Frázový statistický překlad z angličtiny do češtiny

Ondřej Bojar
obo@cuni.cz

9. říjen, 2006

Osnova

- **Frázový statistický překlad krok po kroku**
  - Frázový překlad
  - Hledání nejlepší hypotézy (beam search decoding)
  - Ladění vah (minimum error rate training)
- Frázový statistický překlad o více faktorech
- Experimenty s překladem do češtiny
- Malý rozbor chyb
- Závěr
- O workshopu

Anglické části pocházejí z prezentací Philippa Koehna.

Translation

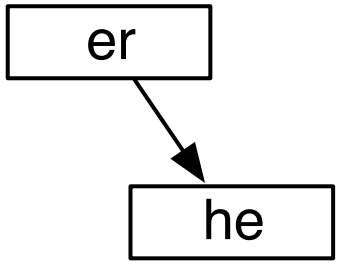
- Task: **translate this sentence** from German into English

er **geht** **ja** **nicht** **nach** **hause**

Translation step 1

- Task: translate this sentence from German into English

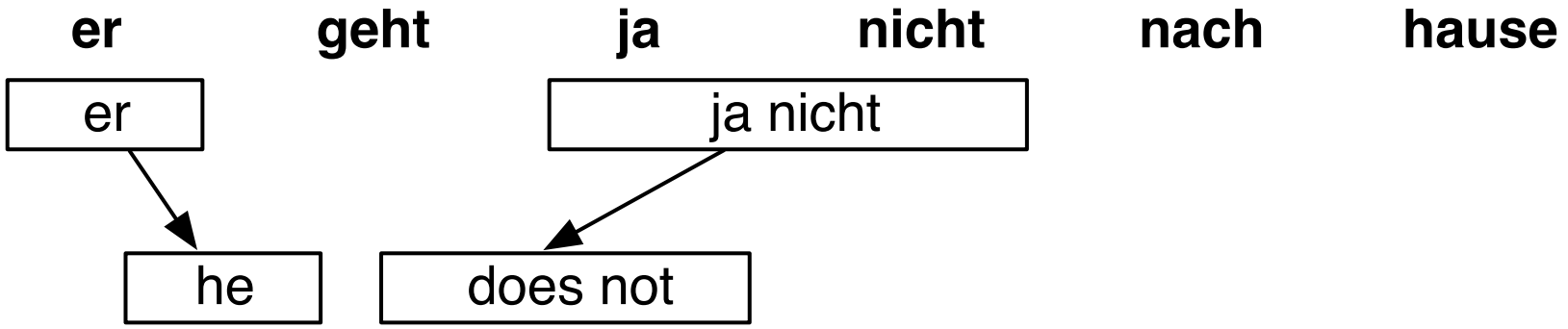
er geht ja nicht nach hause



- **Pick** phrase in input, **translate**

Translation step 2

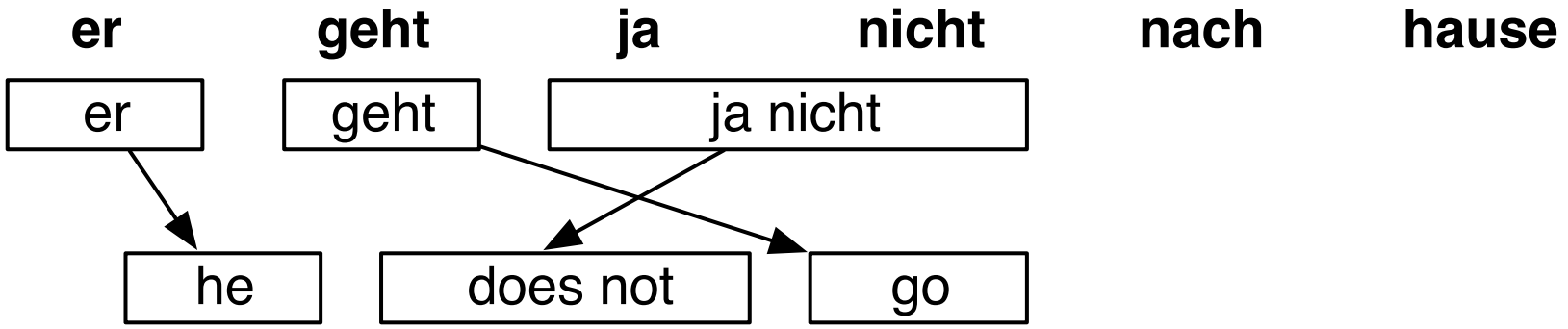
- Task: translate this sentence from German into English



- Pick phrase in input, translate
  - it is allowed to pick words **out of sequence** (**reordering**)
  - phrases may have multiple words: **many-to-many** translation

Translation step 3

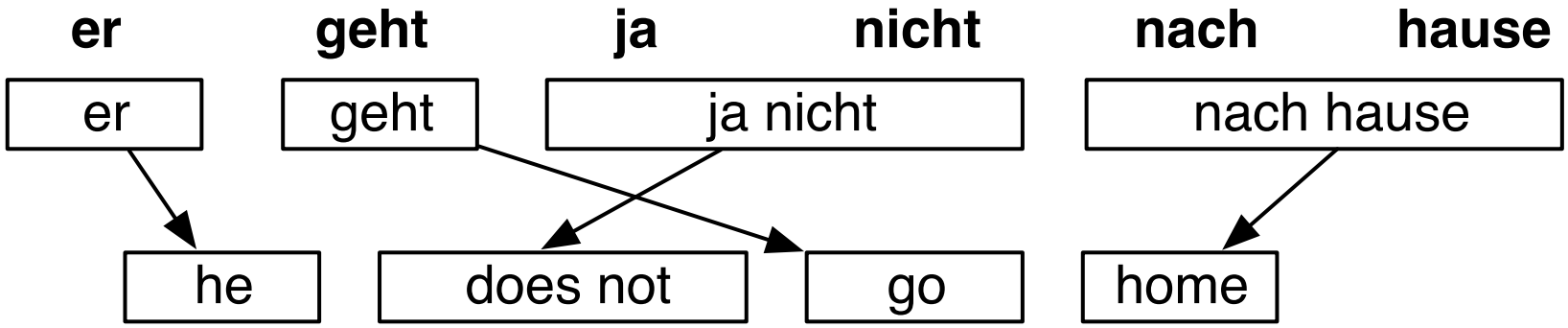
- Task: translate this sentence from German into English



- Pick phrase in input, translate

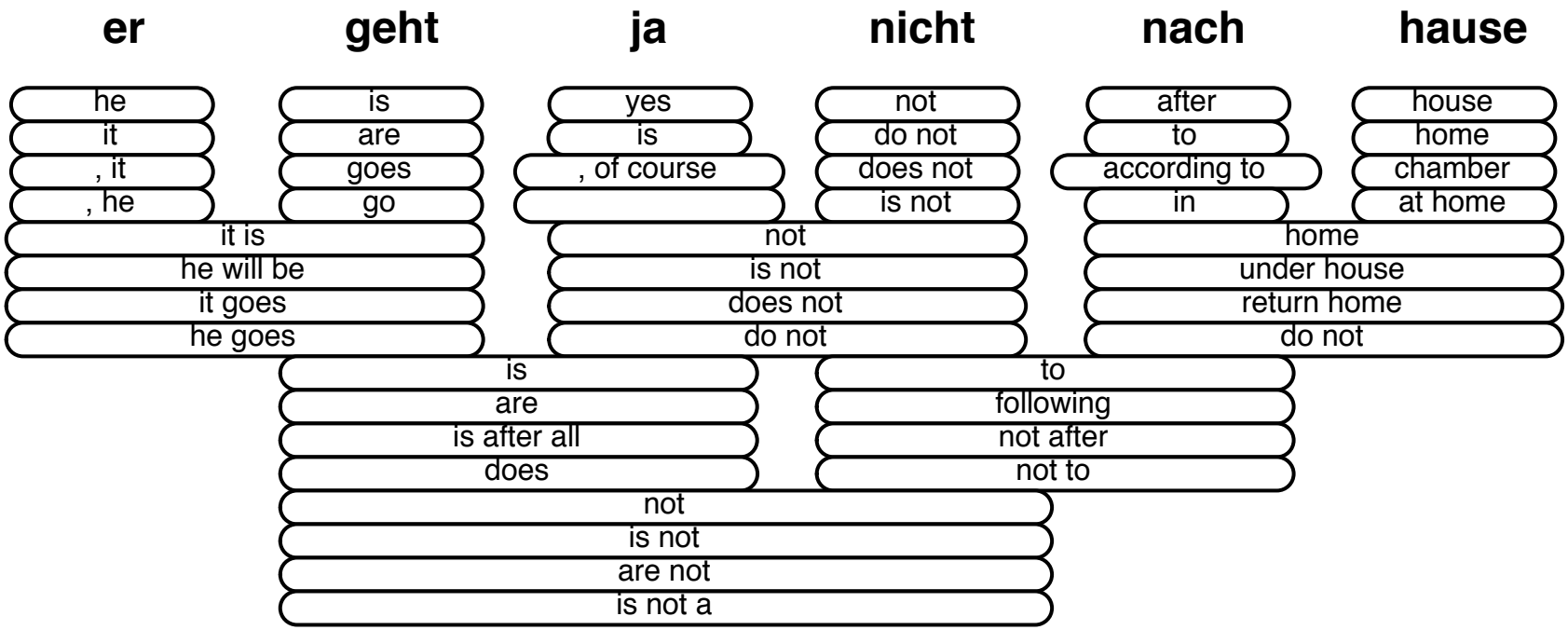
Translation step 4

- Task: translate this sentence from German into English



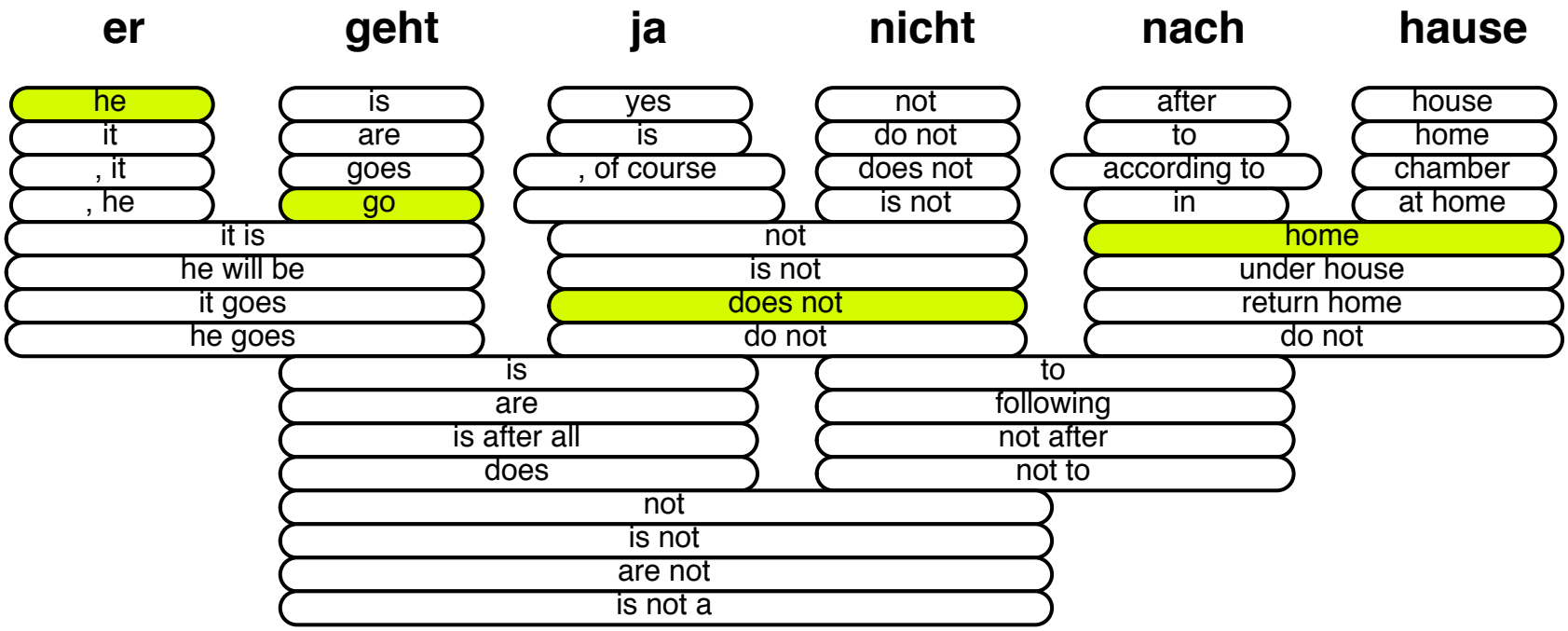
- Pick phrase in input, translate

Translation options



- **Many translation options** to choose from
  - in Europarl phrase table: **2727 matching phrase pairs** for this sentence
  - by pruning to the top 20 per phrase, **202 translation options** remain

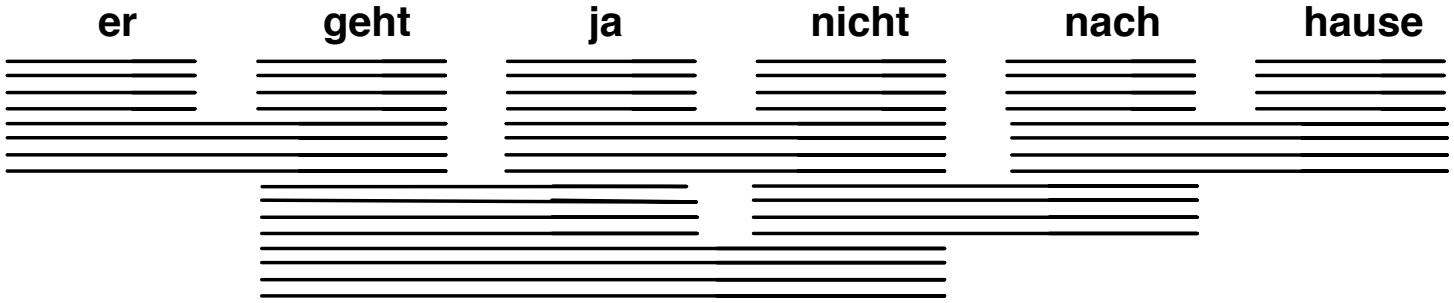
Translation options



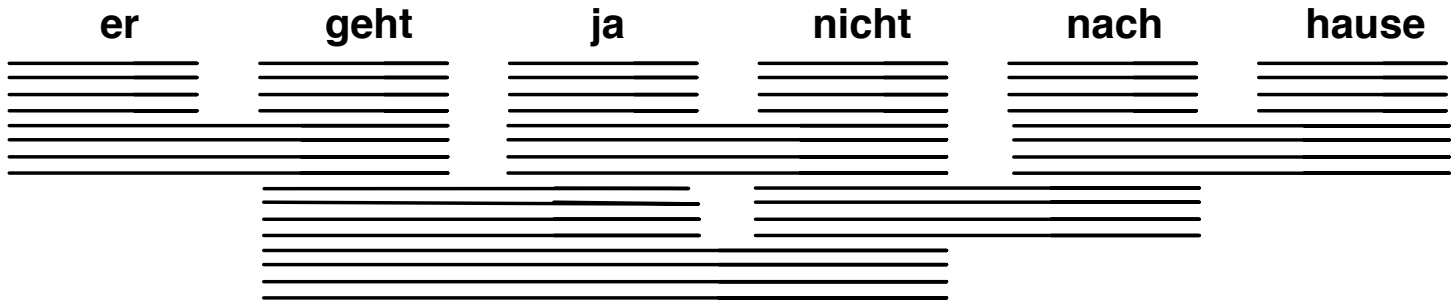
- The machine translation decoder does not know the right answer

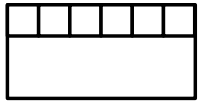
→ **Search problem** solved by heuristic beam search

Decoding process: precompute translation options

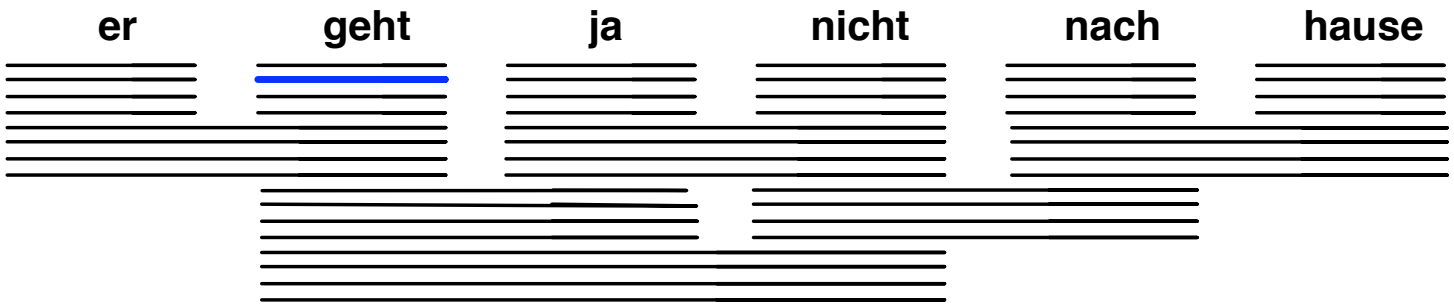


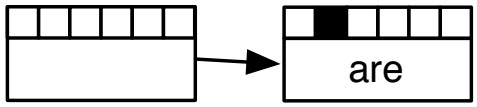
Decoding process: start with initial hypothesis



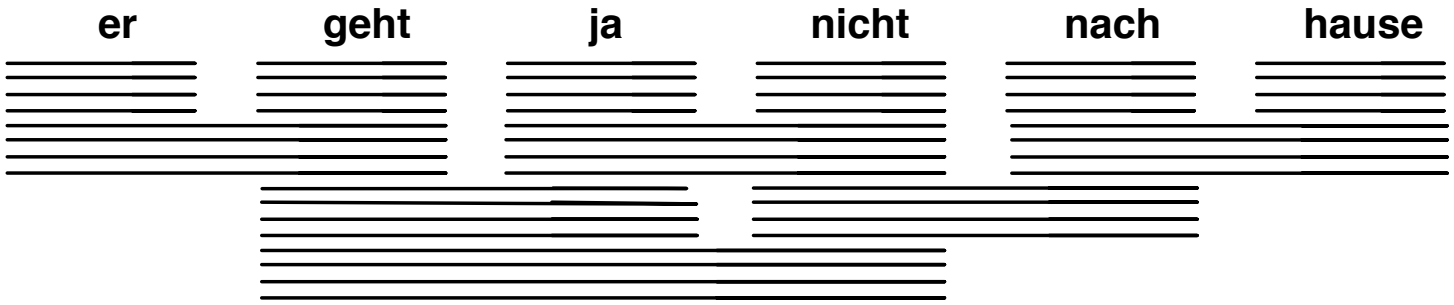


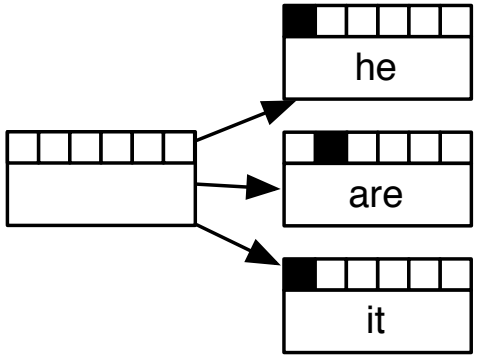
Decoding process: hypothesis expansion



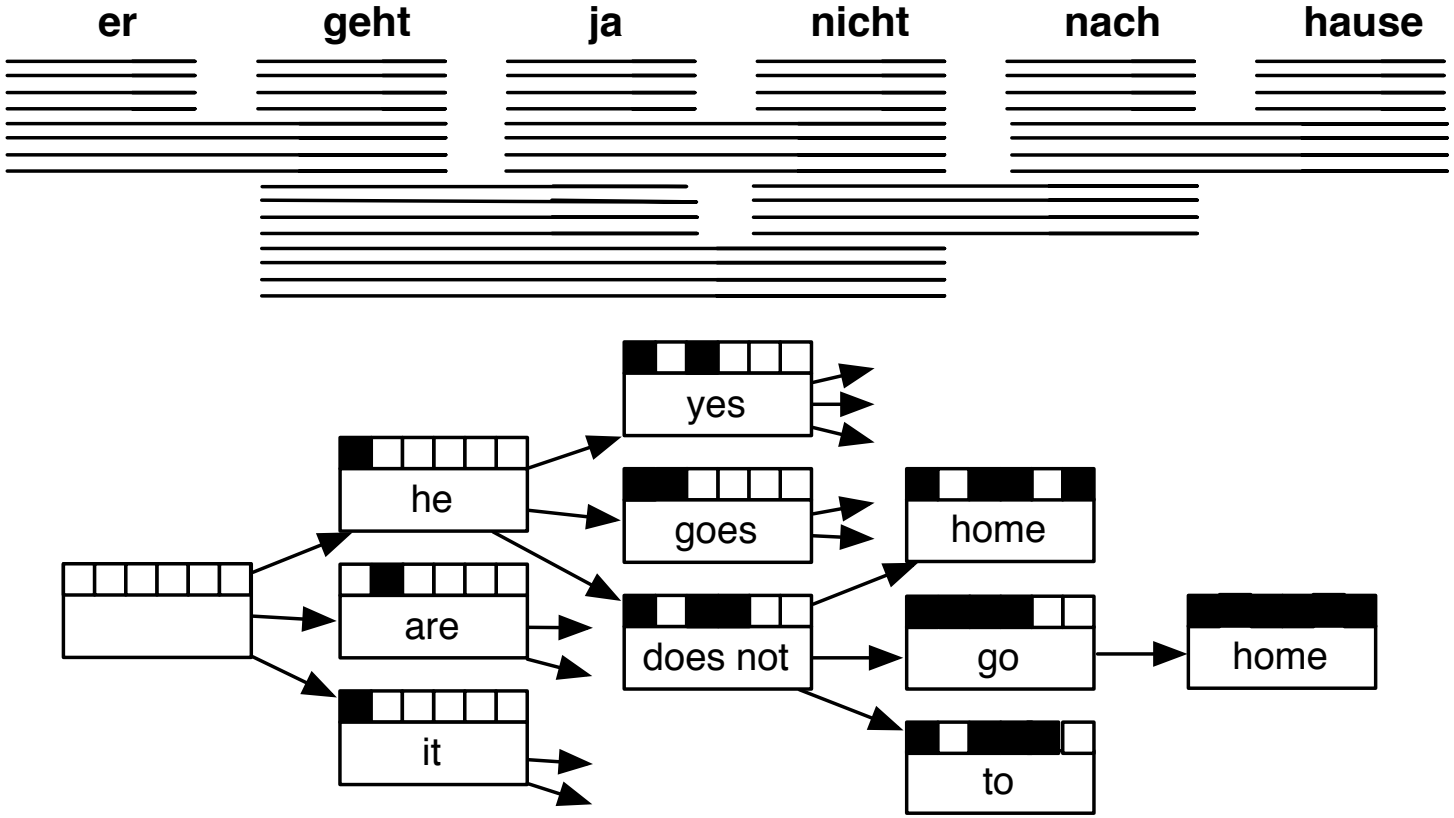


Decoding process: hypothesis expansion

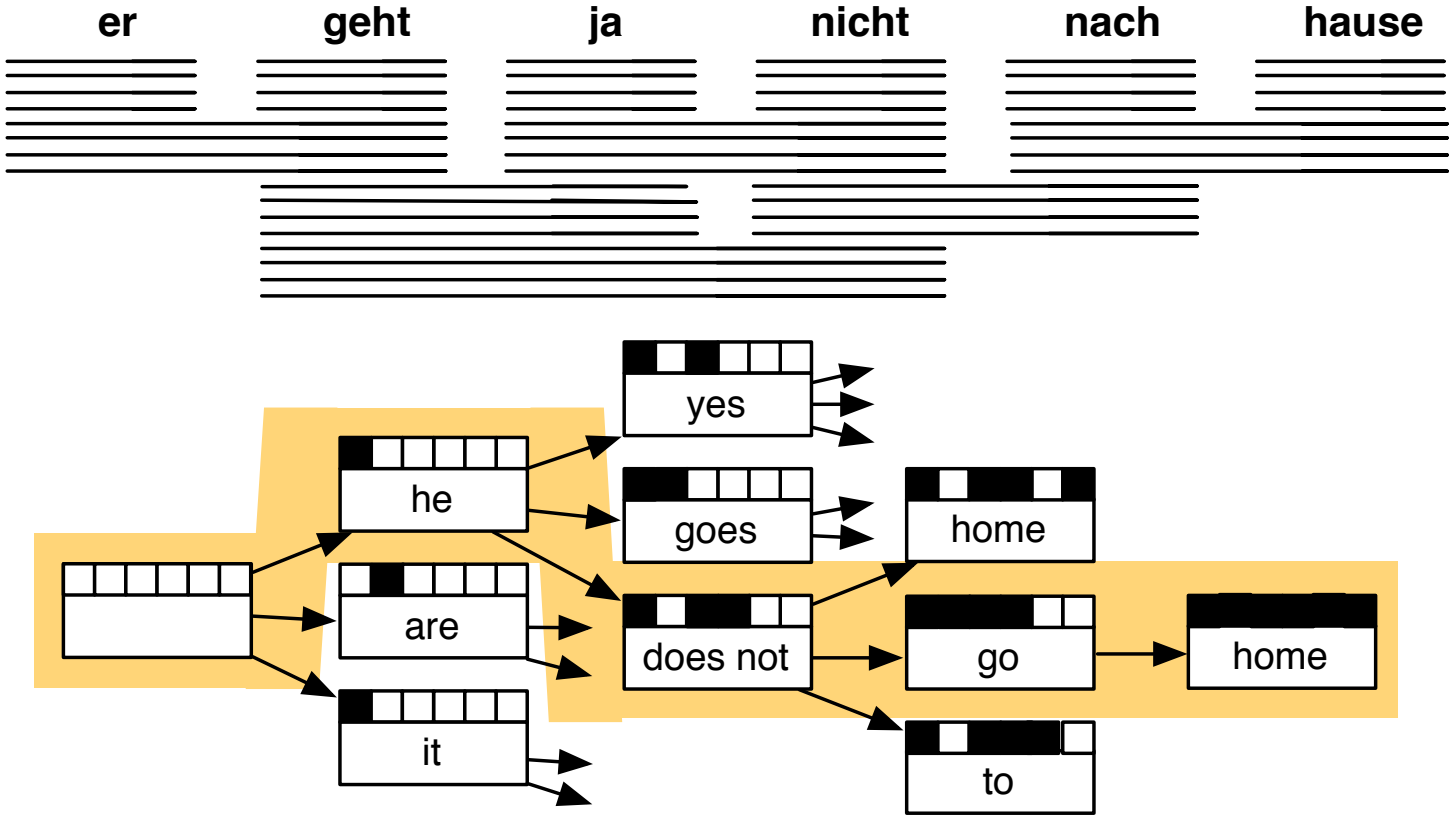




Decoding process: hypothesis expansion



Decoding process: find best path



Complexity of search

- Search space is **exponential** with number of input words
- Pruning required
  - organize hypotheses in **stacks**
  - each stack contains all hypotheses with the **same number of input words** translated
  - only the **top** n hypothesis are kept in a stack (Moses default: 200)
  - only hypothesis **worse by factor** α are kept in the stack (default: 0.03)
  - when comparing hypotheses, **future cost** has to be considered

  $\longrightarrow$ heuristic beam search is **polynomial** with sentence length
- Typically, a **reordering limit** is used (maximum movement)

  $\longrightarrow$ search is **linear** with sentence length

Knowledge sources

- Many different **knowledge sources** useful
  - language model
  - reordering (distortion) model
  - phrase translation model
  - word translation model
  - word count
  - phrase count
  - drop word feature
  - phrase pair frequency
  - additional language models
  - additional features

Set feature weights

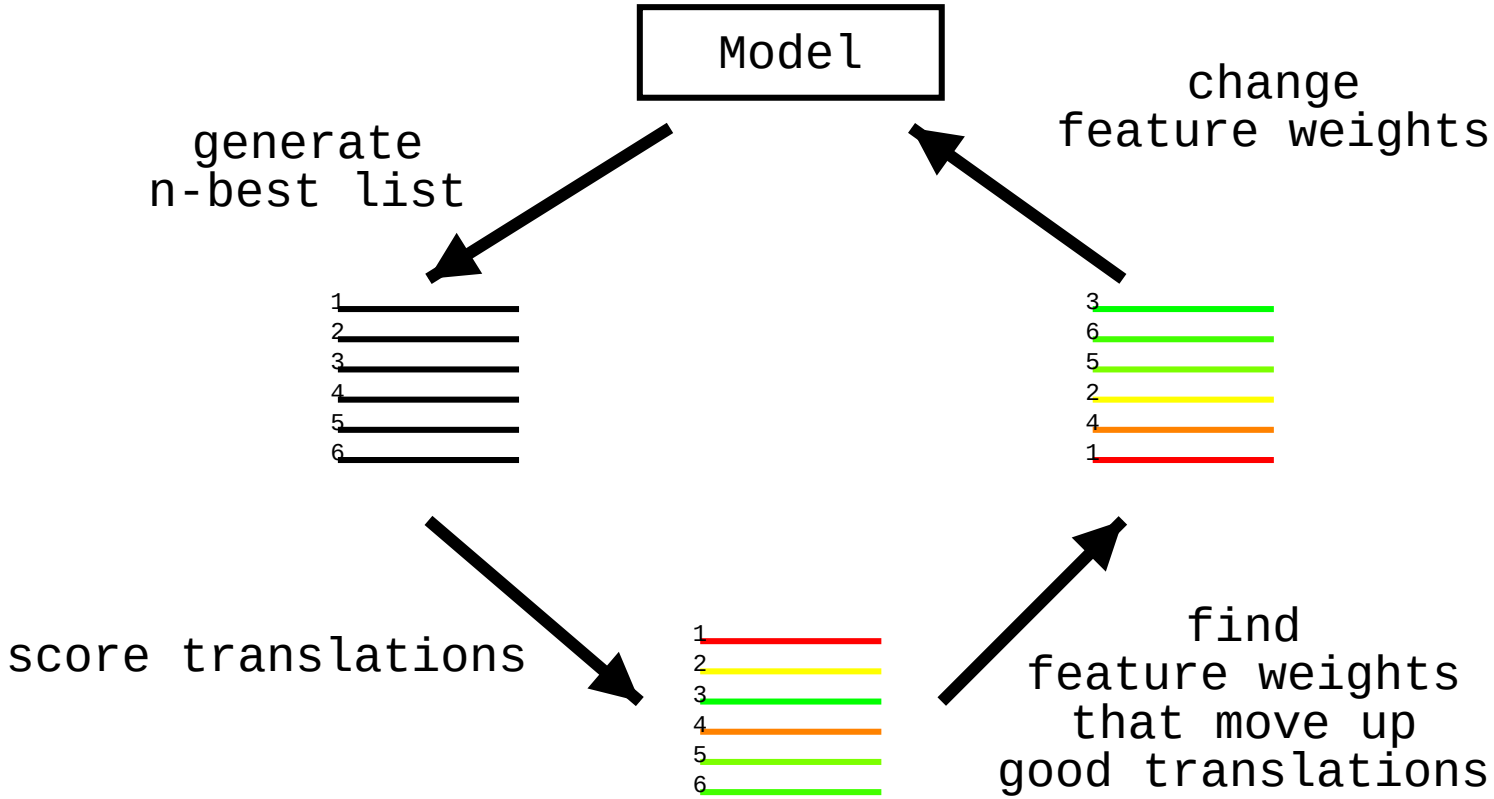
- Contribution of components p_i determined by weight λ_i
- Methods
  - **manual setting** of weights: try a few, take best
  - **automate** this process
- Learn weights
  - set aside a **development corpus**
  - set the weights, so that **optimal translation performance** on this development corpus is achieved
  - requires **automatic scoring** method (e.g., BLEU)

Learning task

- Task: **find weights**, so that feature vector of the correct translations **ranked first**

rank	TRANSLATION	LM	TM	WP	SER
1	Mary not give slap witch green .	-17.2	-5.2	-7	1
2	Mary not slap the witch green .	-16.3	-5.7	-7	1
3	Mary not give slap of the green witch .	-18.1	-4.9	-9	1
4	Mary not give of green witch .	-16.5	-5.1	-8	1
5	Mary did not slap the witch green .	-20.1	-4.7	-8	1
6	Mary did not slap green witch .	-15.5	-3.2	-7	1
7	Mary not slap of the witch green .	-19.2	-5.3	-8	1
8	Mary did not give slap of witch green .	-23.2	-5.0	-9	1
9	Mary did not give slap of the green witch .	-21.8	-4.4	-10	1
10	Mary did slap the witch green .	-15.5	-6.9	-7	1
11	**Mary did not slap the green witch .**	**-17.4**	**-5.3**	**-8**	**0**
12	Mary did slap witch green .	-16.9	-6.9	-6	1
13	Mary did slap the green witch .	-14.3	-7.1	-7	1
14	Mary did not slap the of green witch .	-24.2	-5.3	-9	1
15	Mary did not give slap the witch green .	-25.2	-5.5	-9	1
rank	**translation**	**feature vector**			

Learn feature weights



Osnova

- Frázový statistický překlad krok po kroku
- **Frázový statistický překlad o více faktorech**
  - Motivace: morfologie
  - Krok za krokem
  - Další nápady na využití více faktorů
- Experimenty s překladem do češtiny
- Malý rozbor chyb
- Závěr
- O workshopu

Motivace pro víc faktorů: zlepšit tvarosloví

- Statistický překlad do morfologicky bohatých jazyků funguje hůř než opačně.
  Viz např. (Koehn, 2005).

Motivace pro překlad angličtina→čeština:

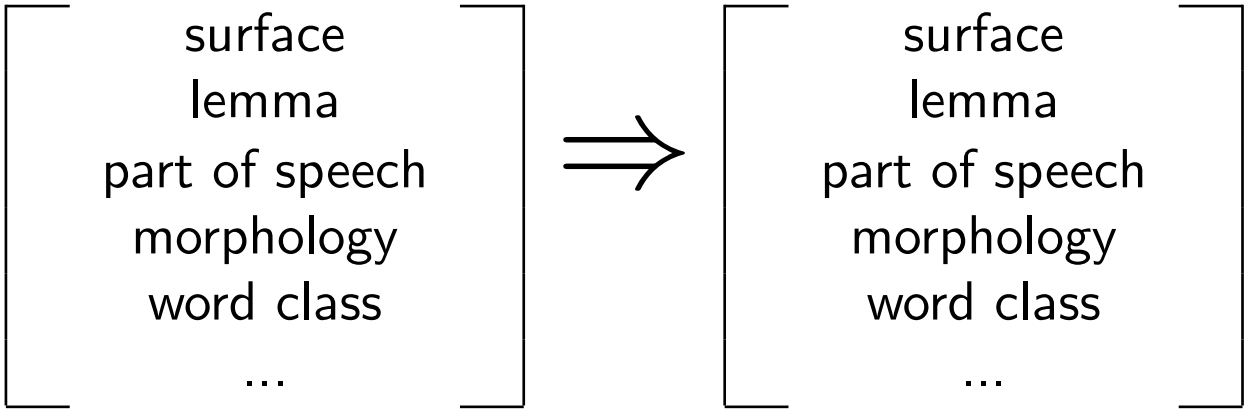
Běžné BLEU	25%
BLEU bez ohledu na slovní tvary[1]	33%

⇒ Dokonalým tvarováním slov by bylo možné zlepšit skóre o 8 bodů.

[1] BLEU lematizovaného výstupu proti lematizovaným referencím.

Factored translation models

- **Factored represention** of words



- Benefits
  - **generalization**, e.g. by translating lemmas, not surface forms
  - **richer model**, e.g. using syntax for reordering, language modeling

Translation process: example

Input: Häuser → (Häuser, Haus, NNS)

1. **Translation step**: lemma ⇒ lemma
   (?, house, ?), (?, home, ?)
2. **Generation step**: lemma ⇒ part-of-speech
   (?, house, NN), (?, house, NNS), (?, home, NN), (?, homes, NNS)
3. **Translation step**: part-of-speech ⇒ part-of-speech
   ~~(?, house, NN)~~, (?, house, NNS), ~~(?, home, NN)~~, (?, homes, NNS)
4. **Generation step**: lemma, part-of-speech ⇒ surface
   (houses, house, NNS), (homes, home, NNS)

Pořadí překladových kroků je pevně určeno konfigurací.

Another idea: Subject-verb agreement

- Lexical n-gram language model would prefer

  the paintings of the old man is beautiful

  old man is is a **better trigram** than old man are

- Correct translation

  the paintings of the old man are beautiful
  - SBJ-plural - - - - V-plural -

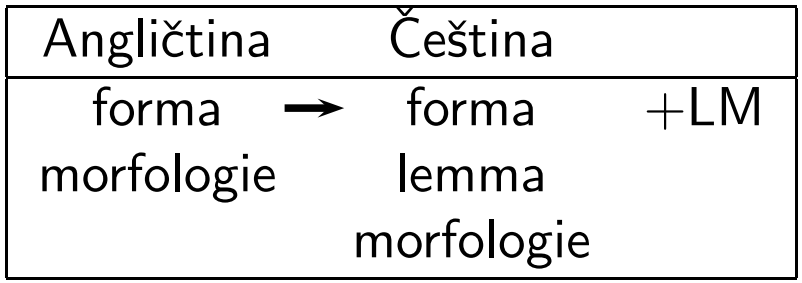
- **Special tag** that tracks *count* of *subject* and *verb*
  p(-,SBJ-plural,-,-,-,-,V-plural,-) > p(-,SBJ-plural,-,-,-,-,V-singular,-)
  "skip" language model: p(SBJ-plural,V-plural) > p(SBJ-plural,V-singular)

Osnova

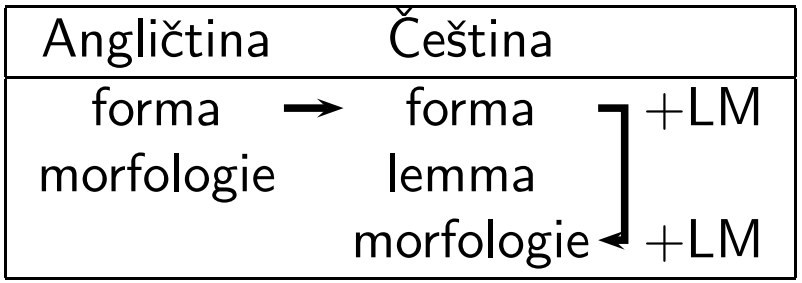
- Frázový statistický překlad krok po kroku
- Frázový statistický překlad o více faktorech
- **Experimenty s překladem do češtiny**
  - Scénáře překladu
  - Jemnost morfologie
  - Více dat?
- Malý rozbor chyb
- Závěr
- O workshopu

Testované scénáře překladu

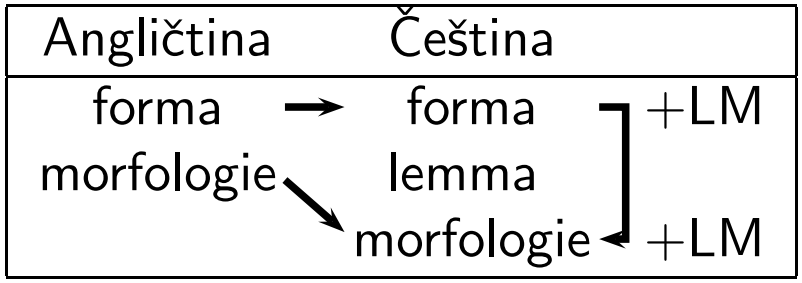
Pouze překlad (P)



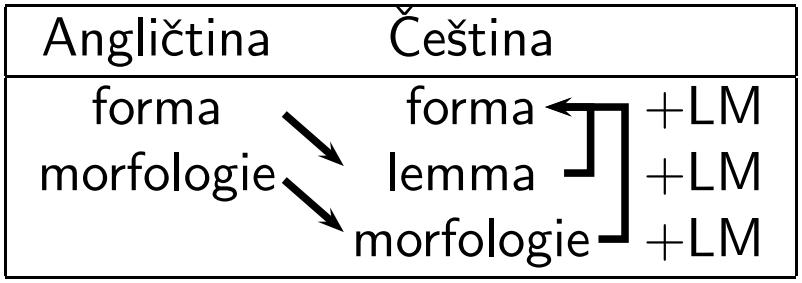
Překlad+kontrola (P+K)



2*překlad+kontrola (P+P+K)



2*překlad+generování (P+P+G)



Výsledky scénářů

	Dev (std)	Dev (opt)	Test (opt)
Baseline: Pouze překlad (P)	25.68	29.24	25.23
2*překlad+generování (P+P+G)	23.93	30.34	25.94
2*překlad+kontrola (P+P+K)	25.12	30.73	26.43
Překlad+kontrola (P+K)	23.51	30.88	27.23

⇒ Přidání faktorů (a jazykových modelů) pro morfologii vždy pomohlo.

⇒ Čím složitější scénář, tím horší výsledek.

Důvodem jsou zřejmě chyby v hledání (search errors):

víc faktorů ⇒ větší prostor hypotéz (možná stejným povrchem) ⇒ hloubka zásobníku nemusí stačit

Míra detailu v reprezentaci morfologie (P+K)

Tagset	Typů viděno	Popis
plné zn.	1098	Plná morfologická značka, 15 pozic, teoreticky 4000 různých značek.
POS	173	Slovní druh a poddruh, u {N,A,P,R} navíc pád.
CNG01	571	Jako POS, ale {N,A,P,R} odlišují pád, číslo a rod.
CNG02	707	Pád, číslo a rod odlišen u {N,A,P,R,C,V}, lemma interpunkce přidáno ke značce.
CNG03	899	Jména a číslovky vyjadřují pád, číslo a rod, zvýrazněno zvratné se/si. Slovesa vyjadřují číslo, rod, čas a vid, zvýrazněno $být$. Předložky vyjadřují pád i lemma, lemma přidáno ke značce {Z,T,I}, tvar čísel u C=.

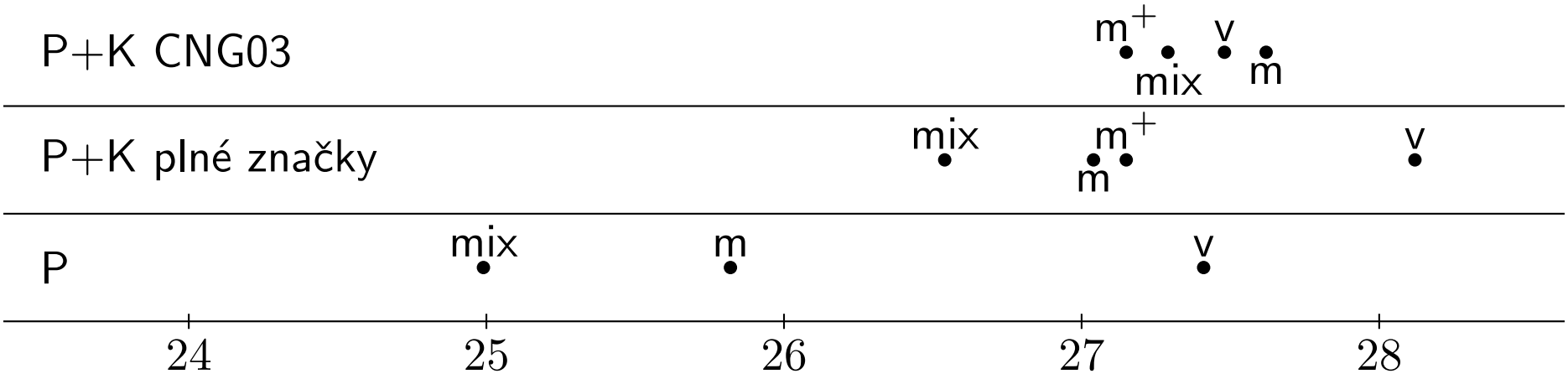
Výsledky P+K s různě jemnou morfologií

	Dev (std)	Dev (opt)	Test (opt)
Baseline: P (jediný faktor)	26.52	28.77	25.82
P+K, CNG01	22.30	29.86	26.14
P+K, POS	21.77	30.27	26.57
P+K, plné značky	22.56	29.65	27.04
P+K, CNG02	23.17	30.77	27.45
P+K, CNG03	23.27	30.75	27.62

⇒ Není chybou přidat detailní morfologii.

⇒ Lépe však funguje specifická sada značek opírající se o znalost jazyka i morf. systému.

Více dat ve scénáři P+K



m=malá data, 20k vět v doméně

m^+=malá data, 20k vět v doméně, ale navíc oddělené jazykové modely (840k vět mimo doménu)

v=velká data, 20k vět v doméně a 840k vět mimo ni, oddělené jazykové modely vět

mix=velká data, 20k vět v doméně a 840k vět mimo, smíšeno do společného jazykového modelu

Osnova

- Frázový statistický překlad krok po kroku
- Frázový statistický překlad o více faktorech
- Experimenty s překladem do češtiny
- **Malý rozbor chyb**
- Závěr
- O workshopu

Dosáhli jsme cíle zlepšit morfologii?

- Na malých datech dosaženo významného zlepšení:
  English→Czech: 20k vět, BLEU zvýšeno z 25.82% na 27.62%
  nebo až na 28.12% při použití dodatečných dat.
- Stále nedosahujeme úrovně lematizovaného BLEU (35%).
- Lokální shoda je již celkem dobrá:

  Mikrostudie: shoda přídavného a podstatného jména
  74% adjektiv ve shodě, 2% v neshodě, (v dalších případech chybělo subst. ap.)

⇒ Kde tedy zůstávají chyby v morfologii?

Mikrostudie chyb v překladu angličtina→čeština

Mikrostudie na nejlepším dosaženém výstupu (BLEU 28,12%), intuitivní metrika:

- studováno 15 vět, z celk. počtu 77 dvojic sloveso-slovesné doplnění ve *vstupním* textu:

překlad	. . . zachoval význam	. . . nebyl srozumitelný	. . . úplně chyběl
slovesa	43%	14%	21%
doplnění	79%	12%	6%

Navíc z případů, kdy sloveso i doplnění byly přeloženy správně, mělo 44% negramatickou nebo nesprávnou vazbu.

Ukázkové chyby

Vstup:	Keep on investing.
Výstup MT:	Pokračovalo investování. (grammar correct here!)
Glosa:	Continued investing. (Meaning: The investing continued.)
Správně:	Pokračujte v investování.

⇒ jazykový model vyhrál neprávem ⇒ nutno zohlednit valenci na zdrojové straně.

Vstup:	brokerage firms rushed out ads . . .			
Výstup MT:	brokerské	firmy	vyběhl	reklamy
Glosa:	brokerage	firms$_{pl.fem}$	ran$_{sg.masc}$	ads$_{pl.nom,pl.acc,pl.voc,sg.gen}$
Správná možnost 1:	brokerské	firmy	vyběhly	s reklamami$_{pl.instr}$
Správná možnost 2:	brokerské	firmy	vydaly	reklamy$_{pl.acc}$

Data na cílové straně možná dost bohatá k identifikaci: vyběhnout–s–$instr$

Určitě nebudou dost bohatá pro všechny morfologické a lexikální varianty:

vyběhl–s–reklamou, vyběhla–s–reklamami, vyběhl–s–prohlášením, vyběhli–s–oznámením, . . .

Osnova

- Frázový statistický překlad krok po kroku
- Frázový statistický překlad o více faktorech
- Experimenty s překladem do češtiny
- Malý rozbor chyb
- **Závěr**
  - Porovnání s angličtinou
  - Návrat do reality (ukázka překladu)
  - Shrnutí
  - Celkové poučení
- O workshopu

Porovnání s angličtinou a lidmi

	Do angličtiny	Do češtiny	Rozdíl
Lidé	55.3±6.0	46.3±4.3	-9.0
P, 20k vět	28.50	25.23	-3.27
P+K, 20k vět	28.66	27.23	-1.43
P, 860k vět, směs domén	34.12	25.40	-8.72

- BLEU lidí počítáno 5x a průměrováno
- větší rozptyl do angličtiny, protože i původní text sloužil jako jedna z referencí
- nižší skóre do češtiny může ukazovat na větší tvaroslovnou a slovoslednou volnost, ale také možná jen, že do angličtiny překládali přímočařeji
- P+K podle očekávání pomohlo víc do češtiny
- do angličtiny pomohlo víc dat a překvapivě nevadilo smísit domény

Ukázka překladu do češtiny (BLEU 28,12)

jsme asi navštívit , pokud reklama funguje .
těžko na patách pátečního propadu akciového 190-point a nejistota , že to následovalo , několik velkých brokerských firem , které jsou kolébat nové reklamy , že známá zpráva : pokračovalo investování , trh je docela dobře .
jejich posláním je , aby udrželi klienti prchají z trhu , jako individuální investoři udělali v droves po krachu v říjnu
právě dny po krachu v roce 1987 , hlavní brokerské firmy vyběhl reklamy na klidné investory .
tentokrát kolem , jsou pohybující ještě rychlejší .

We are about to see if advertising works .
Hard on the heels of Friday 's 190-point stock-market plunge and the uncertainty that 's followed , a few big brokerage firms are rolling out new ads trumpeting a familiar message : Keep on investing , the market 's just fine .
Their mission is to keep clients from fleeing the market , as individual investors did in droves after the crash in October
Just days after the 1987 crash , major brokerage firms rushed out ads to calm investors .
This time around , they are moving even faster .

Shrnutí

- Frázový překlad do češtiny není zcela beznadějný.

  Podobně jako u jiných jazyků lokální shody přijatelné, problém celkové koherence.

- Kontrola morfologie ve vícefaktorovém překladu pomáhá (čj, nj, špaň.).

  . . . a to i v případě unsupervised morfologie (třídy slov místo značek, čínšť.).

  – Efekt se snižuje při použití více dat (čj, špaň.).
  – Efekt se snižuje při užití složitějšího scénáře.

- Jiná testovaná využití více faktorů zatím příliš nepomohla.

  aj⟶nj: celková "struktura" věty, shoda čísla slovesa a subjektu

  aj⟶čj: povrchová valence

- Data mimo doménu mohou i ublížit.

Celkové poučení (znovu a znovu a znovu!)

1. Potřebuješ úkol a metriku kvality výstupu.
   Čím blíž je úkol ke každodenním problémům, tím lépe.
2. Napřed zkus *nejjednodušší* způsob, jak problém řešit.
   Čím víc jsi vzdělán/vychován, tím *těžší* bývá mít jednoduché nápady.
3. Testuj na málo datech, i když jich máš hodně.
   Ladění očividných chyb je rychlejší.
   Případné zlepšení bude markantnější ;-)
4. Opravuj chyby *časté*, nikoli chyby nápadné!
   Mikroevaluace je velmi užitečná. Projdi 10 vzorků výstupu, pojmenuj chyby, posčítej.

Osnova

- Frázový statistický překlad krok po kroku
- Frázový statistický překlad o více faktorech
- Experimenty s překladem do češtiny
- Malý rozbor chyb
- Závěr
- **O workshopu**

Celkově o workshopu

Výsledky týmu SMT na workshopu: `http://www.clsp.jhu.edu/ws2006/`

- Moses: systém pro strojový překlad, zřejmě bude LGPL.
- Vícefaktorové modely pro frázový překlad.
- Confusion networks (aproximace slovních grafů) pro překlad víceznačného vstupu.

Budoucnost:

- Euromatrix:
  - Edinburgh: Další využití více faktorů (reordering, lepší modely celkové koherence).
  - ÚFAL: Stromečkový dekodér (navazuje na disertaci M. Čmejrka).
- MIT (M. Collins, B. Cowan):
  - "Překlad valenčních rámců", doplnění pak frázově.
- RWTH Aachen, ITC irst:
  - Překlad mluvené řeči, confusion networks.

Jak dosáhnout úspěšného workshopu

Všechno musí být připraveno předem.

- Moses byl vyvíjen 9 měsíců předem, workshop "jen" ověřil použitelnost.
  Přesto první dva tři týdny *neproběhl jediný experiment*.
- Data byla připravena předem.
  Přesto se třetí týden ukázalo, že španělština má otazníky místo diakritiky a v češtině chybí půlka dat.

Literatura

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, September.