Target-Side Context for Discriminative Models in Statistical MT

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

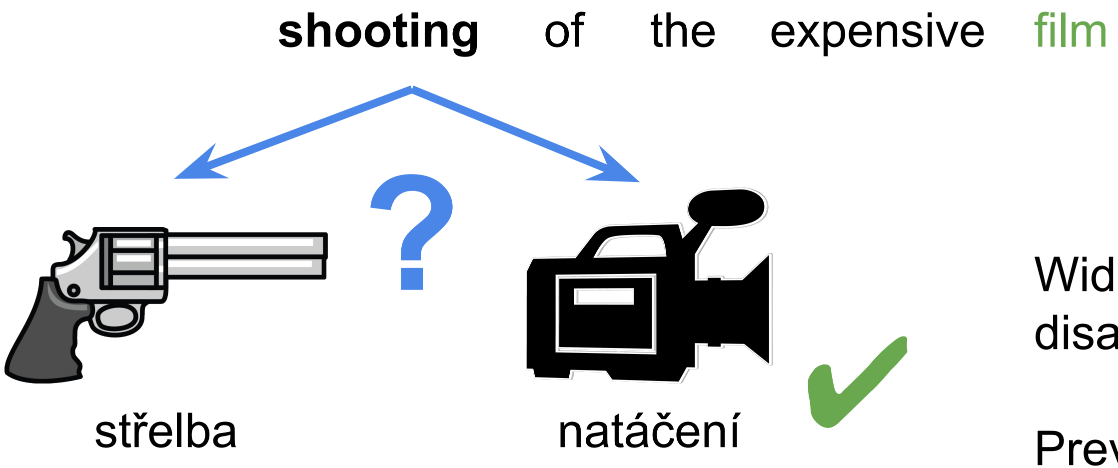
CIS Talk

June 28, 2016

Outline

- **Motivation**
- Model Description
- Integration in Phrase-Based Decoding
- Experimental Evaluation
- Analysis, Discussion

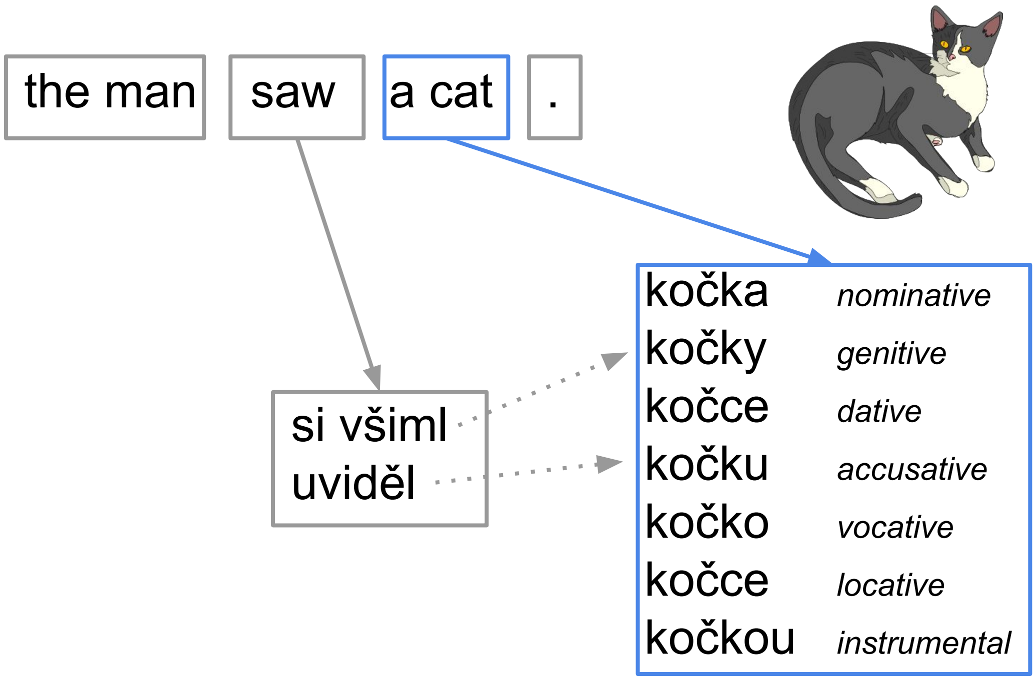
Why Context Matters in MT: Source



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

Previous work has looked at using source context in MT.

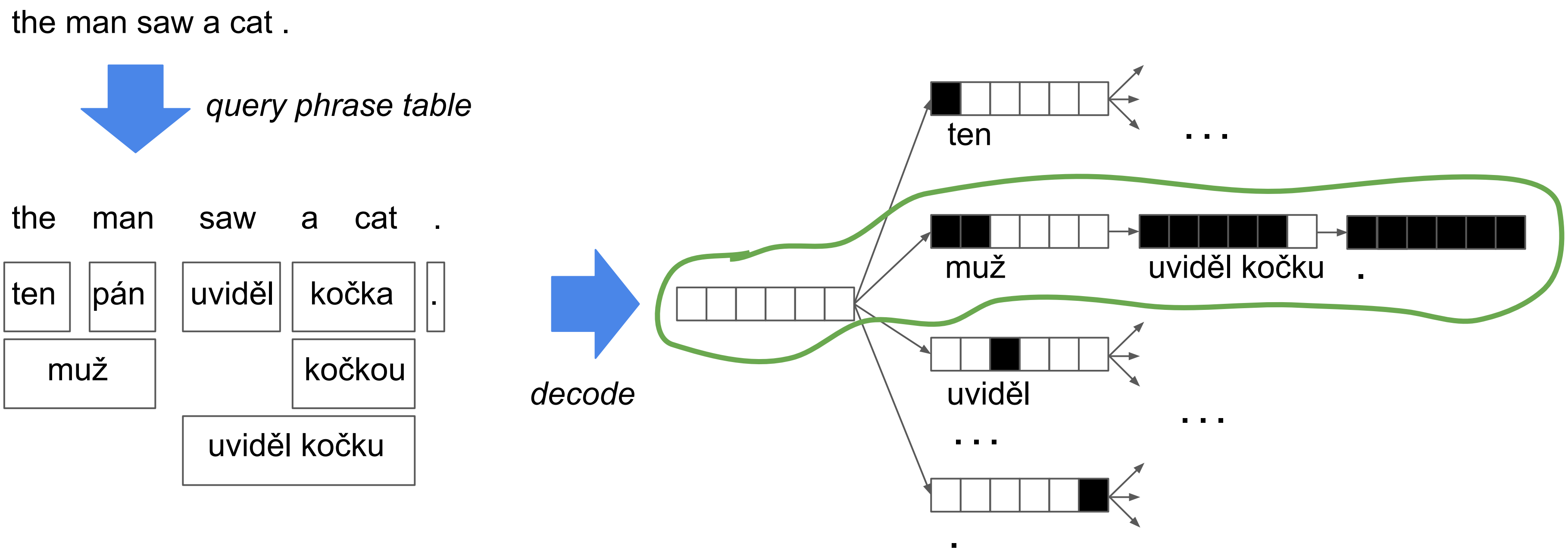
Why Context Matters in MT: Target



Correct case depends on how we translate the previous words.

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

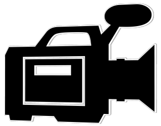
Phrase-Based MT: Quick Refresher



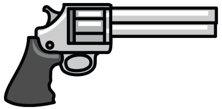
$P_{LM} = P(\text{muž}|\langle s\rangle) \cdot P(\text{uviděl kočku} \mid \langle s\rangle\ \text{muž}) \cdot \ldots \cdot P(\langle /s\rangle \mid \text{kočku .})$

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
- Analysis, Discussion

A Discriminative Model of Source and Target Context

Let F, E be the source and target sentence.

Model the following probability distribution:

$$P(E|F) \propto \prod_{(\bar{e_i},\bar{f_i})\in(E,F)} P(\bar{e_i}|\bar{f_i}, F, e_{prev}, e_{prev-1})$$

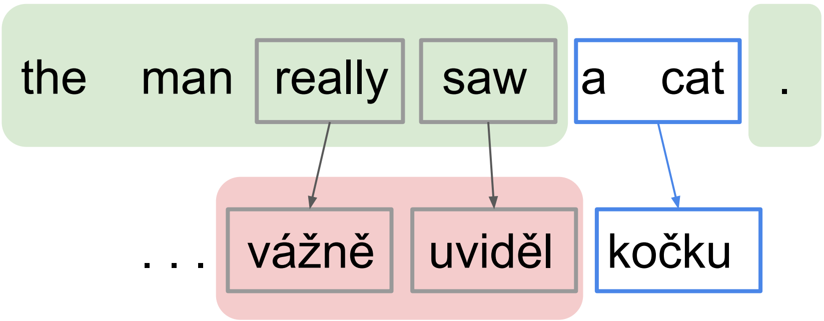
target phrase: $\bar{e_i}$; source phrase: $\bar{f_i}$; source context: F; target context: e_{prev}, e_{prev-1}

Where:

weight vector: w; feature vector: fv

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

Model Features (1/2)



Label Independent (S = shared):

- source window: *swin-1^saw swin-2^really ...*
- source internal: *sin^a sin^cat*
- source indicator: *sind^a_cat*

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

Label Dependent (T = translation):

- target internal: *tin^kočku*
- target indicator: *tind^kočku*

Full Feature Set: { S×T U S U T }

sin^cat&tin^kočku ... sind^a_cat&tind^kočku ... blng^saw^uviděl&tind^kočku ... tcwin-1^uviděl&tind^kočku ... sind^a_cat ... tind^kočku

Model Features (2/2)

- train a single model where each class is defined by label-dependent features
- our feature set is richer than surface forms
- **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)

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

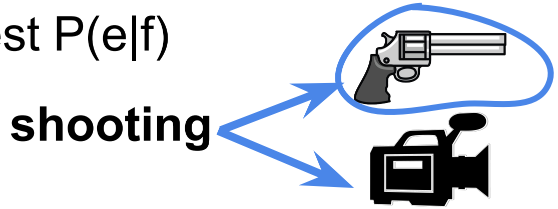
- Vowpal Wabbit
- quadratic feature combinations generated automatically
- objective function: logistic loss
- setting: `--csoaa_ldf mc`
- 10 iterations over data
    - select best model based on held-out accuracy
- no regularization

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

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)



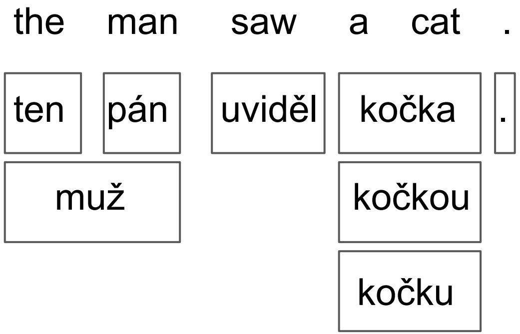
- English-Czech translation, tested on WMT13 test set

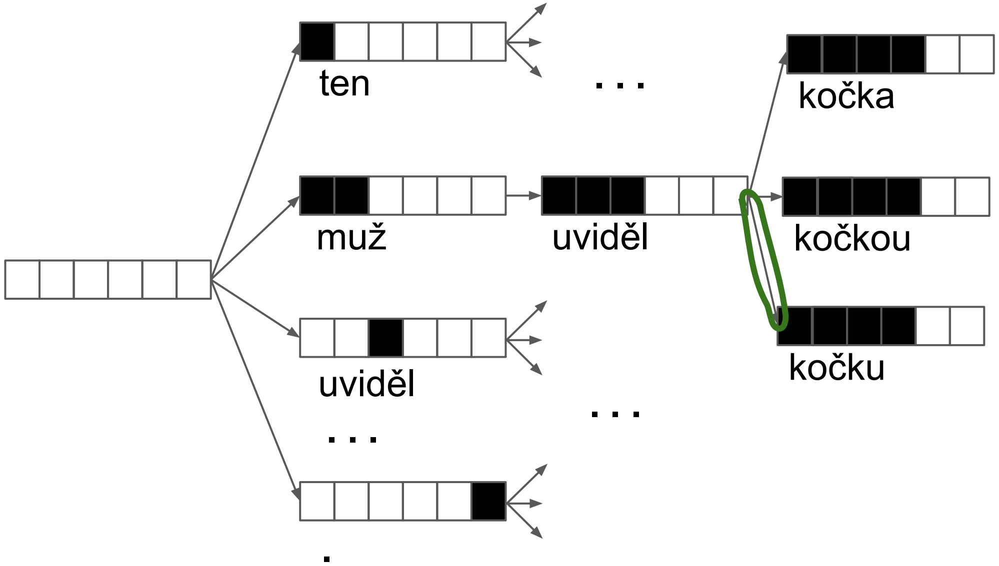
Model	Accuracy
baseline	51.5
+source context	66.3
+target context	74.8*

Outline

- Motivation
- Model Description
- **Integration in Phrase-Based Decoding**
- Experimental Evaluation
- Analysis, Discussion

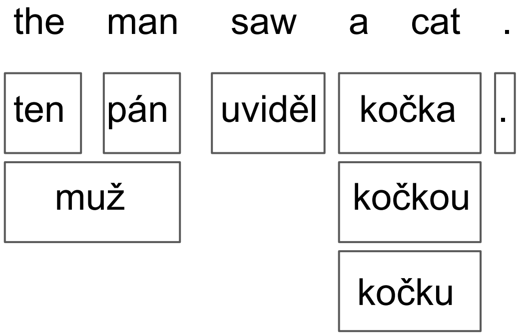
Decoding with the Context Model

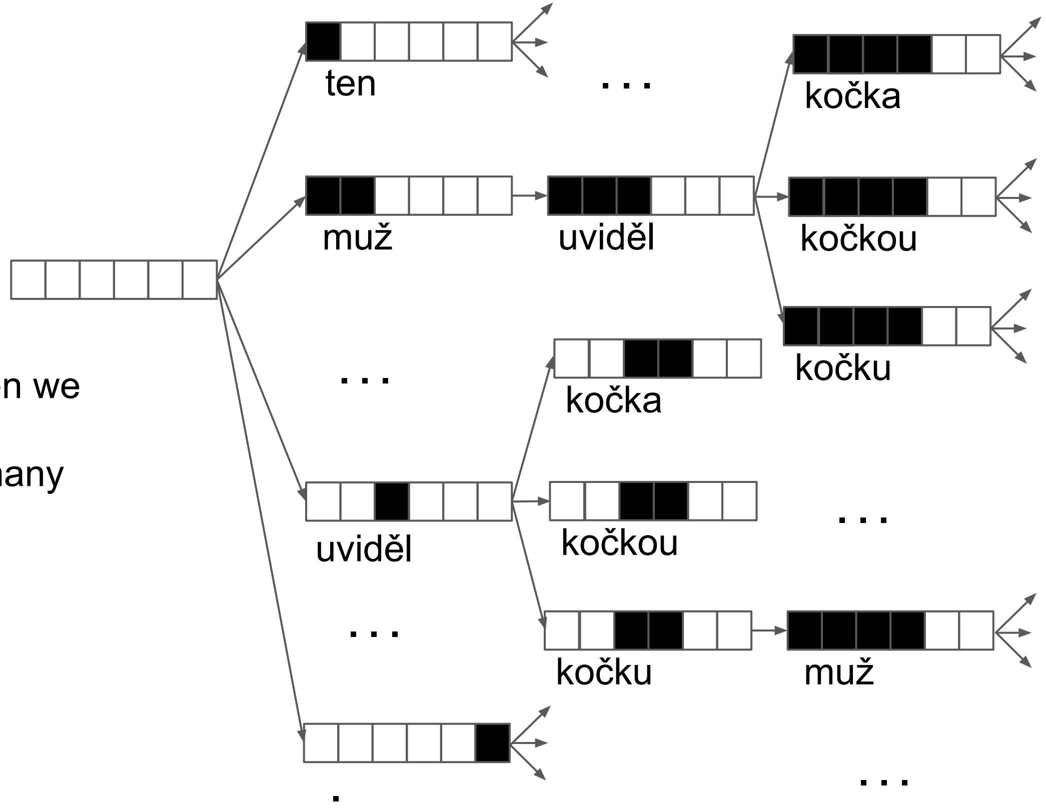




$$P_{disc}(\text{kočku}|\text{muž uviděl}, \text{cat}, \text{the man saw a cat}) = \frac{\exp(\text{score}(\text{kočku}|\text{muž uviděl}, \text{cat}, \text{the man saw a cat}))}{\sum_{i \in \left\{ \substack{\text{kočka,} \\ \text{kočkou,} \\ \text{kočku}} \right\}} \exp(\text{score}(i|\text{muž uviděl}, \text{cat}, \text{the man saw a cat}))}$$

Challenges in Decoding

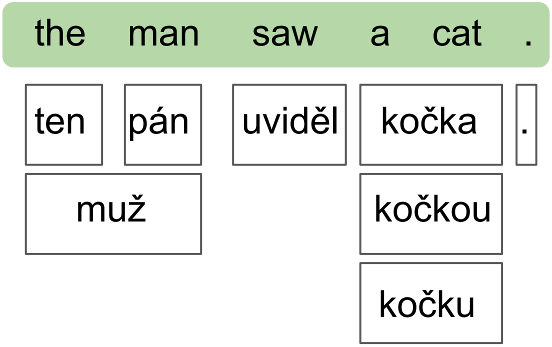


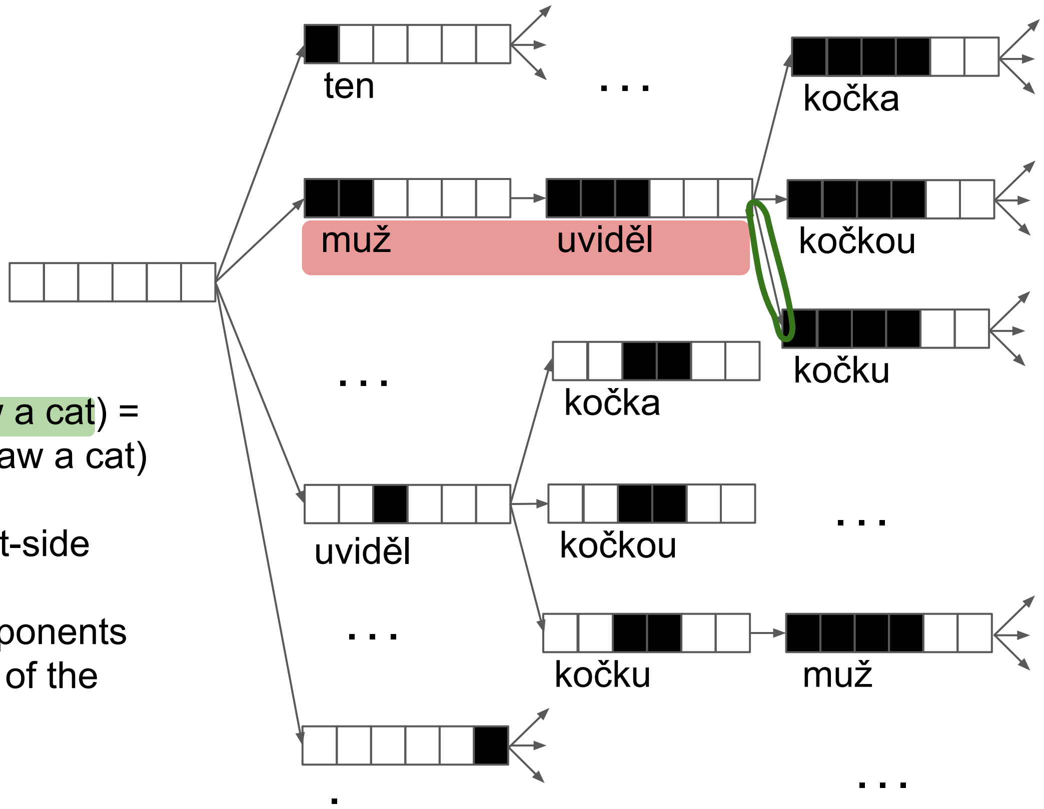


- **source** context remains constant when we decode a single sentence
- each translation option evaluated in many different **target** contexts
  - as many as a language model

- normalization: evaluate all possible translations of a given phrase to get a conditional probability distribution

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

Trick #2: Cache Feature Vectors

- each translation option ("kočku") will be seen multiple times during decoding, in different contexts
  - generate features T (internal to the translation option) before decoding
  - get back feature hashes from VW
  - store them in cache as integer arrays for future use
- target-side contexts repeat within a single search ("muž uviděl" -> *)
  - generate features S_{tgt} the first time we see that particular context during decoding
  - store their hashes in cache

Trick #3: Cache Final Results

- our score needs to be locally normalized
  - over possible translations of the current source phrase in the current context
- compute score for all possible translations at once, normalize
- store scores of all translations in cache
  - the decoder will probably evaluate other translations as well 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

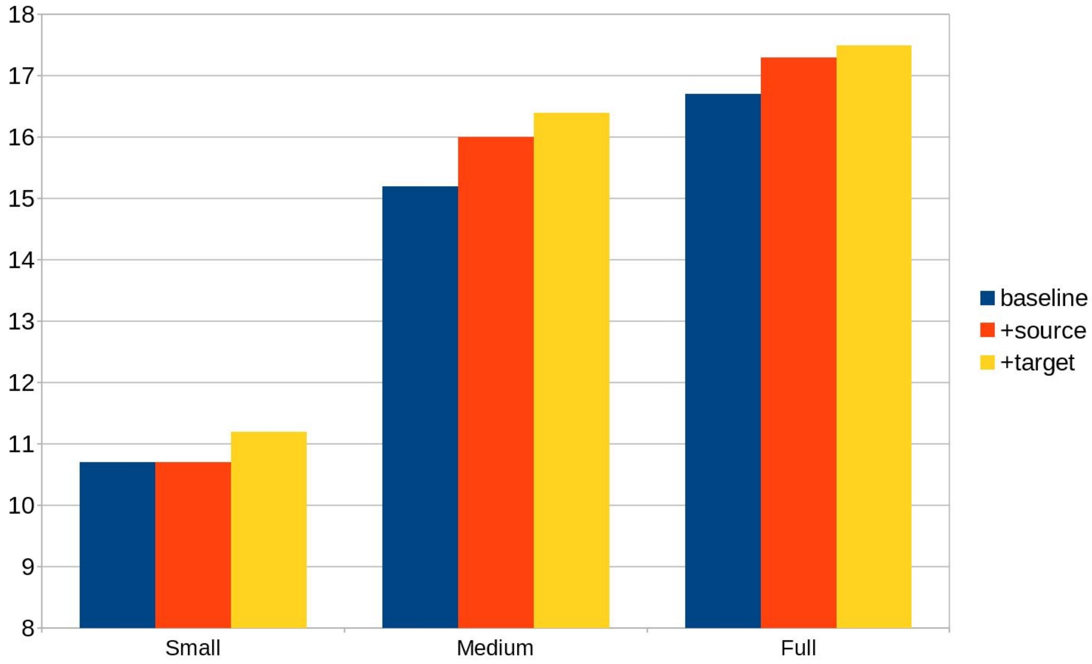
Outline

- Motivation
- Model Description
- Integration in Phrase-Based Decoding
- **Experimental Evaluation**
- Analysis, Discussion

Scaling to Large Data (1/2)

- not clear whether discriminative models help when large parallel data is available
- English-Czech translation, train on subsets of CzEng 1.0
- 5-gram LM, tune on WMT13, test on WMT14
- system variants:
  - baseline
  - +source
  - +target
- settings:
  - small -- 200k
  - medium -- 5M
  - full -- 14.8M

Scaling to Large Data (2/2)

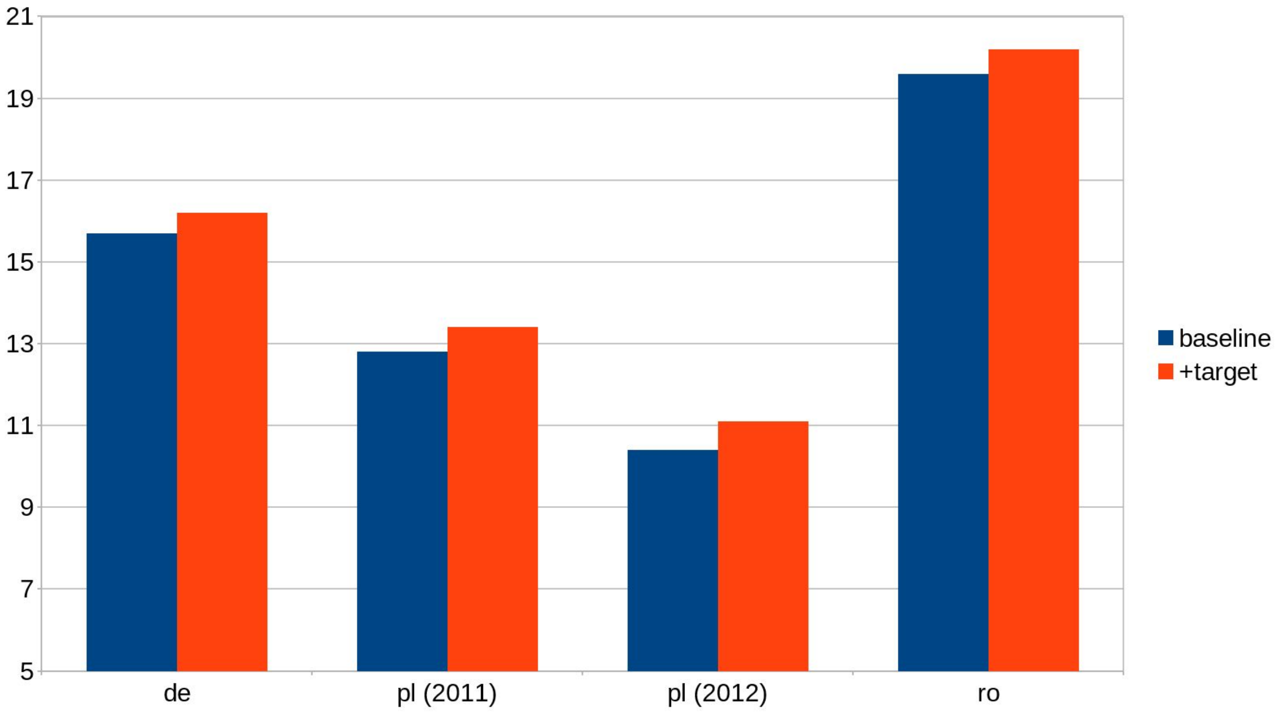


- BLEU scores on WMT14
- average over 5 independent optimization runs

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

Additional Language Pairs (2/2)



- average test BLEU over 5 independent optimization runs

Outline

- Motivation
- Model Description
- Integration in Phrase-Based Decoding
- Experimental Evaluation
- **Analysis, Discussion**

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	1 is better	2 is better
baseline vs. +source	52	26	26
baseline vs. +target	52	18	**34**

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 .* ✔

System Outputs: Discussion

- source-context model improves:
  - semantics
  - often also morphology and syntax
- target-context helps overall agreement and coherence on top of the source-context model

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
- improvement consistent for three other language pairs
- model freely available as part of the Moses toolkit

Thank you!

Questions?