Target-Side Context for Discriminative Models in Statistical MT

Aleš Tamchyna,
Alexander Fraser,
Ondřej Bojar,
Marcin Junczys-Dowmunt

ACL 2016

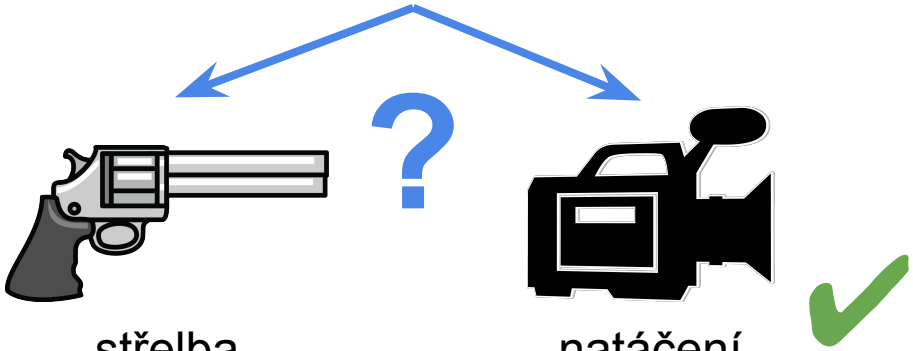
August 9, 2016

Outline

- **Motivation**
- Model Description
- Integration in Phrase-Based Decoding
- Experimental Evaluation
- Conclusion

Why Context Matters in MT: Source

shooting of the expensive film

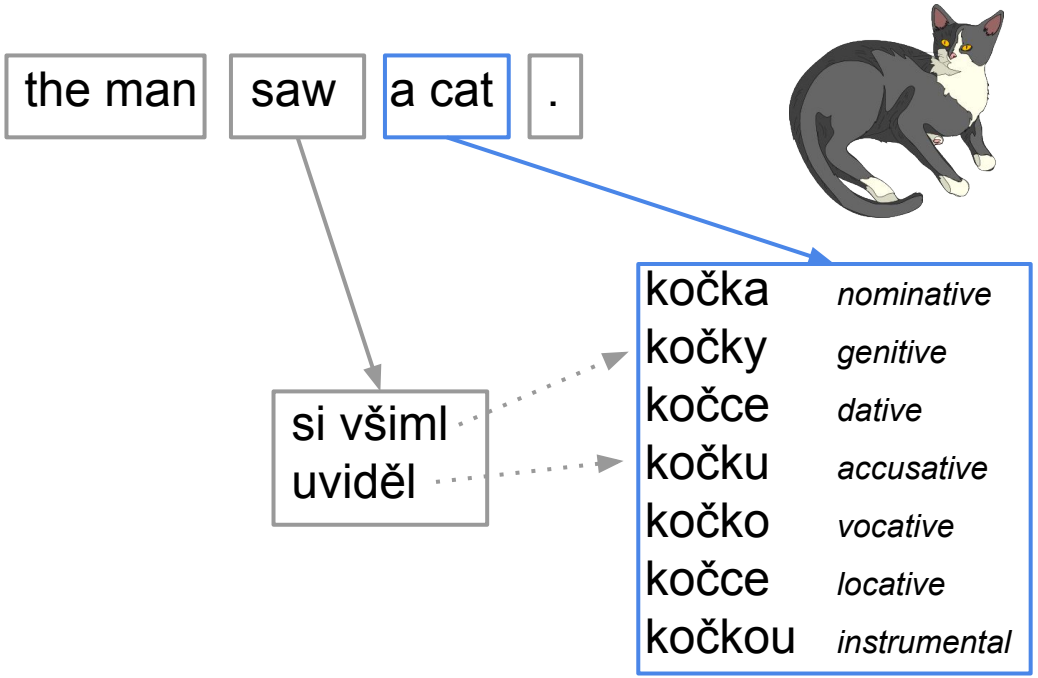


střelba natáčení

Wider **source** context required for disambiguation of word **sense.**

Previous work has looked at using source context in MT.

Why Context Matters in MT: Target



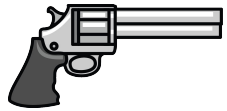
Correct case depends on how we translate the previous words.

Wider **target** context required for disambiguation of word **inflection**.

How Does PBMT Fare?

shooting of the **film** . **natáčení** filmu . ✔

shooting of the expensive **film** . **střelby** na drahý film . ✘



the man saw **a cat** . muž uviděl **kočku**$_{acc}$. ✔

the man saw **a black cat** . muž spatřil **černou**$_{acc}$ **kočku**$_{acc}$. ✔

the man saw **a yellowish cat** . muž spatřil **nažloutlá**$_{nom}$ **kočka**$_{nom}$. ✘

Outline

- Motivation
- **Model Description**
- Integration in Phrase-Based Decoding
- Experimental Evaluation
- Conclusion

A Discriminative Model of Source and Target Context

Let F, E be the source and target sentence.

Model the following probability distribution:

$$P(E|F) \propto \prod_{(\bar{e}_i, \bar{f}_i) \in (E,F)} P(\bar{e}_i | \bar{f}_i, F, e_{prev}, e_{prev-1})$$

target phrase

source phrase

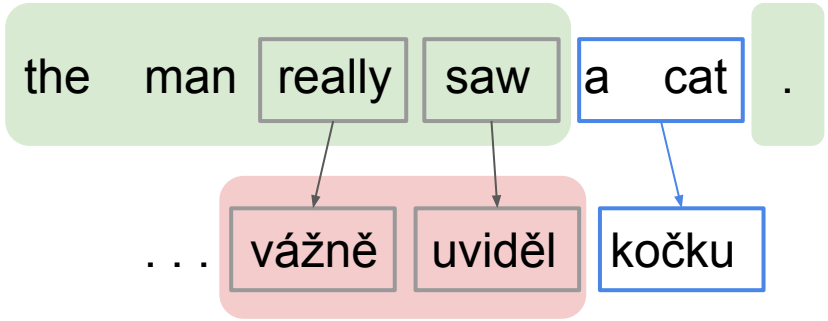
source context *target context*

Where:

$$P(\bar{e}_i | \bar{f}_i, F, e_{prev}, e_{prev-1}) = \frac{\exp(w \cdot \text{fv}(\bar{e}_i, \bar{f}_i, F, e_{prev}, e_{prev-1}))}{\sum_{\bar{e}' \in \text{GEN}(\bar{f}_i)} \exp(w \cdot \text{fv}(\bar{e}', \bar{f}_i, F, e_{prev}, e_{prev-1}))}$$

weight vector *feature vector*

Model Features (1/2)



Label Independent (S = shared):

- source window: *-1^saw -2^really ...*
- source words: *a cat*
- source phrase: *a_cat*

- context window: *-1^uviděl -2^vážně*
- context bilingual: *saw^uviděl really^vážně*

Label Dependent (T = translation):

- target words: *kočku*
- target phrase: *kočku*

Full Feature Set: { S×T ∪ S ∪ T }

cat&kočku ...a_cat&kočku ... saw^uviděl&kočku ... -1^uviděl&kočku ... a_cat ... kočku

Model Features (2/2)

- train a single model where each class is defined by label-dependent features
- **source:** form, lemma, part of speech, dependency parent, syntactic role
- **target:** form, lemma, (complex) morphological tag (e.g. `NNFS1-----A----`)
- Allows to learn e.g.:
  - subjects (role=Sb) often translate into nominative case
  - nouns are usually accusative when preceded by an adjective in accusative case
  - lemma “cat” maps to lemma “kočka” regardless of word form (inflection)

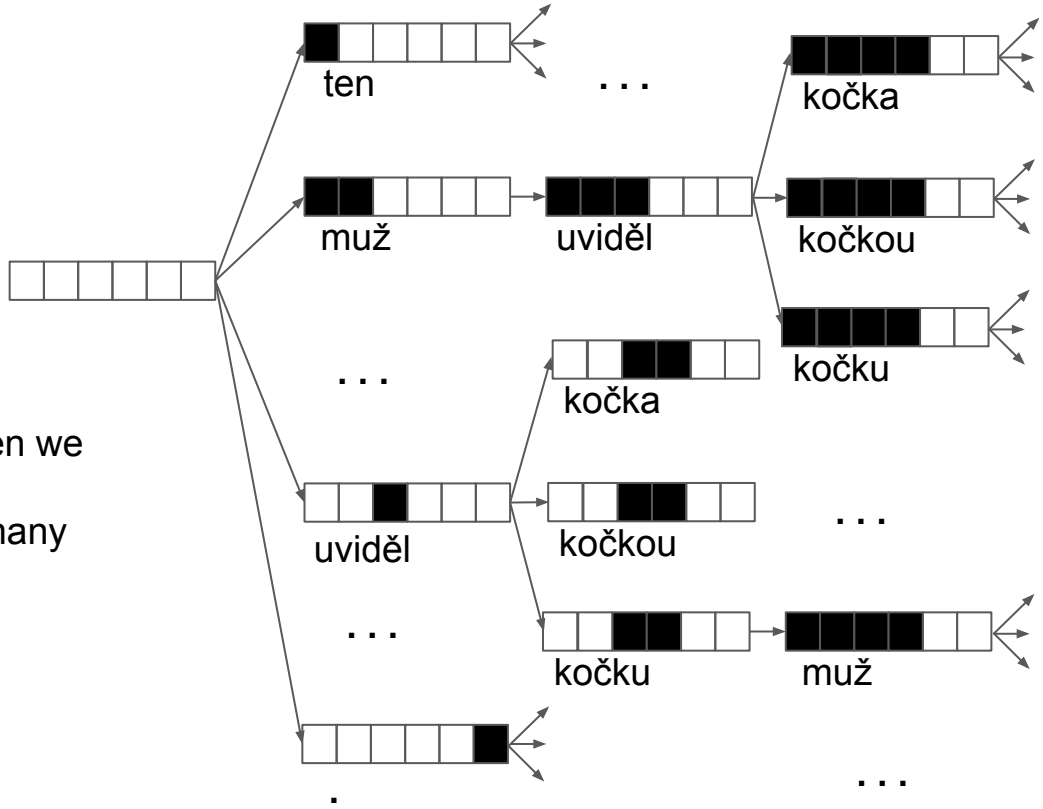
Outline

- Motivation
- Model Description
- **Integration in Phrase-Based Decoding**
- Experimental Evaluation
- Conclusion

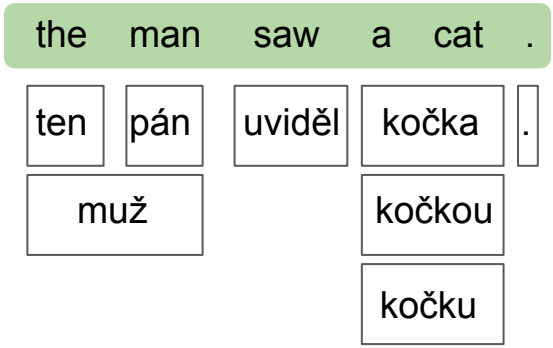
Challenges in Decoding

the	man	saw	a	cat	.
ten	pán	uviděl	kočka	.	
muž			kočkou		
			kočku		

- **source** context remains constant when we decode a single sentence
- each translation option evaluated in many different **target** contexts
  - as many as a language model

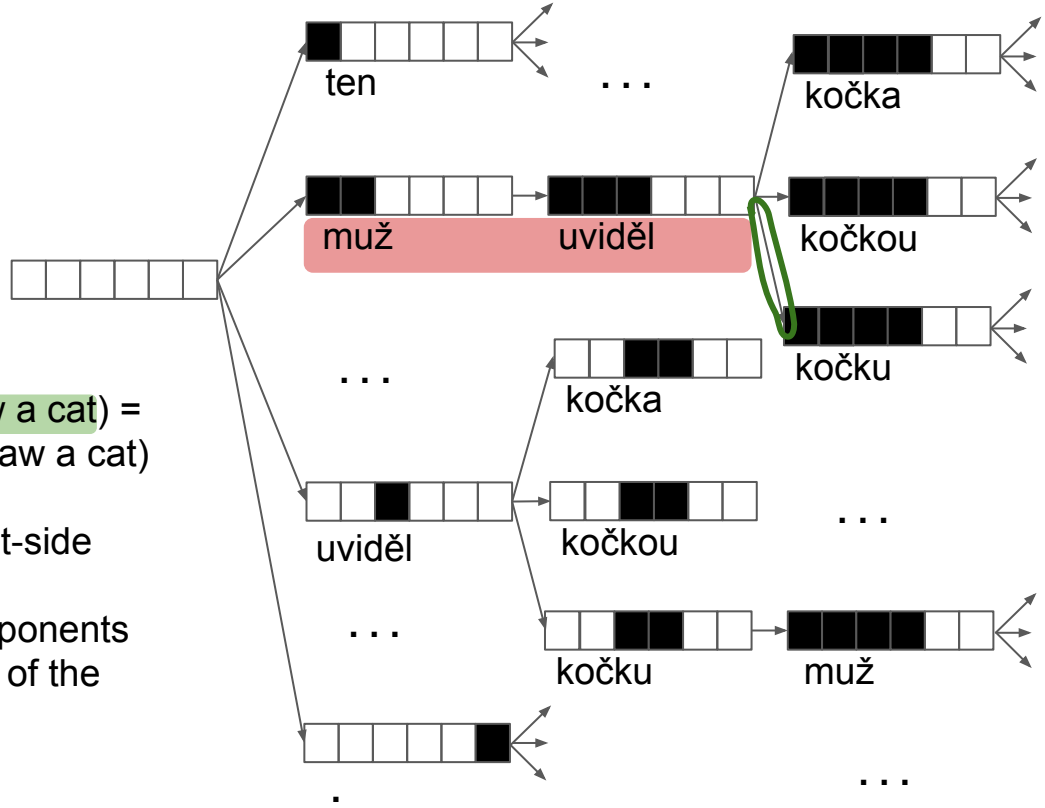


Trick #1: Source- and Target-Context Score Parts



score(kočku|muž uviděl, a cat, the man saw a cat) =
w · fv(kočku, muž uviděl, a cat, the man saw a cat)

- most features do not depend on target-side context “muž uviděl”
- divide the feature vector into two components
- pre-compute source-context only part of the score before decoding



Tricks #2 and #3

- **Cache feature vectors**
  - each **translation option** (“kočku”) will be seen multiple times during decoding
    - cache its feature vector before decoding
  - **target-side contexts** repeat within a single search (“muž uviděl” -> *)
    - cache context features for each new context
- **Cache final results**
  - pre-compute and store scores for all possible translations of the current phrase
    - needed for normalization anyway

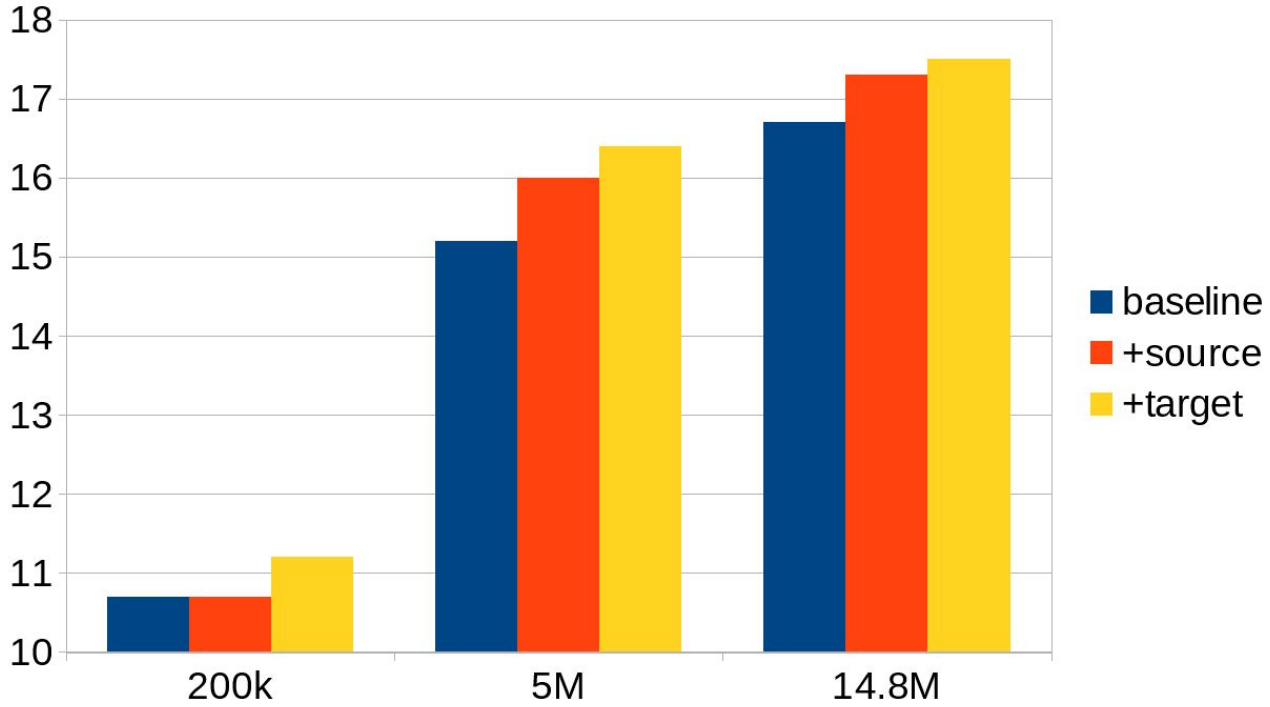
Evaluation of Decoding Speed

Integration	Avg. Time per Sentence
baseline	0.8 s
naive: only #3	13.7 s
+tricks #1, #2	2.9 s

Outline

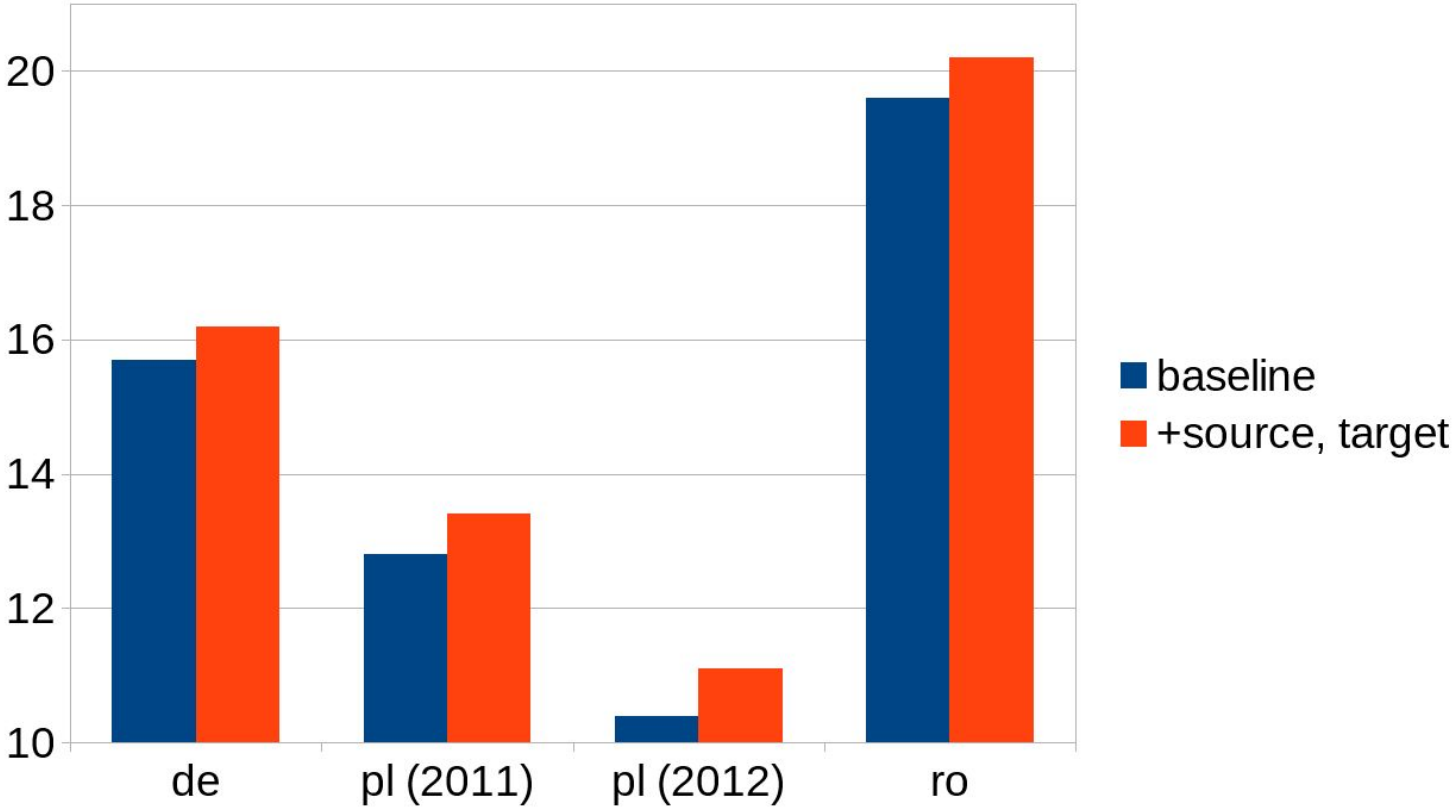
- Motivation
- Model Description
- Integration in Phrase-Based Decoding
- **Experimental Evaluation**
- Conclusion

Scaling to Large Data



- BLEU scores, English-Czech translation
- training data: subsets of CzEng 1.0

Additional Language Pairs



Manual Evaluation

- blind evaluation of system outputs, 104 random test sentences
- English-Czech translation
- sample BLEU scores: 15.08, 16.22, 16.53

Setting	**Equal**	**Baseline is better**	**New is better**
baseline vs. +source	52	26	26
baseline vs. +target	52	18	**34**

Conclusion

- novel discriminative model for MT that uses both source- and target-side context information
- (relatively) efficient integration directly into MT decoding
- significant improvement of BLEU for English-Czech even on large-scale data
- consistent improvement for three other language pairs
- model freely available as part of the Moses toolkit

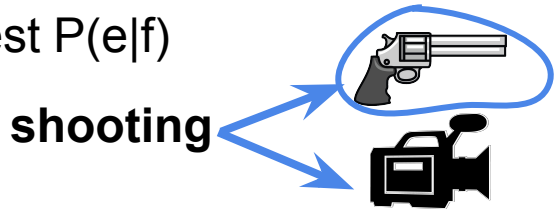
Thank you!

Questions?

Extra slides

Intrinsic Evaluation

- the task: predict the correct translation in the current context
- baseline: select the most frequent translation from the candidates, i.e., translation with the highest P(e|f)



- English-Czech translation, tested on WMT13 test set

Model	**Accuracy**
baseline	51.5
+source context	66.3
+target context	74.8*

Model Training: Parallel Data

gunmen fled after the shooting .	pachatelé po střelbě uprchli .
...	
shooting of an expensive film .	natáčení drahého filmu .
...	
the director left the shooting .	režisér odešel z natáčení .
...	
the man saw a black cat .	muž viděl černou|A4 kočku|N4 .
...	
the black cat noticed the man .	černá|A1 kočka|N1 viděla muže .

Training examples:

\+ střelbě&gunmen střelbě&fled ...
\- natáčení&gunmen natáčení&fled ...

\- střelbě&film střelbě&expensive ...
\+ natáčení&film natáčení&fled ...

\- střelbě&director střelbě&left ...
\+ natáčení&director natáčení&left ...

\- prev=A4&N1 prev=A4&kočka ...
\+ prev=A4&N4 prev=A4&kočku ...

\+ prev=A1&N1 prev=A1&kočka ...
\- prev=A1&N4 prev=A1&kočku ...

Model Training

- Vowpal Wabbit
- quadratic feature combinations generated automatically
- objective function: logistic loss
- setting: `--csoaa_ldf mc`
- 10 iterations over data
  - select best model based on held-out accuracy
- no regularization

Training Efficiency

- huge number of features generated (hundreds of GBs when compressed)
- feature extraction
  - easily parallelizable task: simply split data into many chunks
  - each chunk processed in a multithreaded instance of Moses
- model training
  - Vowpal Wabbit is fast
  - training can be parallelized using VW AllReduce
  - workers train on independent chunks, share parameter updates with a master node
  - linear speed-up
  - 10-20 jobs

Additional Language Pairs (1/2)

- English-German
  - parallel data: 4.3M sentence pairs (Europarl + Common Crawl)
  - dev/test: WMT13/WMT14
- English-Polish
  - not included in WMT so far
  - parallel data: 750k sentence pairs (Europarl + WIT)
  - dev/test: IWSLT sets (TED talks) 2010, 2011, 2012
- English-Romanian
  - included only in WMT16
  - parallel data: 600k sentence pairs (Europarl + SETIMES2)
  - dev/test: WMT16 dev test, split in half

LMs over Morphological Tags

- a stronger baseline: add LMs over tags for better morphological coherence
- do our models still improve translation?
- 1M sentence pairs, English-Czech translation

System	BLEU
baseline	13.0
+tag LM	14.0
+source	14.5
+target	14.8

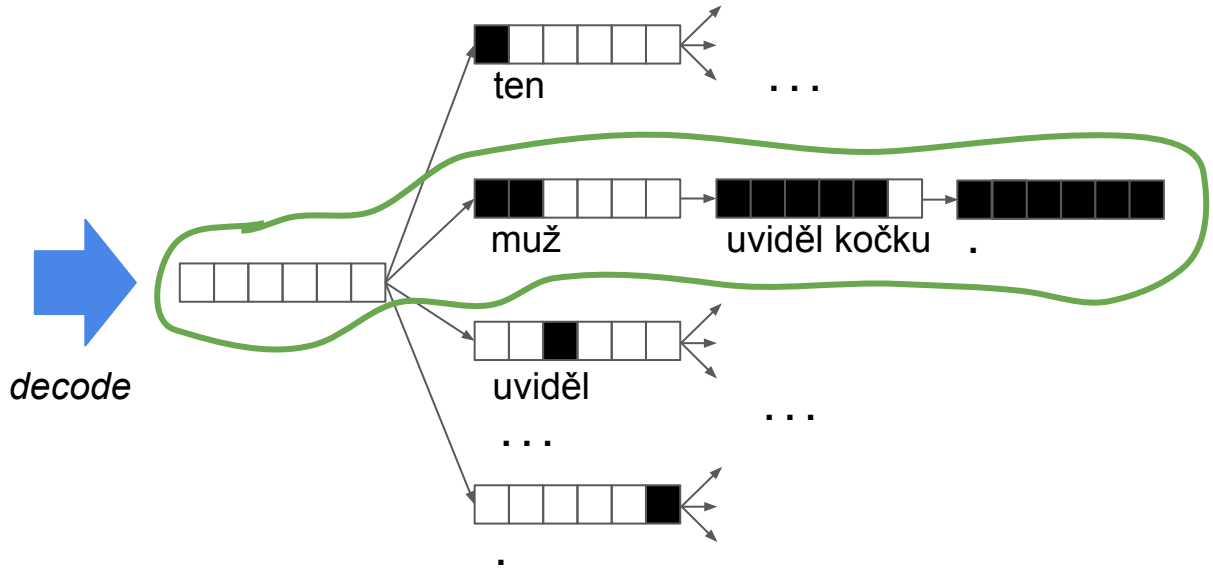
Phrase-Based MT: Quick Refresher

the man saw a cat .

query phrase table

the	man	saw	a	cat	.
ten	pán	uviděl	kočka	.	
muž			kočkou		
		uviděl kočku			

decode



P_{LM} = P(muž|<s>) · P(uviděl kočku | <s> muž) · ... · P(</s> | kočku .)

System Outputs: Example

input: the most intensive mining took place there from 1953 to 1962 .

baseline: nejvíce intenzivní těžba **došlo** tam z roku 1953 **, aby** 1962 .

the_most intensive mining$_{nom}$ ***there_occurred*** *there from 1953* ***, in_order_to*** *1962 .*

+source: nejvíce intenzivní **těžby místo** tam z roku 1953 do roku 1962 .

the_most intensive ***mining***$_{gen}$ ***place*** *there from year 1953 until year 1962 .*

+target: nejvíce intenzivní těžba probíhala od roku 1953 do roku 1962 .

the_most intensive mining$_{nom}$ *occurred from year 1953 until year 1962 .*