Rudolf Rosa, Ondřej Dušek, Michal Novák, Martin Popel
{rosa,odusek,mnovak,popel}@ufal.mff.cuni.cz

Translation Model Interpolation for Domain Adaptation in TectoMT

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

DMTW, Praha, 3 September 2015

Why domain adaptation?

- not to do MT as "Europarl science"
    - reasonably good for e.g. economic news
    - not particularly good for IT helpdesk

Why domain adaptation?

- not to do MT as "Europarl science"

  - reasonably good for e.g. economic news

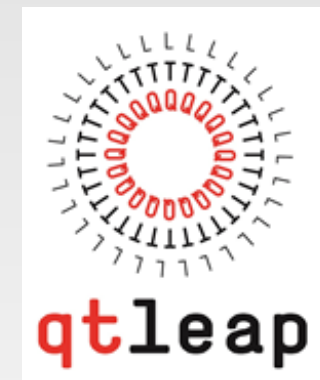 - not particularly good for IT helpdesk

Jak pošlu přílohu mailem?

How do I send an attachment by e-mail?

Click on the paperclip icon and select the file.

Klikněte na ikonu sponky a vyberte soubor.

Why model interpolation?

- available parallel data
    - out-of-domain (Europarl): 2 000 000 sentences
    - in-domain (IT helpdesk):   1 000 sentences

Why model interpolation?

- available parallel data

  - out-of-domain (Europarl): 2 000 000 sentences

  - in-domain (IT helpdesk):   1 000 sentences

- train on *out*+*in* concatenation

  - effect of *in* negligible (overpowered by *out*)

Why model interpolation?

- available parallel data

  - out-of-domain (Europarl): 2 000 000 sentences

  - in-domain (IT helpdesk):   1 000 sentences

- train on *out*+*in* concatenation

  - effect of *in* negligible (overpowered by *out*)

- train on *in* only

  - too weak, low coverage, worse than *out* alone

Why model interpolation?

- available parallel data

    - out-of-domain (Europarl): 2 000 000 sentences

    - in-domain (IT helpdesk):   1 000 sentences

- train on *out*+*in* concatenation

    - effect of *in* negligible (overpowered by *out*)

- train on *in* only

    - too weak, low coverage, worse than *out* alone

- train 2 translation models, one on *in*, one on *out*

    - final TM = (weighted) interpolation of *in*TM and *out*TM

Why TectoMT (deep transfer MT)?

✔ better meaning preservation

✔ better generalization from small data

✗ limited support for TM interpolation

✗ no support for tuning of TM weights

Why TectoMT (deep transfer MT)?

✔ better meaning preservation

✔ better generalization from small data

✗ limited support for TM interpolation – **solved** ✅

✗ no support for tuning of TM weights – **TODO** ☐

Outline

- Motivation

- TectoMT (deep transfer)

- Translation model interpolation in TectoMT

- Evaluation

- Conclusion

TectoMT deep transfer
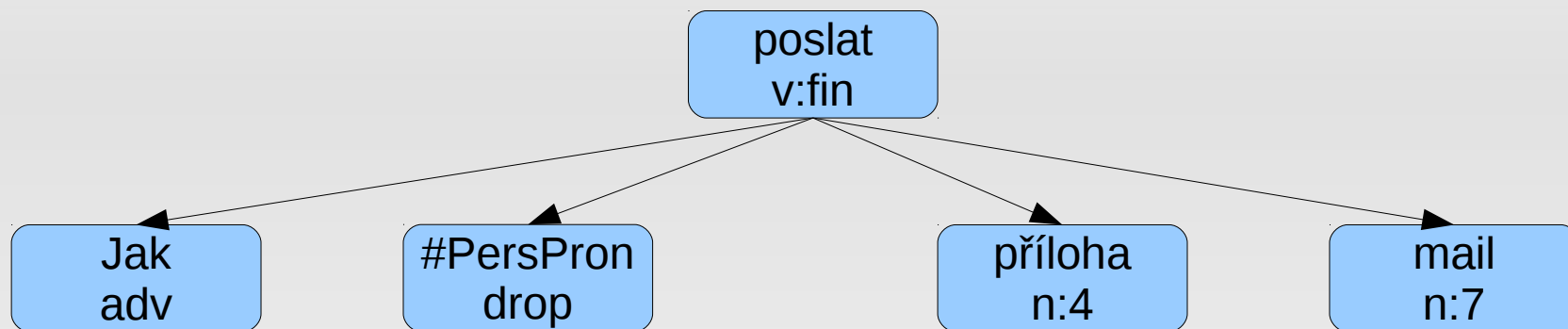
Jak pošlu přílohu mailem?

TectoMT deep transfer

TectoMT deep transfer

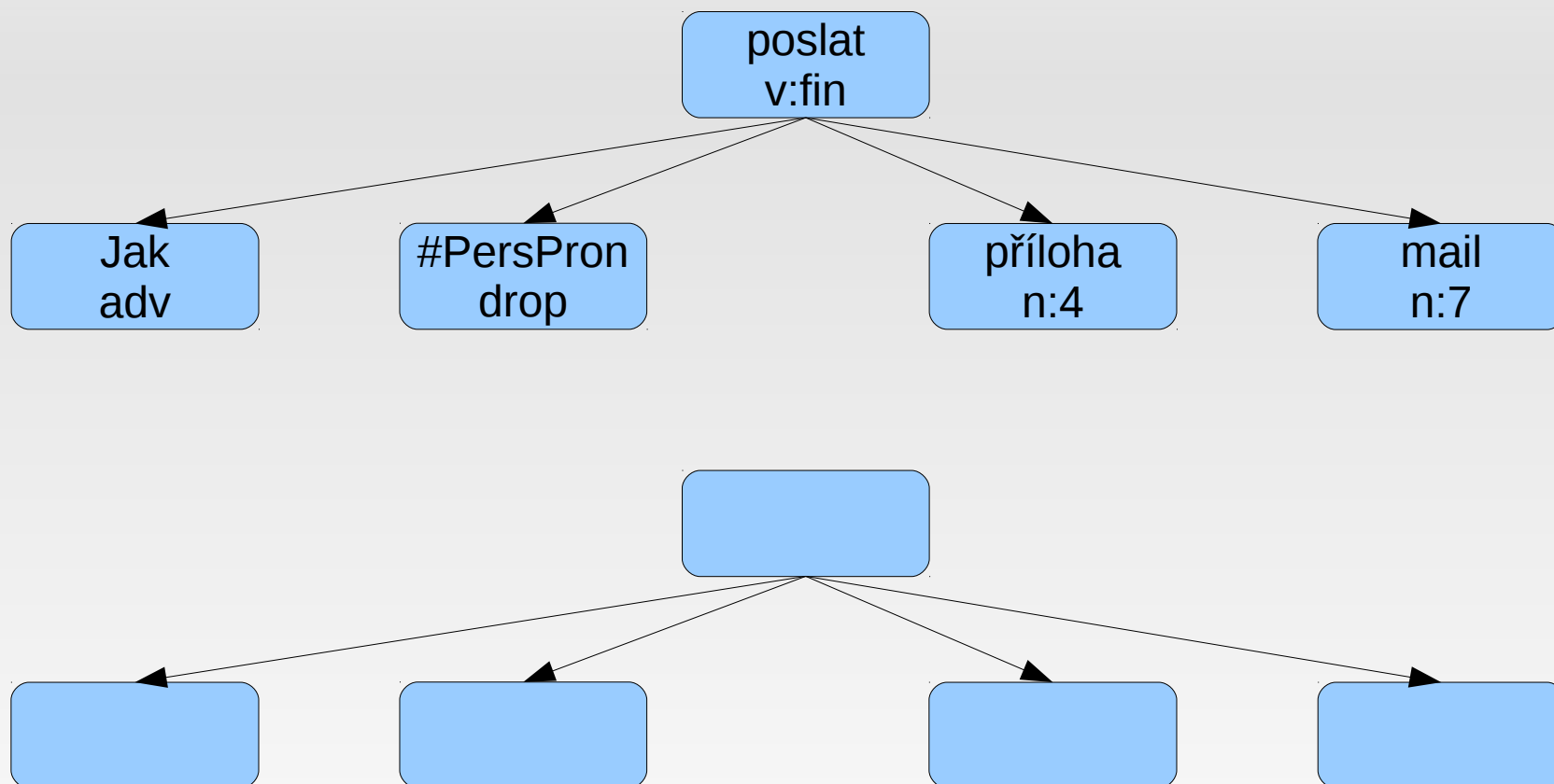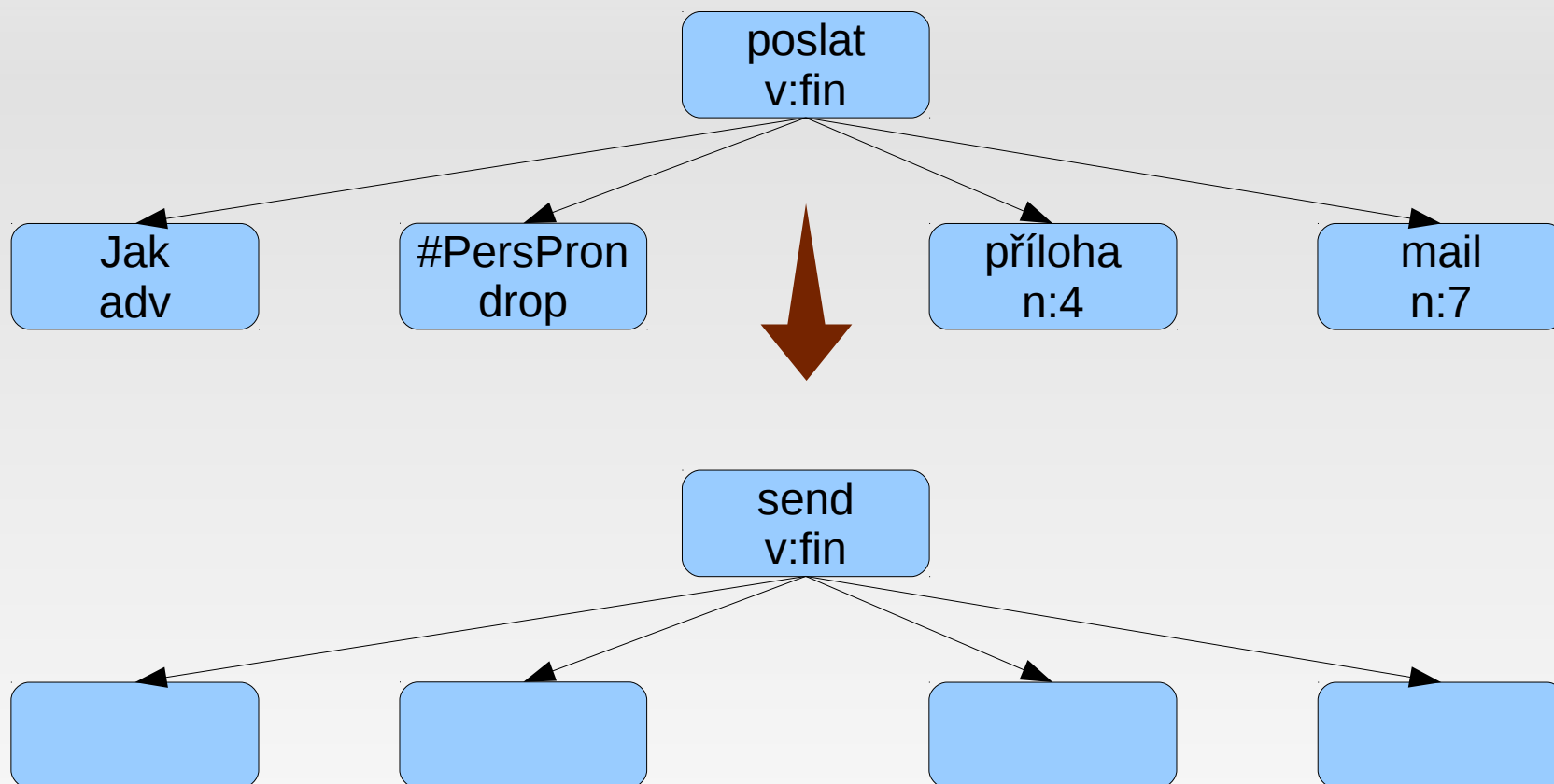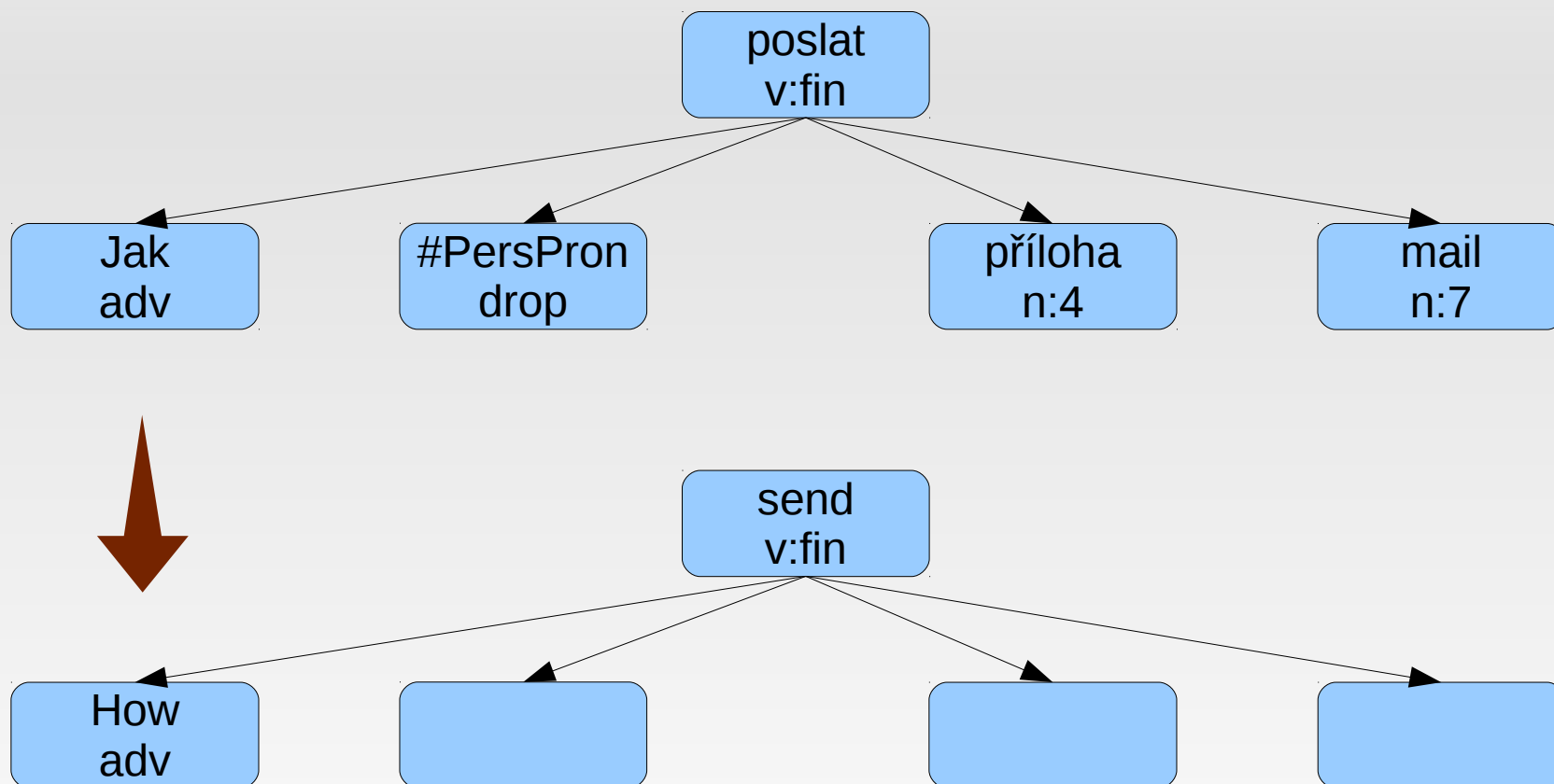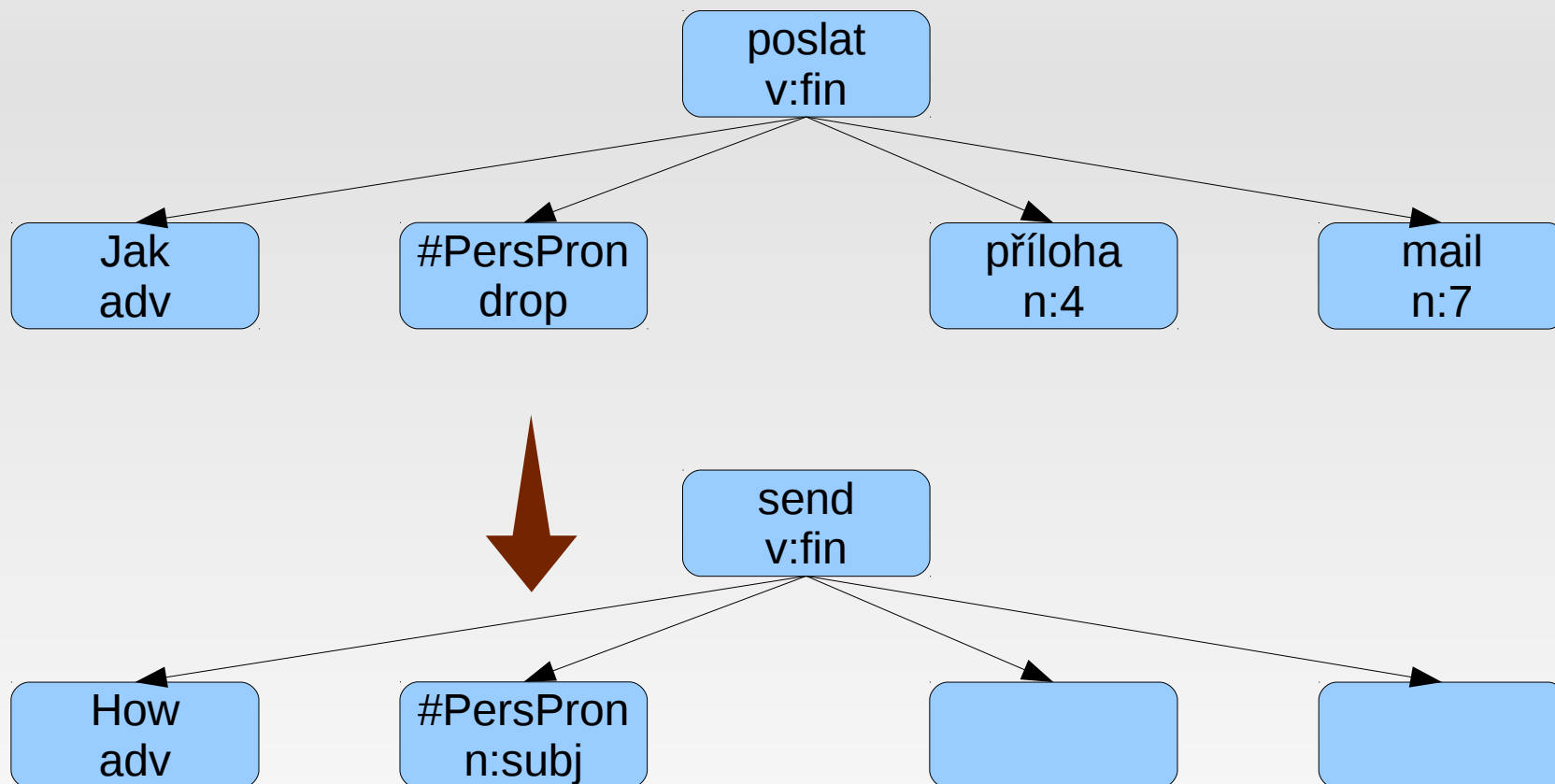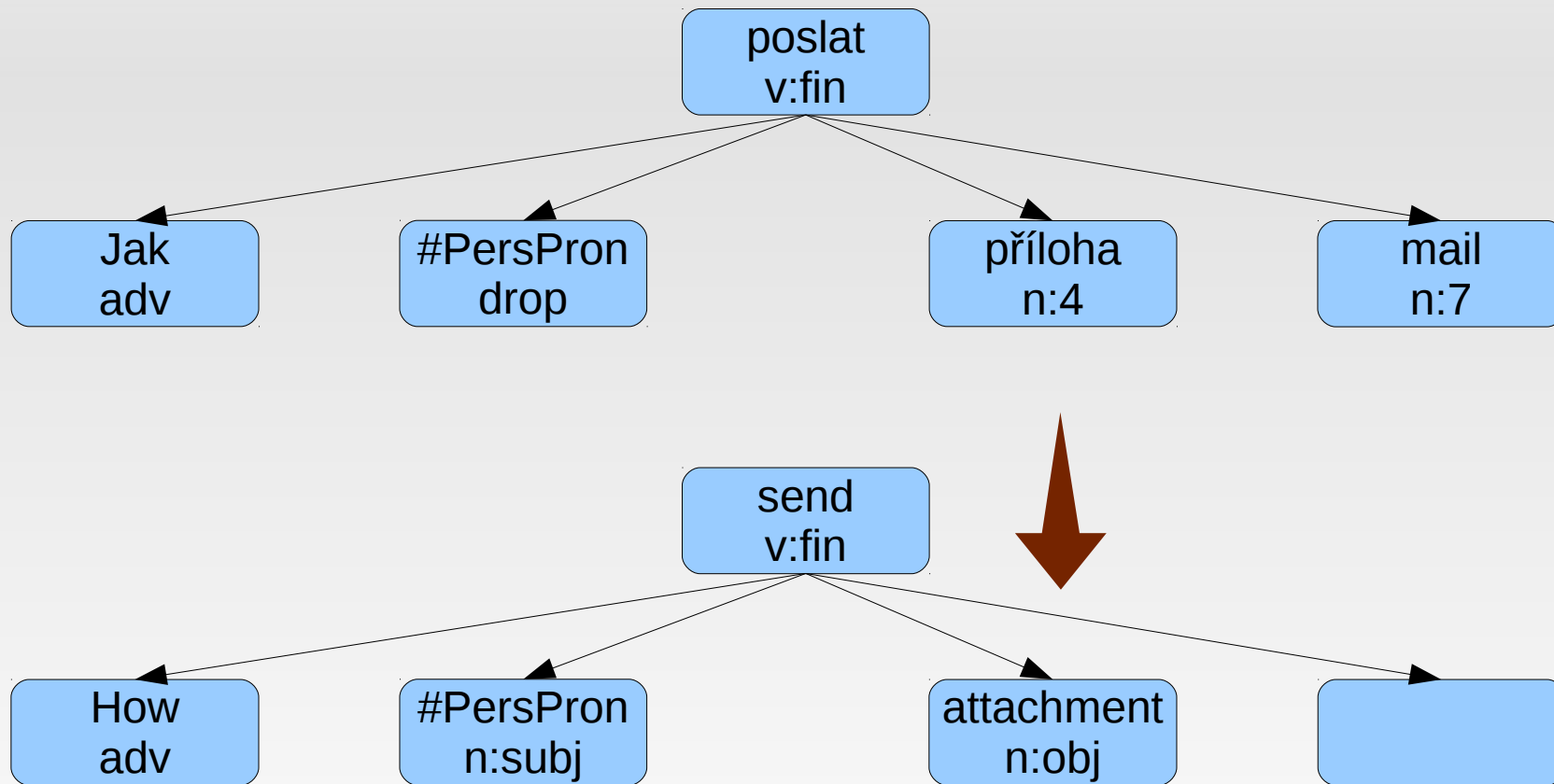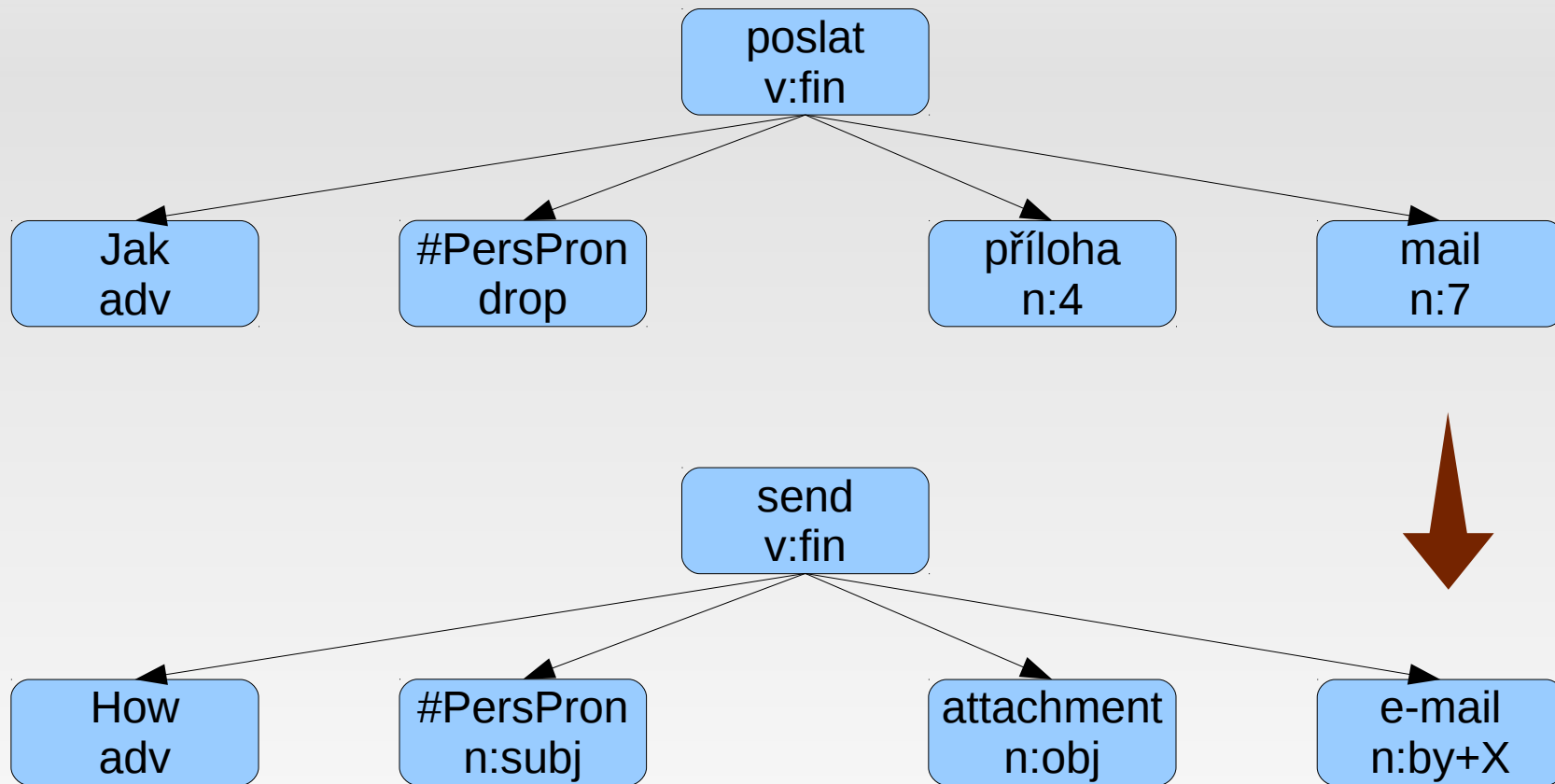# TectoMT deep transfer

Jak   pošlu   přílohu   mailem?

TectoMT deep transfer

Jak   pošlu   přílohu   mailem?
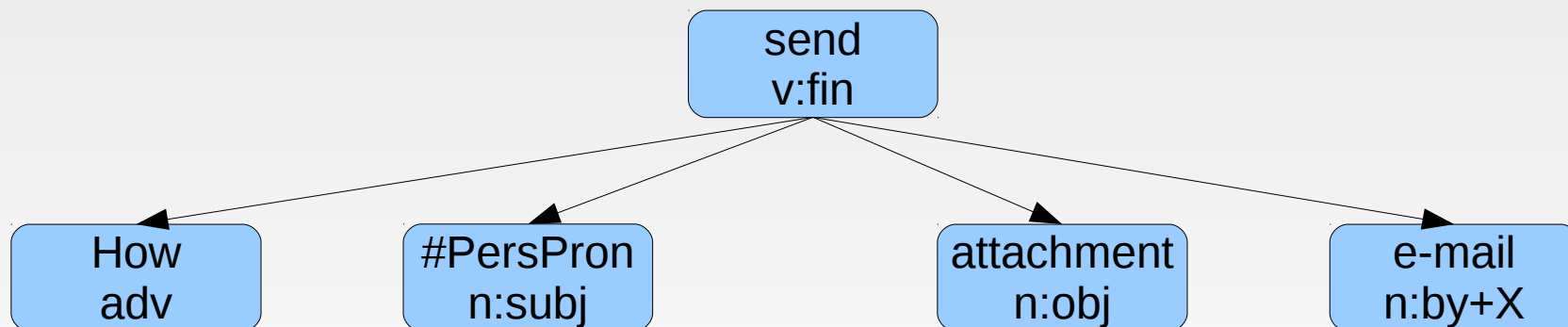
```
                         ┌──────────┐
                         │  poslat  │
                         │  v:fin   │
                         └──────────┘
              ┌──────┬───────┴───────┬──────────┐
       ┌──────────┐ ┌──────────┐ ┌──────────┐ ┌──────────┐
       │   Jak    │ │#PersPron │ │ příloha  │ │   mail   │
       │   adv    │ │   drop   │ │   n:4    │ │   n:7    │
       └──────────┘ └──────────┘ └──────────┘ └──────────┘
```

⬇

```
                         ┌──────────┐
                         │   send   │
                         │  v:fin   │
                         └──────────┘
              ┌──────┬───────┴───────┬──────────┐
       ┌──────────┐ ┌──────────┐ ┌──────────┐ ┌──────────┐
       │   How    │ │          │ │          │ │          │
       │   adv    │ │          │ │          │ │          │
       └──────────┘ └──────────┘ └──────────┘ └──────────┘
```

TectoMT deep transfer

Jak   pošlu   přílohu   mailem?

TectoMT deep transfer

TectoMT deep transfer
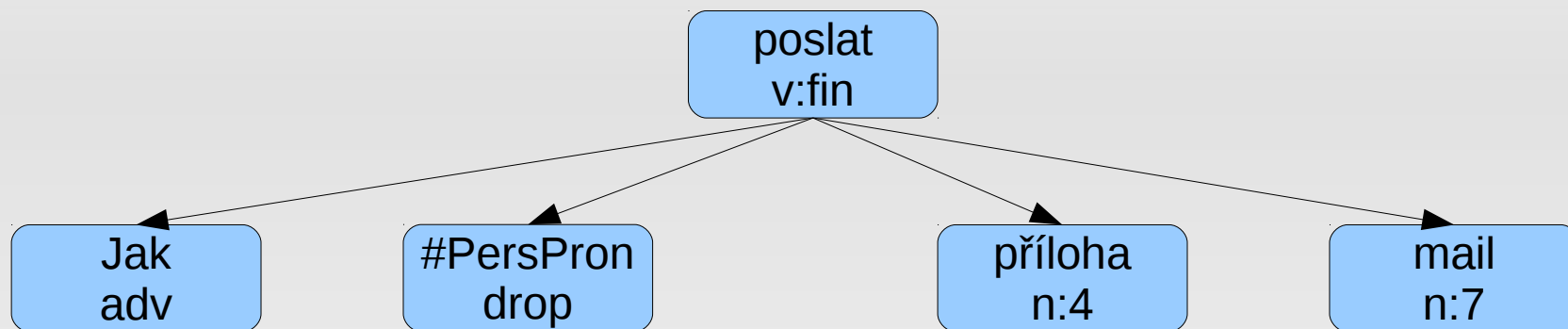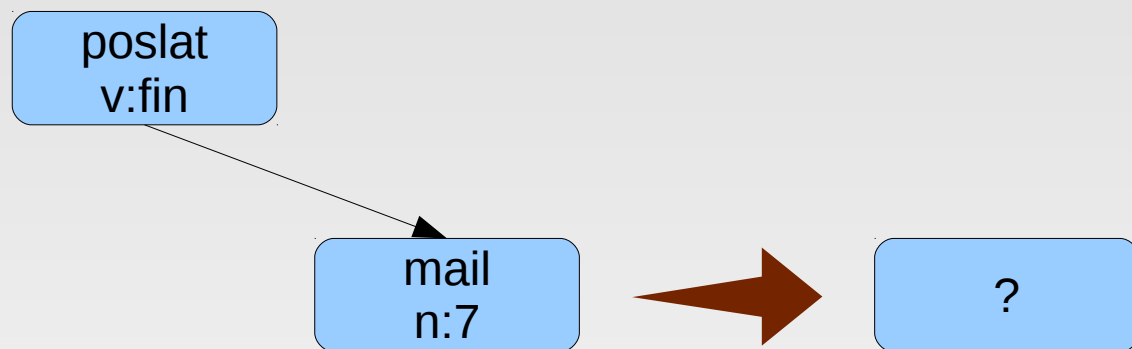
T-lemma and formeme transfer

T-lemma and formeme transfer

poslat
v:fin

mail
n:7

e-mail (n)	0.32
online (adv)	0.25
address (n)	0.11
online (adj)	0.05
...	

n:by+X	0.23
n:with+X	0.22
adv	0.13
n:via+X	0.04
...	

T-lemma and formeme transfer

poslat
v:fin

mail
n:7

e-mail (n)   0.32
online (adv)   0.25
address (n)   0.11
online (adj)   0.05
...

n:by+X   0.23
n:with+X   0.22
adv   0.13
n:via+X   0.04
...

e-mail
n:by+X

Translation model interpolation

$$P(\text{e-mail}|\text{mail}) = \frac{2}{3} MaxEnt_{\text{mail}}(\text{e-mail}, context) + \frac{1}{3} Static(\text{e-mail}, \text{mail})$$

Translation model interpolation

- past:

$$P(\text{e-mail}|\text{mail}) = \frac{2}{3} MaxEnt_{\text{mail}}(\text{e-mail}, context) + \frac{1}{3} Static(\text{e-mail}, \text{mail})$$

- present:

$$P(\text{e-mail}|\text{mail}) = \frac{2}{6} OutMaxEnt_{\text{mail}}(\text{e-mail}, context) + \frac{1}{6} OutStatic(\text{e-mail}, \text{mail})$$

$$+ \frac{2}{6} InMaxEnt_{\text{mail}}(\text{e-mail}, context) + \frac{1}{6} InStatic(\text{e-mail}, \text{mail})$$

Translation model interpolation

- future:

$$P(\text{e-mail}|\text{mail}) = w_1 \cdot OutMaxEnt_{\text{mail}}(\text{e-mail}, context) + w_2 \cdot OutStatic(\text{e-mail}, \text{mail})$$

$$+ w_3 \cdot InMaxEnt_{\text{mail}}(\text{e-mail}, context) + w_4 \cdot InStatic(\text{e-mail}, \text{mail})$$

- interpolation weights tuned on development set

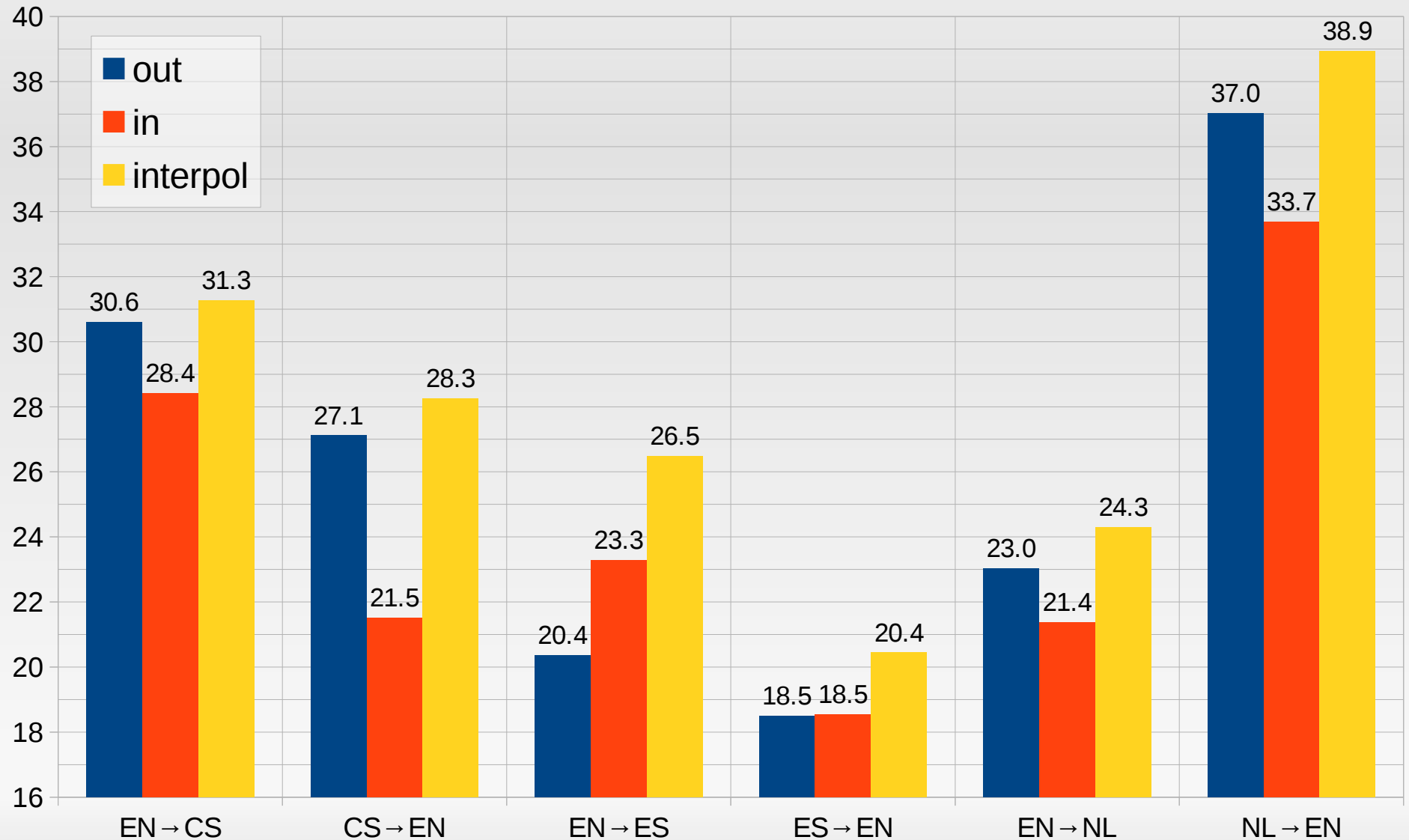# Translation model interpolation

- future:

$$P(\text{e-mail}|\text{mail}) = w_1 \cdot OutMaxEnt_{\text{mail}}(\text{e-mail}, context) + w_2 \cdot OutStatic(\text{e-mail}, \text{mail})$$

$$+ w_3 \cdot InMaxEnt_{\text{mail}}(\text{e-mail}, context) + w_4 \cdot InStatic(\text{e-mail}, \text{mail})$$

- interpolation weights tuned on development set

- not available at the moment:

  - tuning in TectoMT
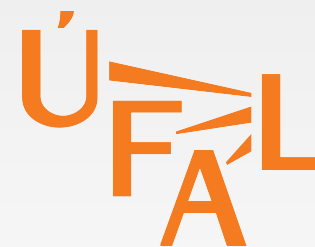
  - development set

Evaluation

Conclusion

- TectoMT deep transfer system
  - IT helpdesk domain (1000 parallel sentences)
  - domain adaptation by TM interpolation
  - no tuning of interpolation weights
  - consistently positive results, up to +3 BLEU
- future work
  - tuning of interpolation weights
  - other domain adaptation methods

Thank you for your attention

Rudolf Rosa, Ondřej Dušek, Michal Novák, Martin Popel
{rosa,odusek,mnovak,popel}@ufal.mff.cuni.cz

**Translation Model Interpolation
for Domain Adaptation
in TectoMT**

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

http://ufal.mff.cuni.cz/rudolf-rosa/