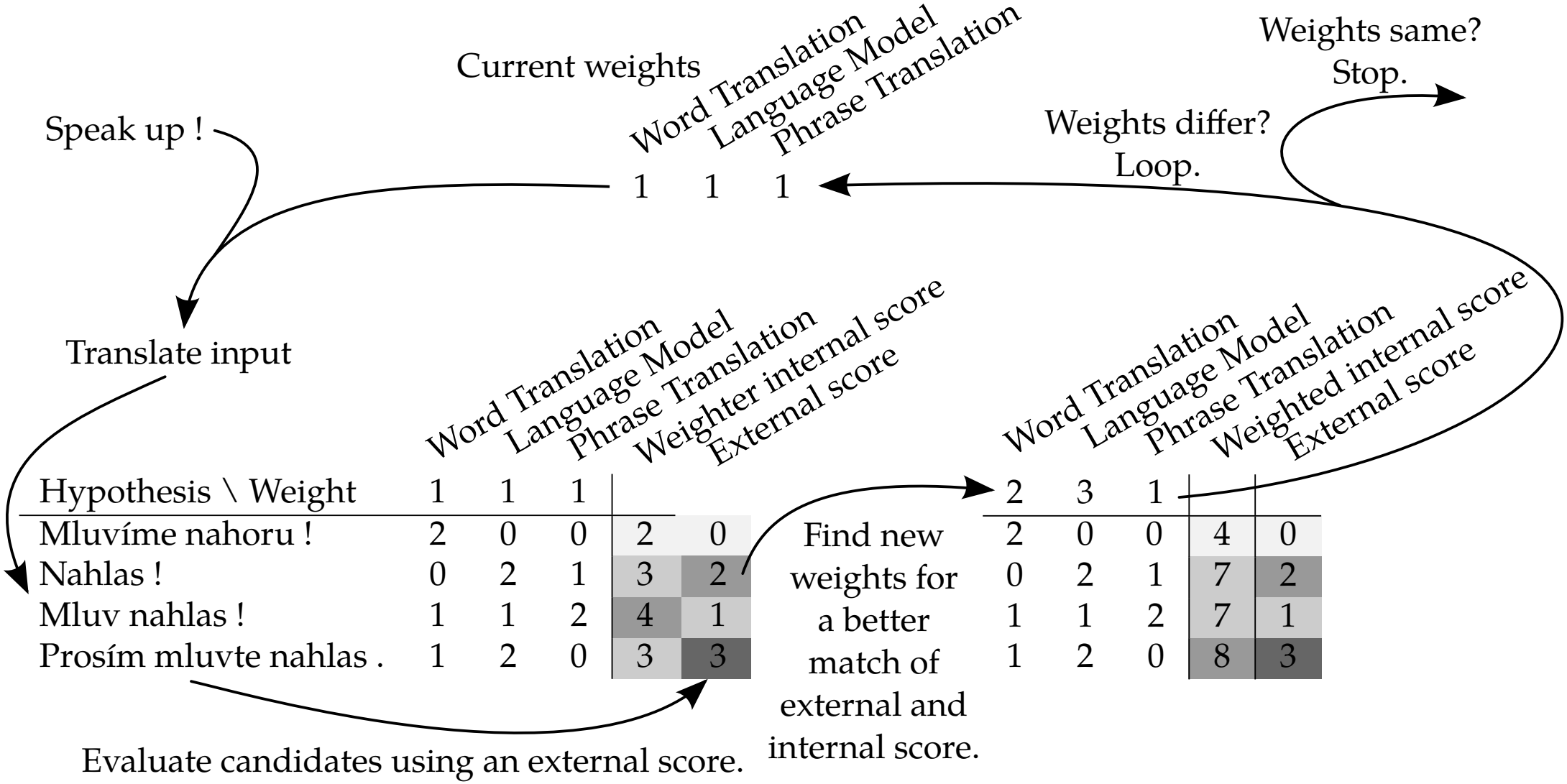
1. Collect translation options (all possible translations per span).
2. Gradually expand partial hypotheses until all input covered.
3. Prune less promising hypotheses.
4. When all input covered, trace back the best path.

Local and Non-Local Features



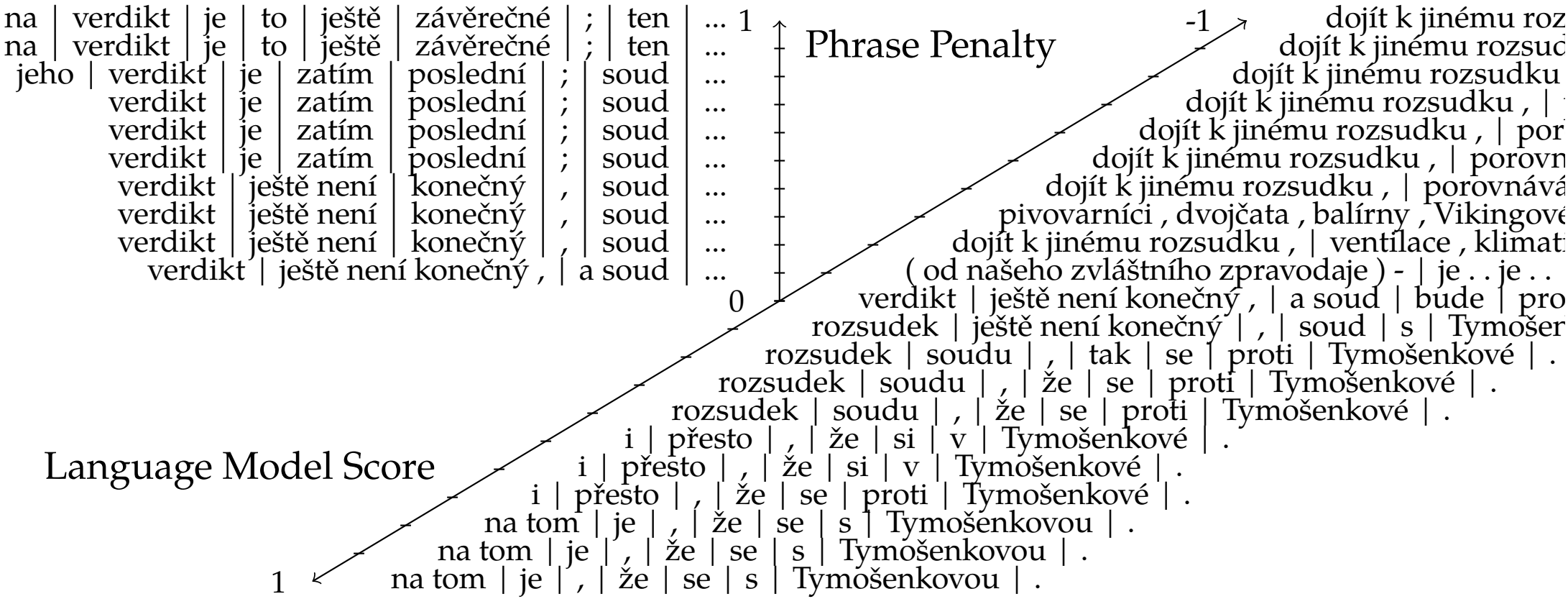
- Local features decompose along hypothesis construction.
  - Phrase- and word-based features.
- Non-local features span the boundaries (e.g. LM).

Weight Optimization: MERT Loop



Minimum Error Rate Training (Och, 2003)

Effects of Weights



- Higher phrase penalty chops sentence into more segments.
- Too strong LM weight leads to words dropped.
- Negative LM weight leads to obscure wordings.

Summary of PBMT

Phrase-based MT:

- is a log-linear model
- assumes phrases relatively independent of each other
- decomposes sentence into contiguous phrases
- search has two parts:
  - lookup of all relevant translation options
  - stack-based beam search, gradually expanding hypotheses

To train a PBMT system:

1. Align words.
2. Extract (and score) phrases consistent with word alignment.
3. Optimize weights (MERT).

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.
- Getting parallel data.
  - Including EM for word alignment.
- Phrase-based MT.
  - Log-linear model.
  - Local and non-local features.
  - MERT.

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

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.

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

Philipp Koehn. 2003. Noun Phrase Translation. Ph.D. thesis, University of Southern California.

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

Adam Lopez. 2009. Translation as weighted deduction. In Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 532–540, Athens, Greece, March. Association for Computational Linguistics.

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.

References

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1):19–51.

Franz Joseph Och. 2002. Statistical Machine Translation: From Single-Word Models to Alignment Templates. Ph.D. thesis, RWTH Aachen University.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In Proc. of the Association for Computational Linguistics, Sapporo, Japan, July 6-7.

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.

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.