Using Parallel Texts and Lexicons for Verbal Word Sense Disambiguation

Ondřej Dušek, Eva Fučíková, Jan Hajič,
Martin Popel, Jana Šindlerová, and Zdeňka Urešová

Institute of Formal and Applied Linguistics
Charles University in Prague

August 25, 2015

Introduction

The task

- given a *verb* in a sentence, assign a sense from a dictionary

Introduction

The task

- given a *verb* in a sentence, assign a sense from a dictionary

The Dow ***fell*** *22.6% on Black Monday.*

Introduction

The task

- given a *verb* in a sentence, assign a sense from a dictionary

*The Dow **fell** 22.6% on Black Monday.*

fall^1	occur on a given date
fall^2	be defeated
fall^3	(phrasal) fall in love
fall^4	be reduced, move down
	...

Introduction

The task

- given a *verb* in a sentence, assign a sense from a dictionary

*The Dow **fell** 22.6% on Black Monday.*

fall^1	occur on a given date
fall^2	be defeated
fall^3	(phrasal) fall in love
fall^4	be reduced, move down
	. . .

Introduction

The task

- given a *verb* in a sentence, assign a sense from a dictionary

*The Dow **fell** 22.6% on Black Monday.*

fall^1	occur on a given date
fall^2	be defeated
fall^3	(phrasal) fall in love
fall^4	be reduced, move down
	. . .

- based on:
    - monolingual context (Dušek et al., 2014)

Introduction

The task

- given a *verb* in a sentence, assign a sense from a dictionary

The Dow ***fell*** *22.6% on Black Monday.*

fall^1	occur on a given date
fall^2	be defeated
fall^3	(phrasal) fall in love
fall^4	be reduced, move down
	...

- based on:
  - monolingual context (Dušek et al., 2014)
  - + parallel context (new)

Introduction

The task

- given a *verb* in a sentence, assign a sense from a dictionary

*The Dow **fell** 22.6% on Black Monday.*

fall^1	occur on a given date
fall^2	be defeated
fall^3	(phrasal) fall in love
fall^4	be reduced, move down
	…

- based on:
  - monolingual context (Dušek et al., 2014)
  - + parallel context (new)
- part of our deep analysis pipeline

Why try parallel texts?

- Translation is, in a way, interpretation

Why try parallel texts?

- Translation is, in a way, interpretation
    - different senses tend to have different translations

Why try parallel texts?

- Translation is, in a way, interpretation
  - different senses tend to have different translations

→ should be helpful for WSD in some cases

Why try parallel texts?

- Translation is, in a way, interpretation
  - different senses tend to have different translations

→ should be helpful for WSD in some cases

EN: *The Dow **fell** 22.6% on Black Monday.*

CS: *Akcie Dow Jones na Černé pondělí **ztratily** 22.6 %.*

Why try parallel texts?

- Translation is, in a way, interpretation
  - different senses tend to have different translations

→ should be helpful for WSD in some cases

EN: *The Dow* ***fell*** *22.6% on Black Monday.*

CS: *Akcie Dow Jones na Černé pondělí* ***ztratily*** *22.6 %.*

fall^1	occur on a given date	*připadnout*
fall^2	be defeated	*padnout, podlehnout*
fall^3	(phrasal) fall in love	*zamilovat se*
fall^4	be reduced, move down	***ztratit***, *spadnout*
	…	

Why try parallel texts?

- Translation is, in a way, interpretation
  - different senses tend to have different translations

  → should be helpful for WSD in some cases

EN: *The Dow **fell** 22.6% on Black Monday.*

CS: *Akcie Dow Jones na Černé pondělí **ztratily** 22.6 %.*

fall[1]	occur on a given date	*připadnout*
fall[2]	be defeated	*padnout, podlehnout*
fall[3]	(phrasal) fall in love	*zamilovat se*
fall[4]	be reduced, move down	***ztratit**, spadnout*
	…	

Why try parallel texts?

- Translation is, in a way, interpretation
  - different senses tend to have different translations

  → should be helpful for WSD in some cases

EN: *The Dow **fell** 22.6% on Black Monday.*

CS: *Akcie Dow Jones na Černé pondělí **ztratily** 22.6 %.*

fall^{1}	occur on a given date	*připadnout*
fall^{2}	be defeated	*padnout, podlehnout*
fall^{3}	(phrasal) fall in love	*zamilovat se*
fall^{4}	be reduced, move down	***ztratit**, spadnout*
	. . .	

- WSD for texts where translation exists
- Pre-annotation
- Any text, using MT?

Theoretical base

- Functional Generative Description
  - a.k.a. tectogrammatics (Sgall et al., 1986)
  - dependency-based

Theoretical base

- Functional Generative Description
  - a.k.a. tectogrammatics (Sgall et al., 1986)
  - dependency-based
- Two structural layers:
  - *a-tree*: surface dependency trees

Theoretical base

- Functional Generative Description
    - a.k.a. tectogrammatics (Sgall et al., 1986)
    - dependency-based
- Two structural layers:
    - *a-tree*: surface dependency trees
    - *t-tree*: deep syntactic dependency trees

Theoretical base

- Functional Generative Description
    - a.k.a. tectogrammatics (Sgall et al., 1986)
    - dependency-based
- Two structural layers:
    - *a-tree*: surface dependency trees
    - *t-tree*: deep syntactic dependency trees
- Prague Dependency Treebanks family
    - manually annotated with *t-trees*

Theoretical base

- Functional Generative Description
  - a.k.a. tectogrammatics (Sgall et al., 1986)
  - dependency-based
- Two structural layers:
  - *a-tree*: surface dependency trees
  - *t-tree*: deep syntactic dependency trees
- Prague Dependency Treebanks family
  - manually annotated with *t-trees*
  - PDT 2.5 – 800kW Czech news texts

Theoretical base

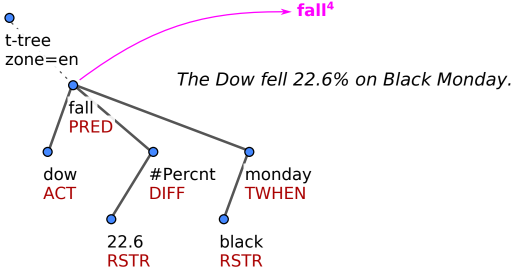
- Functional Generative Description
  - a.k.a. tectogrammatics (Sgall et al., 1986)
  - dependency-based
- Two structural layers:
  - *a-tree*: surface dependency trees
  - *t-tree*: deep syntactic dependency trees
- Prague Dependency Treebanks family
  - manually annotated with *t-trees*
  - PDT 2.5 – 800kW Czech news texts
  - PCEDT 2.0 – 1.1MW parallel English-Czech, based on PTB-WSJ
    - PTB-WSJ translated into Czech

Word sense annotation in FGD

- present in *t-trees*

Word sense annotation in FGD

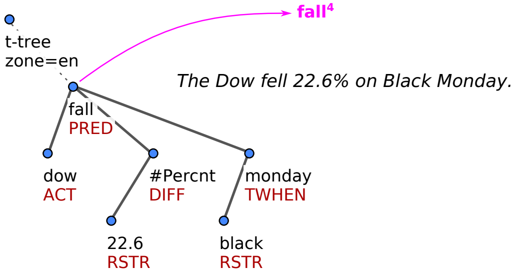
- present in *t-trees*



- t-trees: nodes for content words only
  - t-lemma – deep lemma
  - functor – function label (similar to PropBank labels)
  - grammatemes – grammatical functions (not shown)

Word sense annotation in FGD

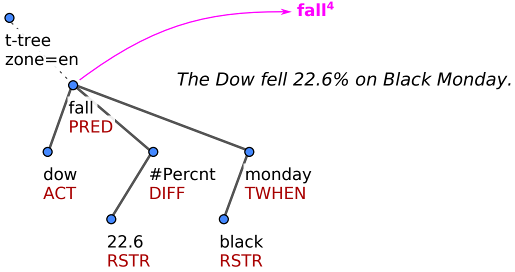
- present in *t-trees*



- t-trees: nodes for content words only
  - t-lemma – deep lemma
  - functor – function label (similar to PropBank labels)
  - grammatemes – grammatical functions (not shown)
  - verbs + some nouns: valency frame ID = sense

Word sense annotation in FGD

- present in *t-trees*



- t-trees: nodes for content words only
  - t-lemma – deep lemma
  - functor – function label (similar to PropBank labels)
  - grammatemes – grammatical functions (not shown)
  - verbs + some nouns: valency frame ID = sense
- ID links to a valency lexicon

Valency lexicons (dictionaries)

- Valency frame:

Valency lexicons (dictionaries)

- Valency frame:
  - sense (ID)

Valency lexicons (dictionaries)

- Valency frame:
    - sense (ID)
    - list of valency arguments (frame members)
        - functor (function label)
        - obligatory (yes/no)
        - subcategorization (surface)

Valency lexicons (dictionaries)

- Valency frame:
  - sense (ID)
  - list of valency arguments (frame members)
    - functor (function label)
    - obligatory (yes/no)
    - subcategorization (surface)
  - notes, examples

Valency lexicons (dictionaries)

- Valency frame:
    - sense (ID)
    - list of valency arguments (frame members)
        - functor (function label)
        - obligatory (yes/no)
        - subcategorization (surface)
    - notes, examples

fall^{4} **ACT**() **?DIFF**() **?ORIG**() **?PAT**()

(move downward: start and endpoints)

Sales fell to $251.2 million from $278.7 million.

ztratit^{3} **ACT**(1) **PAT**(4;na+6)

(pozbýt, 'lose')

ztratit čas čekáním.CAUS; z. na pružnosti

Valency lexicons (dictionaries)

- Valency frame:
  - sense (ID)
  - list of valency arguments (frame members)
    - functor (function label)
    - obligatory (yes/no)
    - subcategorization (surface)
  - notes, examples
- English – EngVallex
  - converted from PropBank
  - over 7k frames, 4k verbs

fall[4] **ACT**() **?DIFF**() **?ORIG**() **?PAT**()

(move downward: start and endpoints)

Sales fell to $251.2 million from $278.7 million.

ztratit[3] **ACT**(1) **PAT**(4;na+6)

(pozbýt, ‘lose’)

ztratit čas čekáním.CAUS; z. na pružnosti

Valency lexicons (dictionaries)

- Valency frame:
  - sense (ID)
  - list of valency arguments (frame members)
    - functor (function label)
    - obligatory (yes/no)
    - subcategorization (surface)
  - notes, examples
- English – EngVallex
  - converted from PropBank
  - over 7k frames, 4k verbs
- Czech – PDT-Vallex
  - built with PDT (and PCEDT), bottom-up
  - almost 12k frames, over 7k verbs

fall^4 **ACT**() **?DIFF**() **?ORIG**() **?PAT**()

(move downward: start and endpoints)

Sales fell to $251.2 million from $278.7 million.

ztratit^3 **ACT**(1) **PAT**(4;na+6)

(pozbýt, 'lose')

ztratit čas čekáním.CAUS; z. na pružnosti

Valency lexicon mapping – CzEngVallex

- English-Czech mapping of verb senses
  - + their arguments

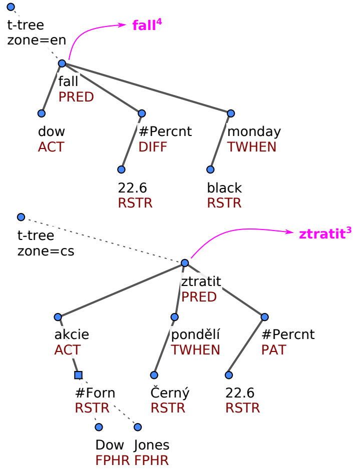
Valency lexicon mapping – CzEngVallex

- English-Czech mapping of verb senses
  - + their arguments
- Manual, based on PCEDT 2.0

Valency lexicon mapping – CzEngVallex

- English-Czech mapping of verb senses
  - + their arguments
- Manual, based on PCEDT 2.0

The Dow fell 22.6% on Black Monday.

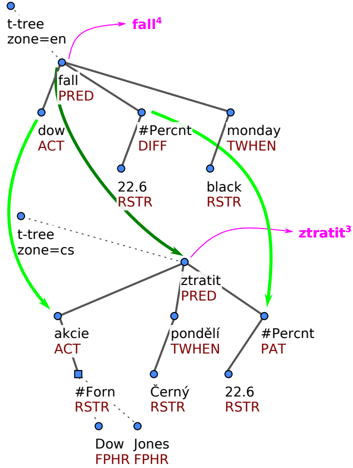


Akcie Dow Jones na Černé pondělí ztratily 22.6 %.

Valency lexicon mapping – CzEngVallex

- English-Czech mapping of verb senses
  - + their arguments
- Manual, based on PCEDT 2.0

The Dow fell 22.6% on Black Monday.



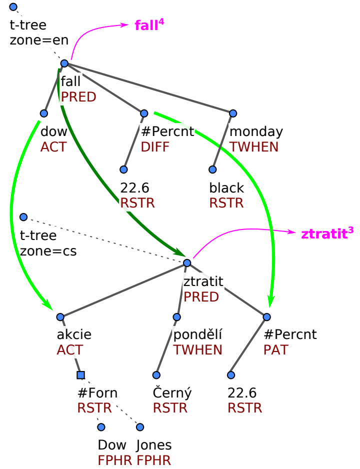
Akcie Dow Jones na Černé pondělí ztratily 22.6 %.

Valency lexicon mapping – CzEngVallex

- English-Czech mapping of verb senses
  - + their arguments
- Manual, based on PCEDT 2.0
- over 19k sense pairs

English	Czech	
3.2k	4.2k	verbs
4.9k	6.7k	senses

The Dow fell 22.6% on Black Monday.



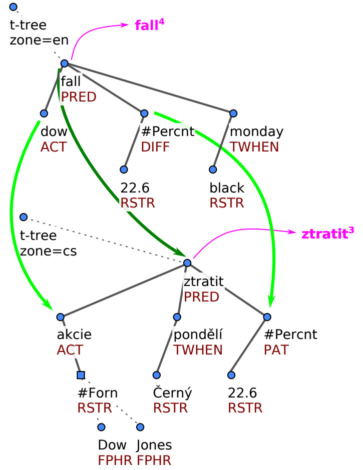
Akcie Dow Jones na Černé pondělí ztratily 22.6 %.

Valency lexicon mapping – CzEngVallex

- English-Czech mapping of verb senses
  - + their arguments
- Manual, based on PCEDT 2.0
- over 19k sense pairs

English	Czech	
3.2k	4.2k	verbs
4.9k	6.7k	senses
66%	72%	verb token PCEDT 2.0 coverage

The Dow fell 22.6% on Black Monday.



Akcie Dow Jones na Černé pondělí ztratily 22.6 %.

Experiments

- Using fully automatic analysis
  - no manual annotation needed, just monolingual/parallel texts

Experiments

- Using fully automatic analysis
  - no manual annotation needed, just monolingual/parallel texts
  - *t-tree* analysis pipeline for Czech and English from Treex NLP toolkit
    1. POS tagging
    2. dependency parsing (→ *a-tree*)
    3. mainly rule-based post-processing (→ *t-tree*)

Experiments

- Using fully automatic analysis
  - no manual annotation needed, just monolingual/parallel texts
  - *t-tree* analysis pipeline for Czech and English from Treex NLP toolkit
    1. POS tagging
    2. dependency parsing (→ *a-tree*)
    3. mainly rule-based post-processing (→ *t-tree*)
- Gold senses projected into automatically analyzed trees for training

Experiments

- Using fully automatic analysis
  - no manual annotation needed, just monolingual/parallel texts
  - *t-tree* analysis pipeline for Czech and English from Treex NLP toolkit
    1. POS tagging
    2. dependency parsing (→ *a-tree*)
    3. mainly rule-based post-processing (→ *t-tree*)
- Gold senses projected into automatically analyzed trees for training
- Data:
  - PCEDT 2.0 (parallel)
  - PDT 2.5 (monolingual, Czech)

Classifier setup

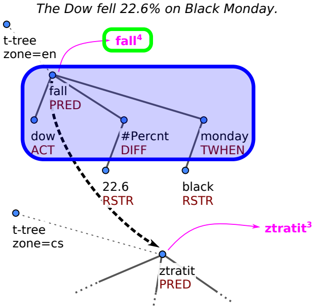
- VowpalWabbit (previous: LibLINEAR)

Classifier setup

- VowpalWabbit (previous: LibLINEAR)

Settings

A. Baseline – features from **monolingual** context (both from *a-tree* and *t-tree*):



Classifier setup

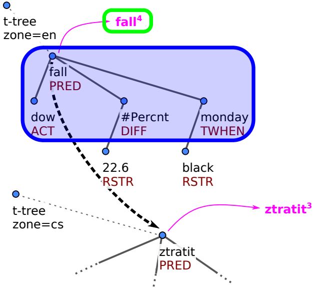
- VowpalWabbit (previous: LibLINEAR)

Settings

A. Baseline – features from **monolingual** context (both from *a-tree* and *t-tree*):

- word form, base form (lemma)
- part-of-speech + morphology
- formemes (morpho-syntactic labels)
- syntactic labels

The Dow fell 22.6% on Black Monday.

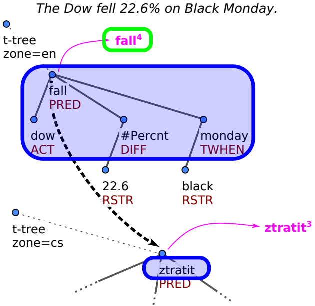


Classifier setup

- VowpalWabbit (previous: LibLINEAR)

Settings

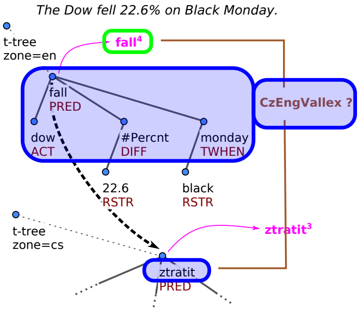
A. Baseline – features from **monolingual** context (both from *a-tree* and *t-tree*):



B. + **Aligned lemma** features – using word alignment
   - 1 feature per lemma

Classifier setup

- VowpalWabbit (previous: LibLINEAR)



Settings

A. Baseline – features from **monolingual** context (both from *a-tree* and *t-tree*):

B. + **Aligned lemma** features – using word alignment

C. + CzEngVallex **valency lexicon** mapping feature

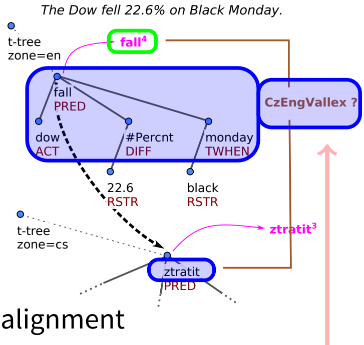
- binary: “this sense and aligned word's lemma are in the mapping”

Classifier setup

- VowpalWabbit (previous: LibLINEAR)

Settings

A. Baseline – features from **monolingual** context (both from *a-tree* and *t-tree*):



B. + **Aligned lemma** features – using word alignment

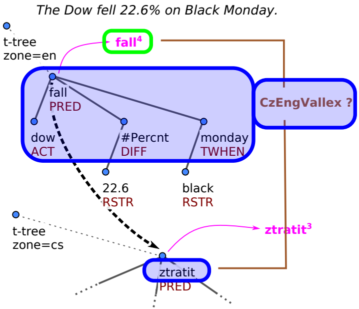
C. + CzEngVallex **valency le**

- binary: “this sense and aligned word's lemma are in the mapping
- differs with target senses

fall[1]	occur on a given date	*připadnout*
fall[2]	be defeated	*padnout, podlehnout*
fall[3]	(phrasal) fall in love	*zamilovat se*
fall[4]	be reduced, move down	***ztratit**, spadnout*
	...	

Classifier setup

- VowpalWabbit (previous: LibLINEAR)



Settings

A. Baseline – features from **monolingual** context (both from *a-tree* and *t-tree*):

B. + **Aligned lemma** features – using word alignment

C. + CzEngVallex **valency lexicon** mapping feature
   - binary:
     “this sense and aligned word's lemma are in the mapping”
   - differs with target senses
   - dense: shared for all lemmas

Evaluation

- Evaluating using F-measure (F1)

Evaluation

- Evaluating using F-measure (F1)
    - TP = identify the verb to be classified and assign correct sense

Evaluation

- Evaluating using F-measure (F1)
  - TP = identify the verb to be classified and assign correct sense
- Additional metrics
  - unlabeled F1-measure: just verb identification
  - sense classification accuracy (total/ambiguous only)

Evaluation

- Evaluating using F-measure (F1)
  - TP = identify the verb to be classified and assign correct sense
- Additional metrics
  - unlabeled F1-measure: just verb identification
  - sense classification accuracy (total/ambiguous only)

Results

English	% F1
previous	80.30
monolingual	82.39
+ aligned lemmas*	82.59
+ val. lexicon**	82.93

* = 95%, ** = 99% significance level

Evaluation

- Evaluating using F-measure (F1)
  - TP = identify the verb to be classified and assign correct sense
- Additional metrics
  - unlabeled F1-measure: just verb identification
  - sense classification accuracy (total/ambiguous only)

Results

English	% F1
previous	80.30
monolingual	82.39
+ aligned lemmas*	82.59
+ val. lexicon**	82.93

+0.5%

* = 95%, ** = 99% significance level

Evaluation

- Evaluating using F-measure (F1)
    - TP = identify the verb to be classified and assign correct sense
- Additional metrics
    - unlabeled F1-measure: just verb identification
    - sense classification accuracy (total/ambiguous only)

Results

English	% F1
previous	80.30
monolingual	82.39
+ aligned lemmas*	82.59
+ val. lexicon**	82.93

+0.5%

Czech	% F1
previous (PDT)	76.65
monolingual (PDT)	77.97
monolingual (PCEDT)	80.22
+ aligned lemmas	80.30
+ val. lexicon*	80.47

* = 95%, ** = 99% significance level

Evaluation

- Evaluating using F-measure (F1)
  - TP = identify the verb to be classified and assign correct sense
- Additional metrics
  - unlabeled F1-measure: just verb identification
  - sense classification accuracy (total/ambiguous only)

Results

English	% F1
previous	80.30
monolingual	82.39
+ aligned lemmas*	82.59
+ val. lexicon**	82.93

+0.5%

* = 95%, ** = 99% significance level

Czech	% F1
previous (PDT)	76.65
monolingual (PDT)	77.97
monolingual (PCEDT)	80.22
+ aligned lemmas	80.30
+ val. lexicon*	80.47

+0.3%

Discussion

- VowpalWabbit – large gain (+2.0/+1.3% F1)

Discussion

- VowpalWabbit – large gain (+2.0/+1.3% F1)
- Parallel features – smaller, but significant gain (+0.5/+0.3% F1)

Discussion

- VowpalWabbit – large gain (+2.0/+1.3% F1)
- Parallel features – smaller, but significant gain (+0.5/+0.3% F1)
    - Czech probably harder (more senses, sparser)
    - English less informative

Discussion

- VowpalWabbit – large gain (+2.0/+1.3% F1)
- Parallel features – smaller, but significant gain (+0.5/+0.3% F1)
    - Czech probably harder (more senses, sparser)
    - English less informative
- Valency lexicon feature better than aligned lemmas only

Discussion

- VowpalWabbit – large gain (+2.0/+1.3% F1)
- Parallel features – smaller, but significant gain (+0.5/+0.3% F1)
  - Czech probably harder (more senses, sparser)
  - English less informative
- Valency lexicon feature better than aligned lemmas only
  - dense feature, helps in rarer verbs

Discussion

- VowpalWabbit – large gain (+2.0/+1.3% F1)
- Parallel features – smaller, but significant gain (+0.5/+0.3% F1)
  - Czech probably harder (more senses, sparser)
  - English less informative
- Valency lexicon feature better than aligned lemmas only
  - dense feature, helps in rarer verbs
- Also cases where the parallel information introduces noise
  - but positive cases prevailing

Examples (English WSD improved by Czech data)

EN: *But those machines are still **considered** novelties, […]*

CS: *Ale tyto stroje […] jsou stále **považovány*** ('believe to be') *za novinky.*

- **consider**^{1} ('think about') monolingual → **consider**^{2} ('believe to be') aligned lemmas, val. lexicon

Examples (English WSD improved by Czech data)

EN: *But those machines are still* ***considered*** *novelties, [...]*

CS: *Ale tyto stroje [...] jsou stále* ***považovány*** ('believe to be') *za novinky.*

- **consider**[1] ('think about') monolingual → **consider**[2] ('believe to be') aligned lemmas, val. lexicon

EN: *This* ***feels*** *more like a one-shot deal.*

CS: *Teď to* ***vypadá*** ('looks like') *spíš na jednorázovou záležitost.*

- **feel**[4] ('have a feeling') monolingual, aligned lemmas → **feel**[5] ('look like') val. lexicon

Examples (parallel information introduces noise)

EN: *Laptops […] have become the fastest-growing […] segment , with sales* ***doubling*** *this year.*

CS: *Laptopy […] se staly, díky letošnímu* ***zdvojnásobení*** (‘doubling’ (noun)) *objemu prodeje, nejrychleji rostoucím segmentem […]*

- **double**[3] (intransitive) → **double**[2] (transitive)
  monolingual aligned lemmas, val. lexicon

Examples (parallel information introduces noise)

EN: *Laptops […] have become the fastest-growing […] segment , with sales* ***doubling*** *this year.*

CS: *Laptopy […] se staly, díky letošnímu* ***zdvojnásobení*** (‘doubling’ (noun)) *objemu prodeje, nejrychleji rostoucím segmentem […]*

- **double**^{3} (intransitive) → **double**^{2} (transitive)
  monolingual → aligned lemmas, val. lexicon

EN: *“We didn’t even get a chance to* ***do*** *the programs we wanted to do.”*

CS: *„Nedali nám žádnou šanci* ***uskutečnit*** (‘accomplish’) *plány, které jsme měli připravené.“*

- **do**^{6} (‘perform a function, run’) → **do**^{2} (‘perform an act’)
  monolingual, aligned lemmas → val. lexicon

Conclusion

- Parallel information (aligned lemma + dictionary mapping) helps in WSD

Conclusion

- Parallel information (aligned lemma + dictionary mapping) helps in WSD
    - small, but significant improvement

Conclusion

- Parallel information (aligned lemma + dictionary mapping) helps in WSD
    - small, but significant improvement
    - using information from valency lexicon mapping works best

Conclusion

- Parallel information (aligned lemma + dictionary mapping) helps in WSD
  - small, but significant improvement
  - using information from valency lexicon mapping works best

Future work

- Try to obtain valency lexicon mapping automatically

Conclusion

- Parallel information (aligned lemma + dictionary mapping) helps in WSD
  - small, but significant improvement
  - using information from valency lexicon mapping works best

Future work

- Try to obtain valency lexicon mapping automatically
- Incorporate MT (use machine-translated parallel texts)
  - this would make it comparable to monolingual WSD

Thank you for your attention

Contact us

Ondřej Dušek, Eva Fučíková, Jan Hajič, Martin Popel,
Jana Šindlerová, and Zdeňka Urešová

Charles University in Prague
Institute of Formal and Applied Linguistics

`odusek@ufal.mff.cuni.cz`

Examples (Czech WSD improved by English data)

CS: *[…] čemu lidé z televizního průmyslu **říkají*** ('call') *stanice „s nejvyšší spontánní znalostí".*

EN: *[…] what people in the television industry **call** a "top of mind" network.*

- **říkat**^{7} ('say') → **říkat**^{4} ('call')
  monolingual aligned lemmas, val. lexicon

Examples (Czech WSD improved by English data)

CS: *[…] čemu lidé z televizního průmyslu **říkají*** (‘call’) *stanice „s nejvyšší spontánní znalostí“.*

EN: *[…] what people in the television industry **call** a “top of mind” network.*

- **říkat**[7] (‘say’) monolingual → **říkat**[4] (‘call’) aligned lemmas, val. lexicon

CS: *Jestliže investor **neposkytne*** (‘does not provide, give, lend’) *dodatečnou hotovost […]*

EN: *If the investor doesn’t **put up** the extra cash […]*

- **poskytnout**[2] (‘give’, light verb) monolingual, aligned lemmas → **poskytnout**[1] (‘provide’) val. lexicon