Maximum Entropy Translation Model in Dependency-Based MT Framework

David Mareček, Martin Popel, Zdeněk Žabokrtský

Charles University in Prague
Institute of Formal and Applied Linguistics

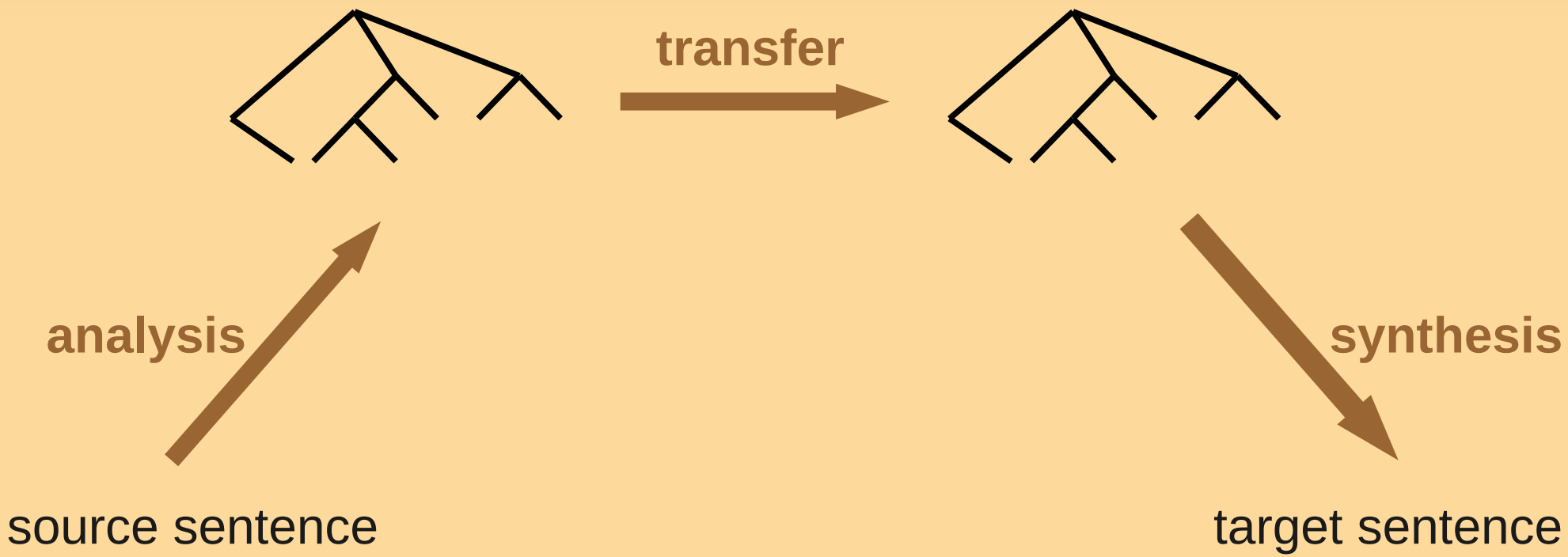
PIRE meeting
July 15, 2010, Uppsala, Sweeden

Outline

- **TectoMT system introduction**
- Translation dictionaries
- Experiments and results

TectoMT

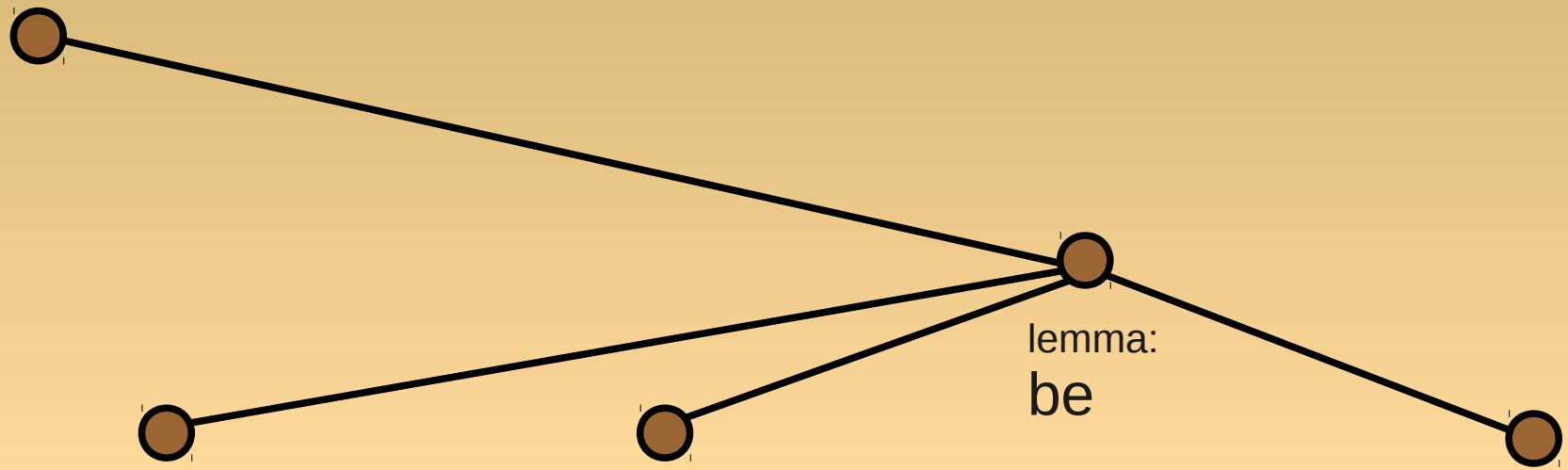
- Analysis-transfer-synthesis translation system
  - Transfer on the level of deep-syntax (tectogrammatical trees)



Tectogrammatical tree

- Dependency tree, where only content words have their own nodes
- Other words (function words) are expressed within the respective content nodes in the form of their attributes
  - Function words: articles, prepositions, auxiliary verbs, modal verbs, punctuation marks, ...
- Three main attributes of the nodes we need:
  - T-lemma
  - Formeme - surface morphosyntactic form of the node
  - Grammatemes - morphological categories

Tectogrammatical tree - example



lemma:
#Perspron

formeme:
n:subj

grammatemes:
gender=fem
number=sg
person=3

lemma:
never

formeme:
adv:

grammatemes:
degcmp=pos

lemma:
be

formeme:
v:fin

grammatemes:
tense=ant
negation=0

lemma:
Sydney

formeme:
n:to+X

grammatemes:
number=sg

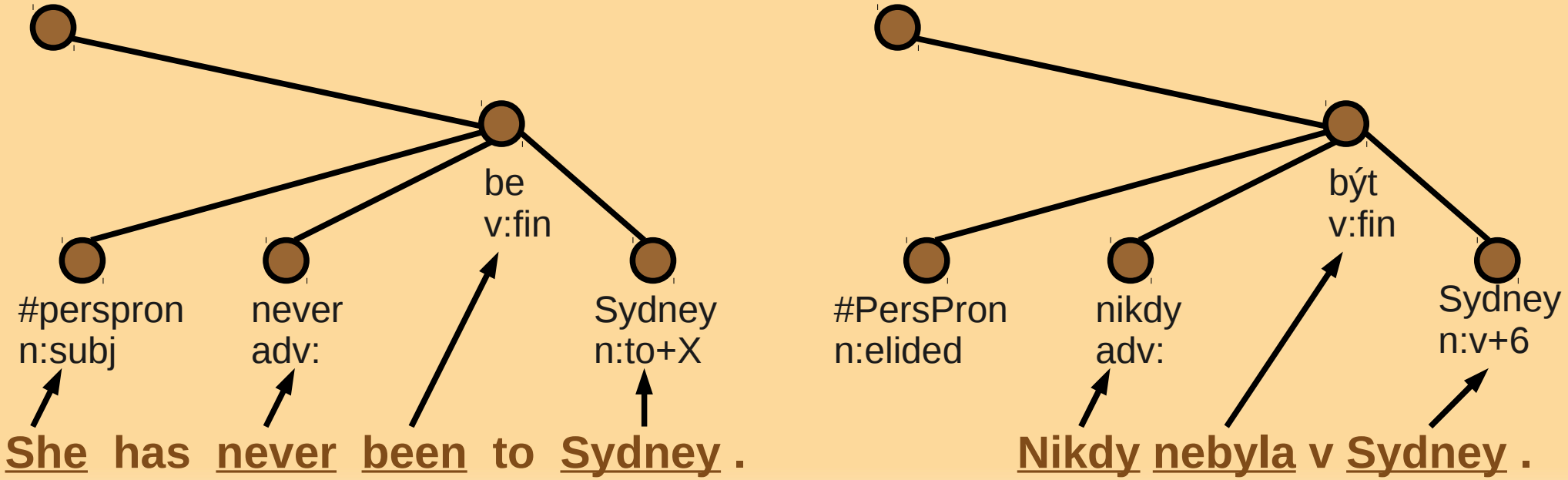
She has never been to Sydney .

Tectogrammatical attributes

- T-lemma
  - mother, read, #PersPron, take_off, put_on, look_after
- Formeme – surface morphosyntactic form of the node
  - English: n:subj, n:obj, n:of+X, n:x+ago, adj:attr, v:fin, adv:, …
  - Czech: n:1, n:2, n:na+4, v:inf, adj:attr, …
- Grammatemes – necessary morphological categories
  - gender, number, person, tense, verb modality, degree of comparison, ...

Why is tectogrammatics good for transfer

- Tectogrammatical trees of corresponding Czech and English sentences are much more similar than their surface shapes
  - Contain content words only
  - Contain also entities elided on surface

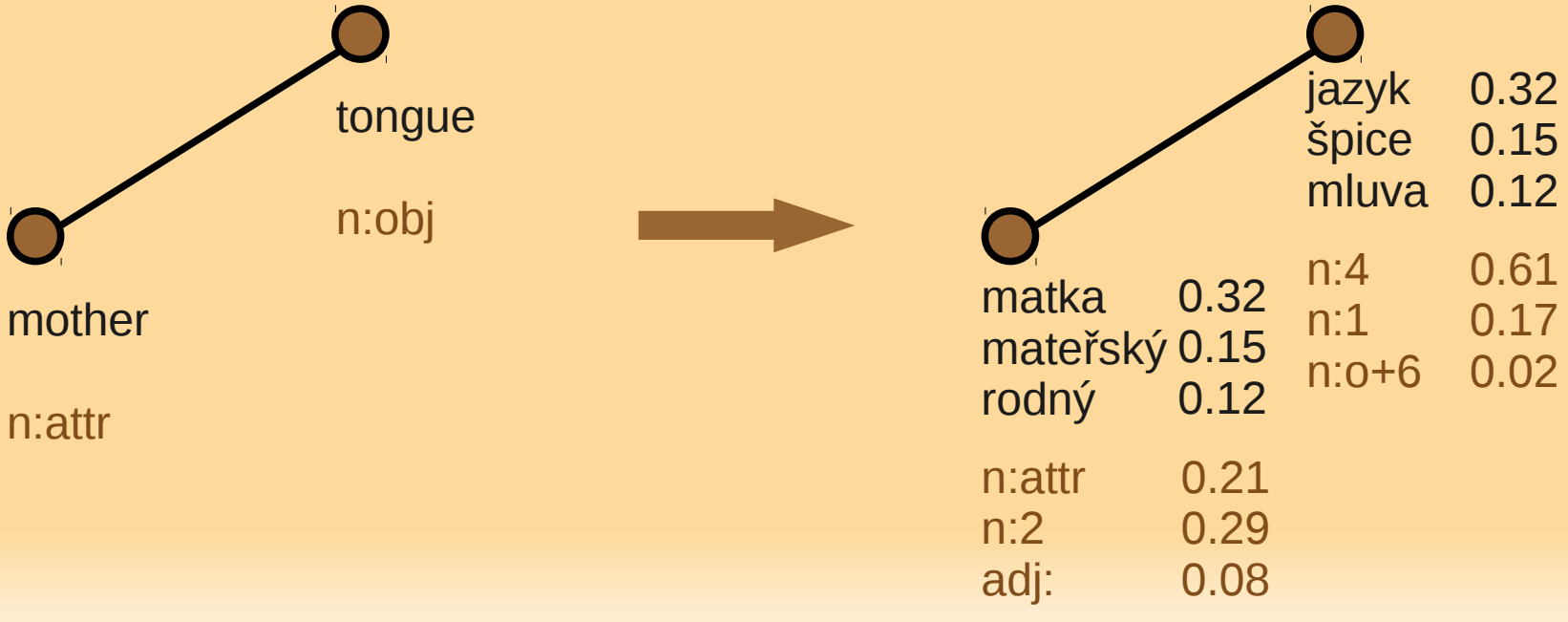


Transfer on tectogrammatics pros & cons

- We can assume that the target structure will be the same (isomorfic) as the source structure
  - than we can simply translate each node in 1:1 manner
  - Of course, there exists 1:2, 2:1 and other mapings
    - ice cream = zmrzlina
    - mother in law = tchýně
  - There are not much of them. These mapings can be solved separately by a special dictionary
- When evaluated, only 8 % of errors were caused by this assumption of isomorfism

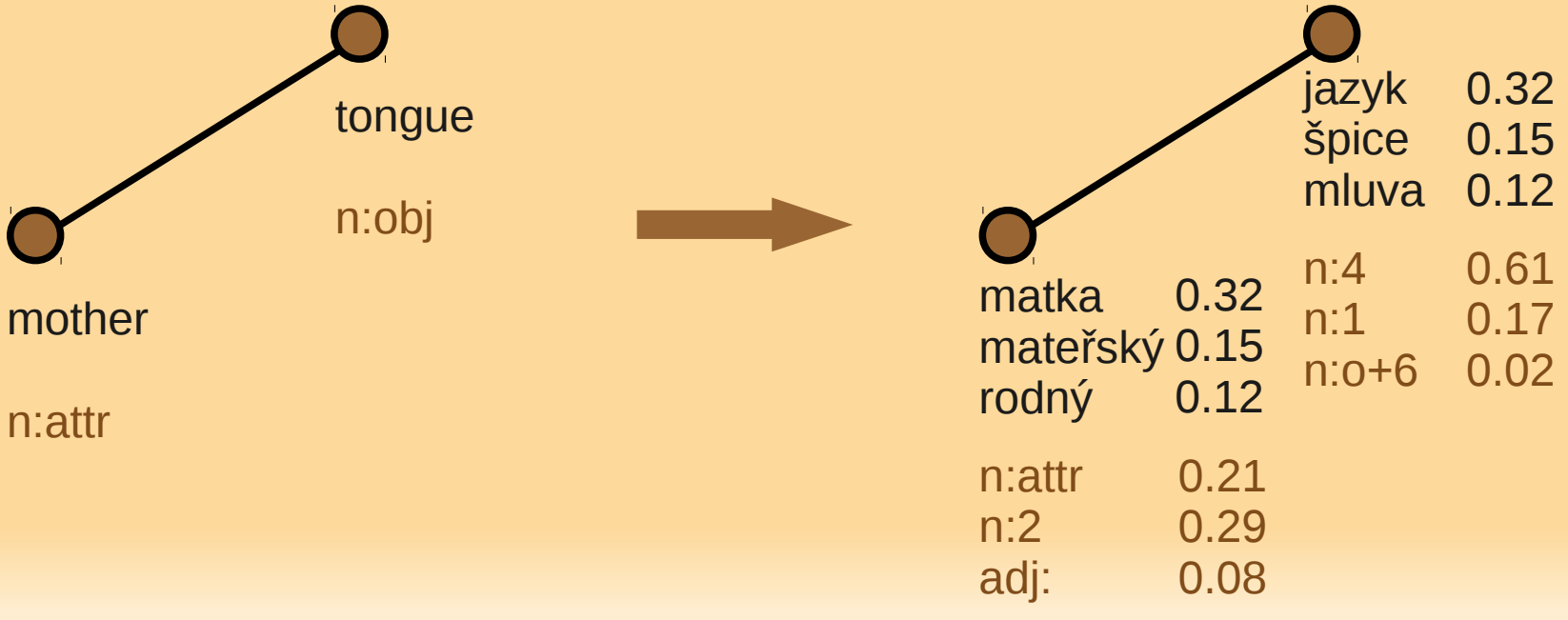
The transfer process

- The topology of the tree and grammatemes (such as person, number, gender, tense, etc.) are preserved
- For each node, its t-lemma and formeme are translated separately
  - For each t-lemma/formeme the set of translation variants is generated from dictionaries



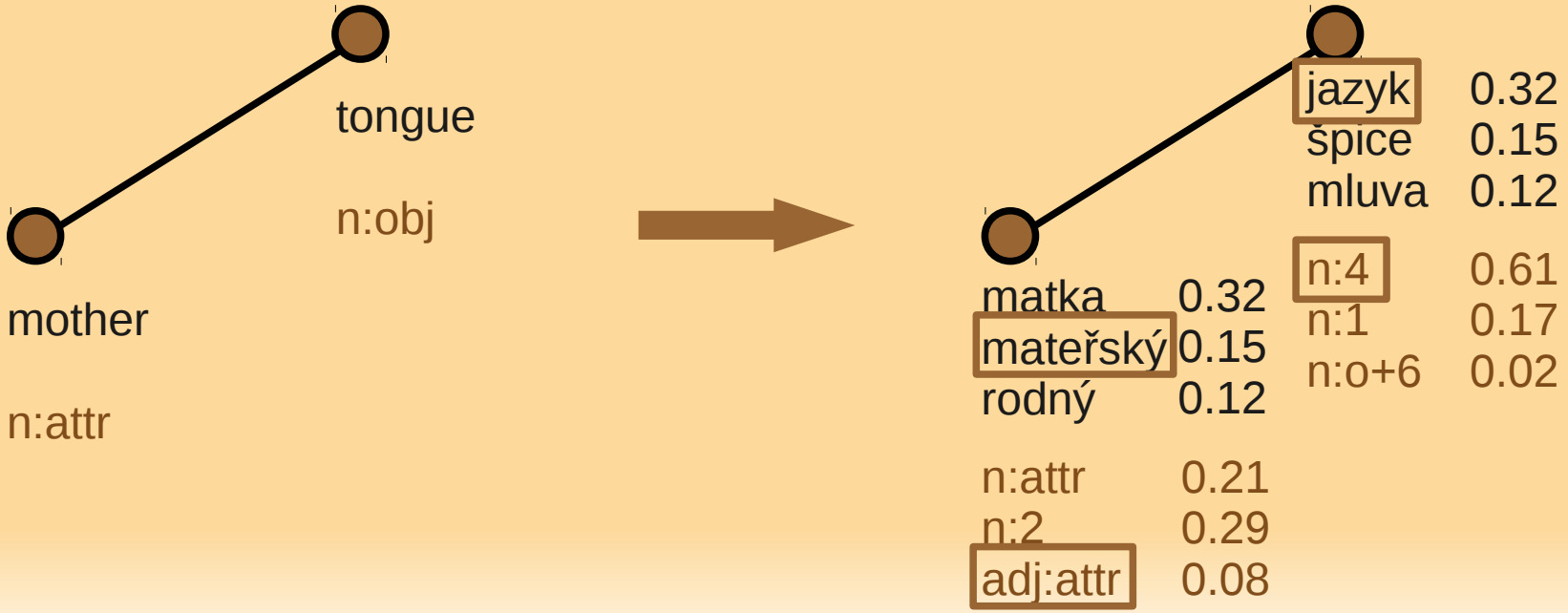
The transfer process

- Each translation variant has a probability assigned from dictionaries
- The translation variants are pruned



The transfer process

- The optimal combination of lemmas and formemes is chosen using TreeLM
  - Hidden-Tree-Markov-Models (HMTM)
  - Viterbi search



Outline

- TectoMT system introduction
- **Translation dictionaries**
- Experiments and results

Translation dictionaries

- For a given source lemma/formeme it returns a set of tranlation hypotheses (with probabilities)
- Translation of lemmas:
  - Static dictionary
  - Context (MaxEnt) dictionary
  - Derivative dictionary (rule-based)
- Translation of formemes:
  - Static dictionary
  - Context (MaxEnt) dictionary

Static dictionary

- Simple dictionary extracted from aligned parallel treebank
  - Extracted all aligned pairs of Czech and English nodes
  - Maximum likelihood estimation
- p(c|e) = count(c,e) / count(e)
- Used both for tranlsation of lemmas and formemes

Context (MaxEnt) dictionary

- Uses also the context of the source node and other attributes

$$p(y|x) = \frac{1}{Z(x)} exp \sum_i \lambda_i f_i(x, y)$$

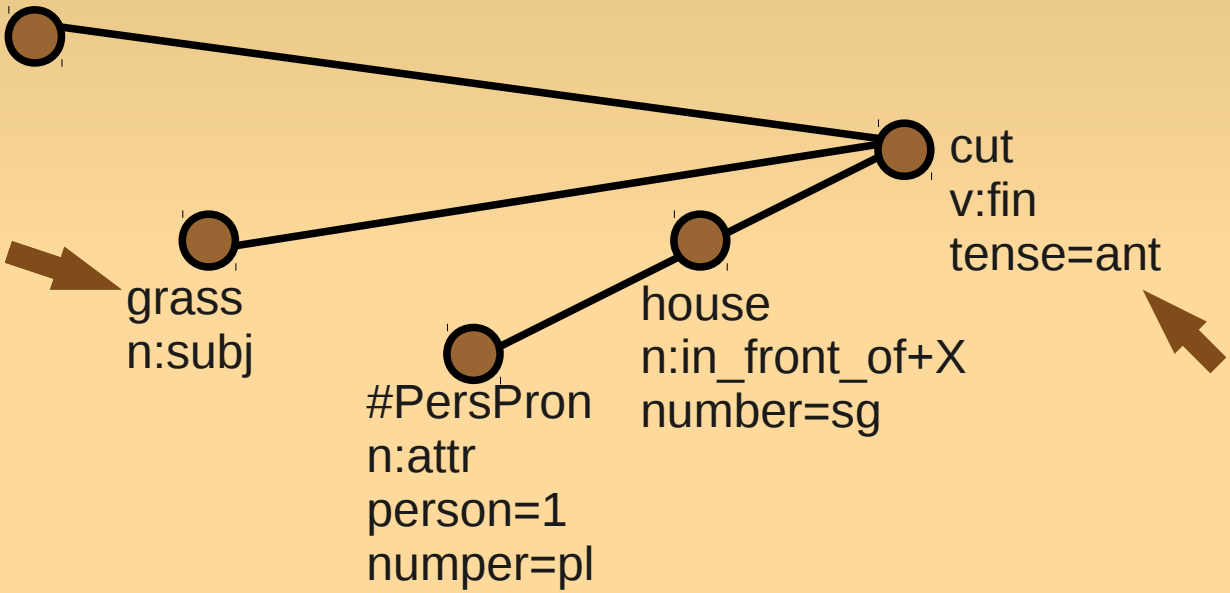
- One MaxEnt model is trained for each source lemma
- Source context features used (x):
  - Local tree context
  - Local linear context
  - Morphological and syntactic categories

Context (MaxEnt) dictionary

- Examples of features:
  - Tense of governing node = past
  - Lemma of the previous node = „cut“
  - Child node formeme = „n:by+X“
  - Has left child = 1

MaxEnt dictionary - example

- The grass in front of our house has been **cut**.



CUT:
kácet (trees)
sekat (grass)
řezat (flowers)
stříhat (hair)
holit (beard)
krájet (bread)
zlevnit (prices)
…

present perfect tense → posekat
present simple → sekat

source_lemma="cut" & child_lemma="grass" & tense="ant"
→ target_lemma="**posekat**"

Derivative dictionaries

- Translation of unknown words using their derivation
  - Translation of adverbs through adjectives:
    - interestingly → interesting → zajímavý → zajímavě
  - Translation of adjectives through verbs
    - translatable → translate → přeložit → přeložitelný

Derivative dictionaries

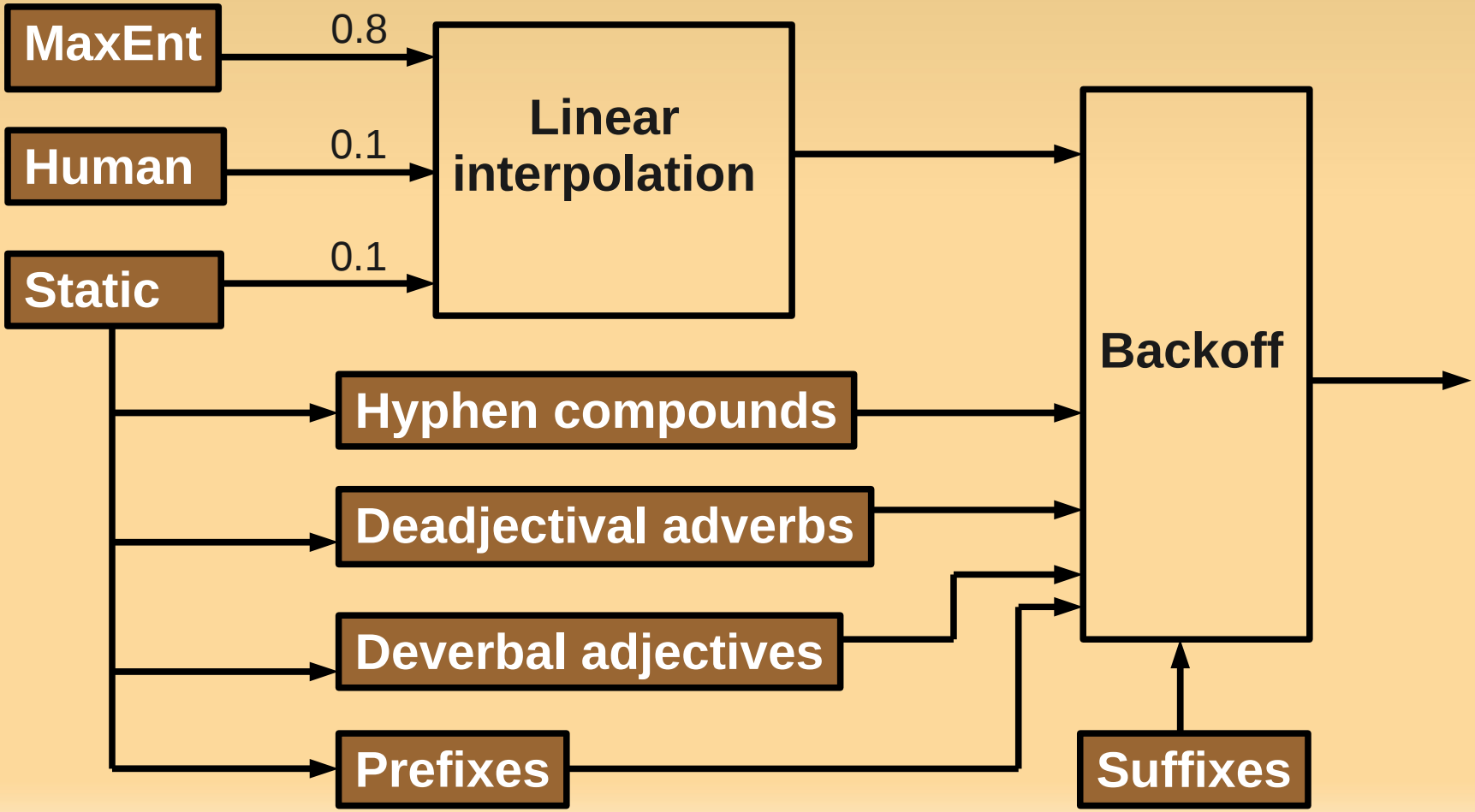
- Translate prefixes separately
  - using lexicon of prefixes
    - Multi-core → vícejádrový
    - Neoclassicism → neoklasicismus
- If we recognize suffix in an unknown word, we translate the suffix only
  - Using lexicon of suffixes
    - Geocentrism → geocentrismus

Derivative dictionaries

- Rules of translation some of hyphen-compounds
  - First part is a number the second is a noun, which is translated as an adjective
    - Two-litre → dvoulitrový
    - 45-year-old → pětačtyřicetiletý
    - Three-fifths → třípětinový
- What probability to return?
  - 1, because it is the only variant we have for the unknown source lemma

Combination of dictionaries

- The described dictionaries are combined in the following way:



Outline

- TectoMT system introduction
- Translation dictionaries
- **Experiments and results**

Experiments and Results

- Resources:
  - For TM: Czech-English parallel corpus CzEng 0.9, approx. 60 megawords on both sides, analyzed up to tectogrammatical layer and aligned
  - For LM: Czech National Corpus, 800 megawords
  - Evaluation data from WMT 2010 test set (2489 sentences)

Dictionary used	BLEU	NIST
Static only (MLE)	11.67	5.023
Static + MaxEnt	12.48	5.234
Static + MaxEnt + Derivative	12.58	5.250

Conclusions

- TectoMT – the analysis-transfer-synthesis MT system with transfer over the deep-syntax was described
- We have focused on the system of translation dictionaries:
  - Static dictionary (MLE)
  - Context dictionary (Maximum Entropy)
  - Derivational dictionary (rule-based)
- We have shown that all the dictionaries improved the quality of machine translation
  - almost 1 BLEU point improvement

Thank you for your attention!