Machine Translation 3: Linguistics in SMT and NMT

Ondřej Bojar
bojar@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

Outline of Lectures on MT

1. Introduction.
   - Why is MT difficult.
   - MT evaluation.
   - Approaches to MT.
   - First peek into phrase-based MT
   - Document, sentence and word alignment.
2. Statistical Machine Translation.
   - Phrase-based: Assumptions, beam search, key issues.
   - Neural MT: Sequence-to-sequence, attention, self-attentive.
3. Advanced Topics.
   - Linguistic Features in SMT and NMT.
   - Multilinguality, Multi-Task, Learned Representations.

Outline of MT Lecture 3

1. Linguistic features for tokens.
   - Factored phrase-based MT.
2. Linguistic structure to organize search.
   - Non-projectivity.
   - TectoMT: transfer-based deep-syntactic model.
3. Combination to make it actually work.
4. Incorporating linguistic features in NMT.
   - Dedicated models or just data hacks.
     - For multi-task, for multilingual MT.
   - Are the models understanding?

Morphological Richness (in Czech)

	Czech	English
Rich morphology	$\geq$ 4,000 tags possible	50 used
	$\geq$ 2,300 tags seen	
Word order	free	rigid

News Commentary Corpus	Czech	English
Sentences	55,676	
Tokens	1.1M	1.2M
Vocabulary (word forms)	91k	40k
Vocabulary (lemmas)	34k	28k

Czech tagging and lemmatization: Hajič and Hladká (1998)
English tagging (Ratnaparkhi, 1996) and lemmatization (Minnen et al., 2001).

Morphological Explosion in Czech

MT chooses output words <u>in a form</u>:

- Czech nouns and adjs.: 7 cases, 4 genders, 3 numbers, . . .
- Czech verbs: gender, number, aspect (im/perfective), . . .

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	kočky	.
	pily	dvě	zelená	pruhovaná	koček	
	. . .	dvou	zelené	pruhované	kočkám	
	viděl	dvěma	zelení	pruhovaní	kočkách	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	. . .		zelených	pruhovaných		
	uviděl		zelenému	pruhovanému		
	uviděla		zeleným	pruhovaným		
	. . .		zelenou	pruhovanou		
	viděl jsem		zelenými	pruhovanými		
	viděla jsem		. . .	. . .		

Morphological Explosion Elsewhere

Compounding in German:

- Rindfleischetikettierungsüberwachungsaufgabenübertragungs-gesetz.
  "beef labelling supervision duty assignment law"

Agglutination in Hungarian or Finnish:

istua	"to sit down" (istun = "I sit down")
istahtaa	"to sit down for a while"
istahdan	"I'll sit down for a while"
istahtaisin	"I would sit down for a while"
istahtaisinko	"should I sit down for a while?"
istahtaisinkohan	"I wonder if I should sit down for a while"

LM over Forms Insufficient

Possible translations differring in morphology:

two	green	striped	cats	
dvou	zelená	pruhovaný	kočkách	← garbage
dva	zelené	pruhované	kočky	← 3grams ok, 4gram bad
dvě	zelené	pruhované	kočky	← correct nominative/accusative
dvěma	zeleným	pruhovaným	kočkám	← correct dative

- 3-gram LM too weak to ensure agreement.
- 3-gram LM possibly already too sparse!

Explicit Morphological Target Factor

- Add morphological tag to each output token:

two	green	striped	cats	
dvou	zelená	pruhovaný	kočkách	← garbage
fem-loc	*neut-acc*	*masc-nom-sg*	*fem-loc*	
dva	zelené	pruhované	kočky	← 3-grams ok, 4-gram bad
masc-nom	*masc-nom*	*masc-nom*		
	fem-nom	*fem-nom*	*fem-nom*	
dvě	zelené	pruhované	kočky	← correct nominative/accusative
fem-nom	*fem-nom*	*fem-nom*	*fem-nom*	
fem-acc	*fem-acc*	*fem-acc*	*fem-acc*	
dvěma	zeleným	pruhovaným	kočkám	← correct dative
fem-dat	*fem-dat*	*fem-dat*	*fem-dat*	

Advantages of Explicit Morphology

- LM over morphological tags generalizes better.
  - p(dvě kočkách) $<$ p(dvě kočky) . . . surely

    But we would need to see all combinations of dva and $kočka$!

    $\Rightarrow$ Better to ask if p(*fem-nom fem-loc*) $<$ p(*fem-nom fem-nom*)
    which is trained on any feminine adj+noun.
- But still does not solve everything.
  - p(dvě zelené) $\gtrless$ p(dva zelené) . . . bad question anyway!

    Not solved by asking if p(*fem-nom fem-nom*) $\gtrless$ p(*masc-nom masc-nom*).
- Tagset size smaller than vocabulary.

  $\Rightarrow$ can afford e.g. 7-grams:

  p(*masc-nom fem-nom fem-nom*) $<$ p(*fem-nom fem-nom fem-nom*)

Any risks?

Factored Phrase-Based MT

- Both input and output words can have more factors.
- Arbitrary number and order of:

  Mapping/Translation steps ($\rightarrow$)

  Translate (phrases of) source factors to target factors.

  two green $\rightarrow$ dvě zelené

  Generation steps ($\downarrow$)

  Generate target factors from target factors.

  dvě $\rightarrow$ *fem*-*nom*; dva $\rightarrow$ *masc*-*nom*

  $\Rightarrow$ Ensures "vertical" coherence.

  Target-side language models (+LM)

  Applicable to various target-side factors.

  $\Rightarrow$ Ensures "horizontal" coherence.

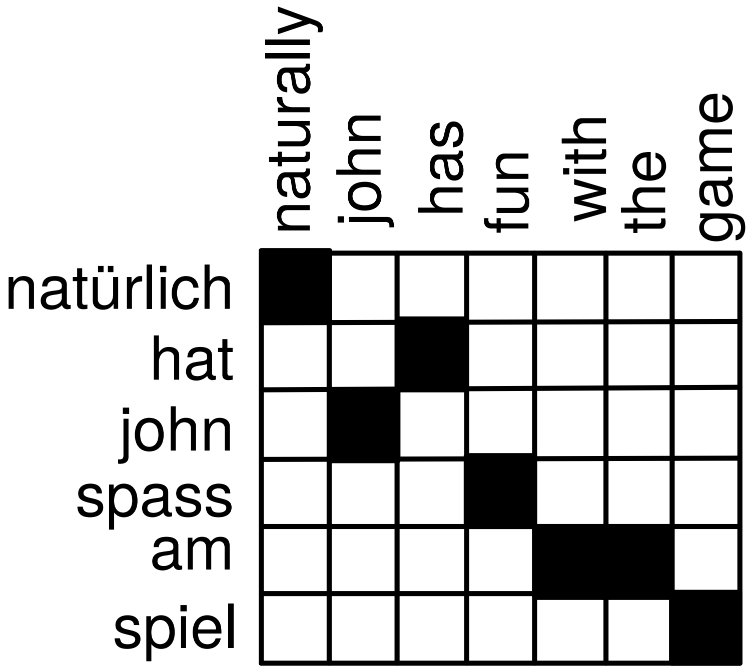
src	tgt	
$f_1 \rightarrow$	$e_1 \downarrow$	+LM
f_2	e_2	

(Koehn and Hoang, 2007)

Factored Phrase Extraction (1/3)

As in standard phrase-based MT:

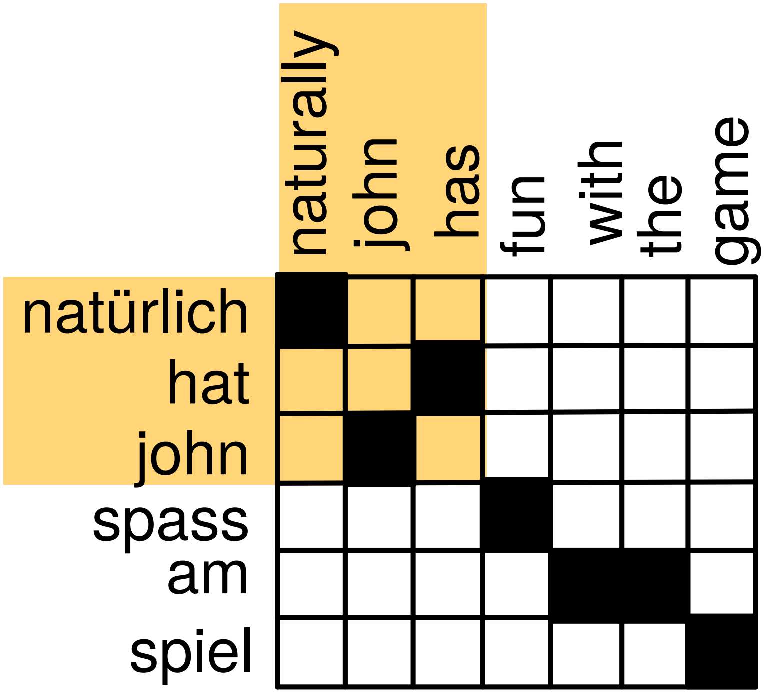
1. Run sentence and word alignment,



Factored Phrase Extraction (2/3)

As in standard phrase-based MT:

1. Run sentence and word alignment,
2. Extract all phrases consistent with word alignment.

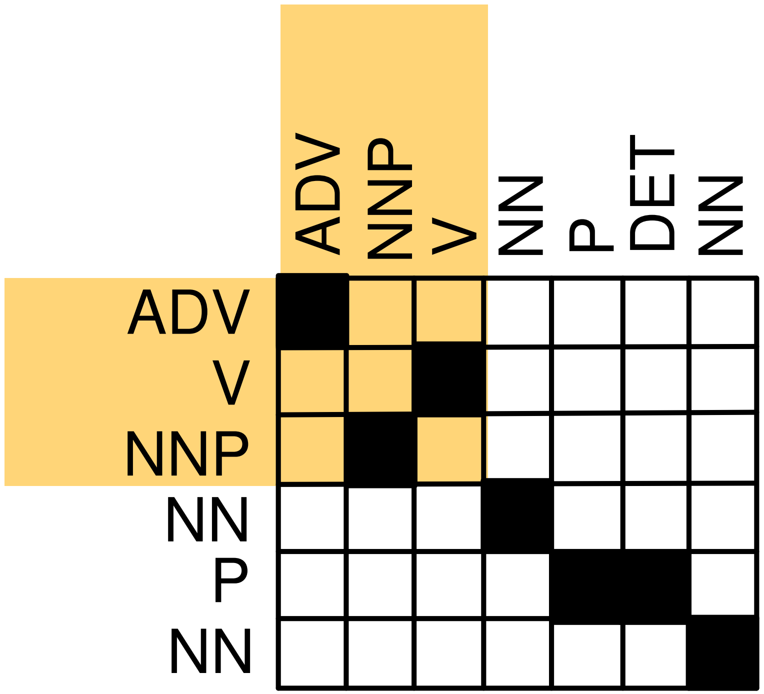


⇒ Extracted: natürlich hat john → naturally john has

Factored Phrase Extraction (3/3)

As in standard phrase-based MT:

1. Run sentence and word alignment,
2. Extract same phrases, just another factor from each word.



$\Rightarrow$ Extracted: ADV V NNP $\rightarrow$ ADV NNP V

Factored Translation Process

Input: (cars, car, NNS)

1. Translation step: lemma $\Rightarrow$ lemma

   (_, auto, _), (_, automobil, _), (_, vůz, _)

2. Generation step: lemma $\Rightarrow$ part-of-speech

   (_, auto, N-sg-nom), (_, auto, N-sg-gen), . . . ,

   (_, vůz, N-sg-nom), . . . , (_, vůz, N-sg-gen) . . .

3. Translation step: part-of-speech $\Rightarrow$ part-of-speech

   (_, auto, N-plur-nom), (_, auto, N-plur-acc), . . . ,

   (_, vůz, N-plur-nom), . . . , (_, vůz, N-sg-gen) . . .

4. Generation step: lemma, part-of-speech $\Rightarrow$ surface

   (auta, auto, N-plur-nom), (auta, auto, N-plur-acc), . . . ,

   (vozy, vůz, N-plur-nom), . . . , (vozu, vůz, N-sg-gen) . . .

Factored Phrase-Based MT

See slides by Philipp Koehn, pages 49–75:

- Decoding
- Experiments
  - incl. Alternative Decoding Paths

Translation Scenarios for En→Cs

Vanilla

English		Czech	
form	→	form	+LM
lemma		lemma	
morphology		morphology	

Translate+Check (T+C)

English		Czech	
form	→	form	+LM
lemma		lemma	
morphology		morphology	+LM

Translate+2·Check (T+C+C)

English		Czech	
form	→	form	+LM
lemma		lemma	+LM
morphology		morphology	+LM

2·Translate+Generate (T+T+G)

English		Czech	
form		form	+LM
lemma	→	lemma	+LM
morphology	→	morphology	+LM

Factored Attempts (WMT09)

Sents	System	BLEU	NIST	Sent/min
2.2M	Vanilla	**14.24**	**5.175**	12.0
2.2M	T+C	13.86	5.110	2.6
84k	T+C+C&T+T+G	10.01	4.360	4.0
84k	Vanilla MERT	10.52	4.506	–
84k	Vanilla even weights	08.01	3.911	–

- In WMT07, T+C worked best.
  + fine-tuned tags helped with small data (Bojar, 2007).
- In WMT08, T+C was worth the effort (Bojar and Hajič, 2008).
- In WMT09, our computers could handle 7-grams of forms.
  ⇒ No gain from T+C.
- T+T+G too big to fit and explodes the search space.
  ⇒ Worse than Vanilla trained on the same dataset.

T+T+G Failure Explained

- Factored models are "**synchronous**", i.e. Moses:
  1. Generates fully instantiated "translation options".
  2. Appends translation options to extend "partial hypothesis".
  3. Applies LM to see how well the option fits the previous words.
- There are too many possible combinations of lemma+tag.

⇒ Less promising ones must be pruned.

! Pruned <u>before</u> the linear context is available.

A Fix: Reverse Self-Training

Goal: Learn from monolingual data to produce new target-side word forms in correct contexts.

	Source English		Target Czech
Para 126k	a cat chased...	=	**kočka** honila...
			kočka honit... (lem.)
	I saw a cat	=	viděl jsem **kočku**
			vidět být kočka (lem.)
Mono 2M	?		četl jsem o **kočce**
			číst být o kočka (lem.)
			Use reverse translation
	I read about a cat	←	backed-off by lemmas.

⇒ New phrase learned: "about a cat" = "o **kočce**".

The Back-off to Lemmas

- The key distinction from self-training used for domain adaptation
  (Bertoldi and Federico, 2009; Ueffing et al., 2007).
- We use simply "alternative decoding paths" in Moses:

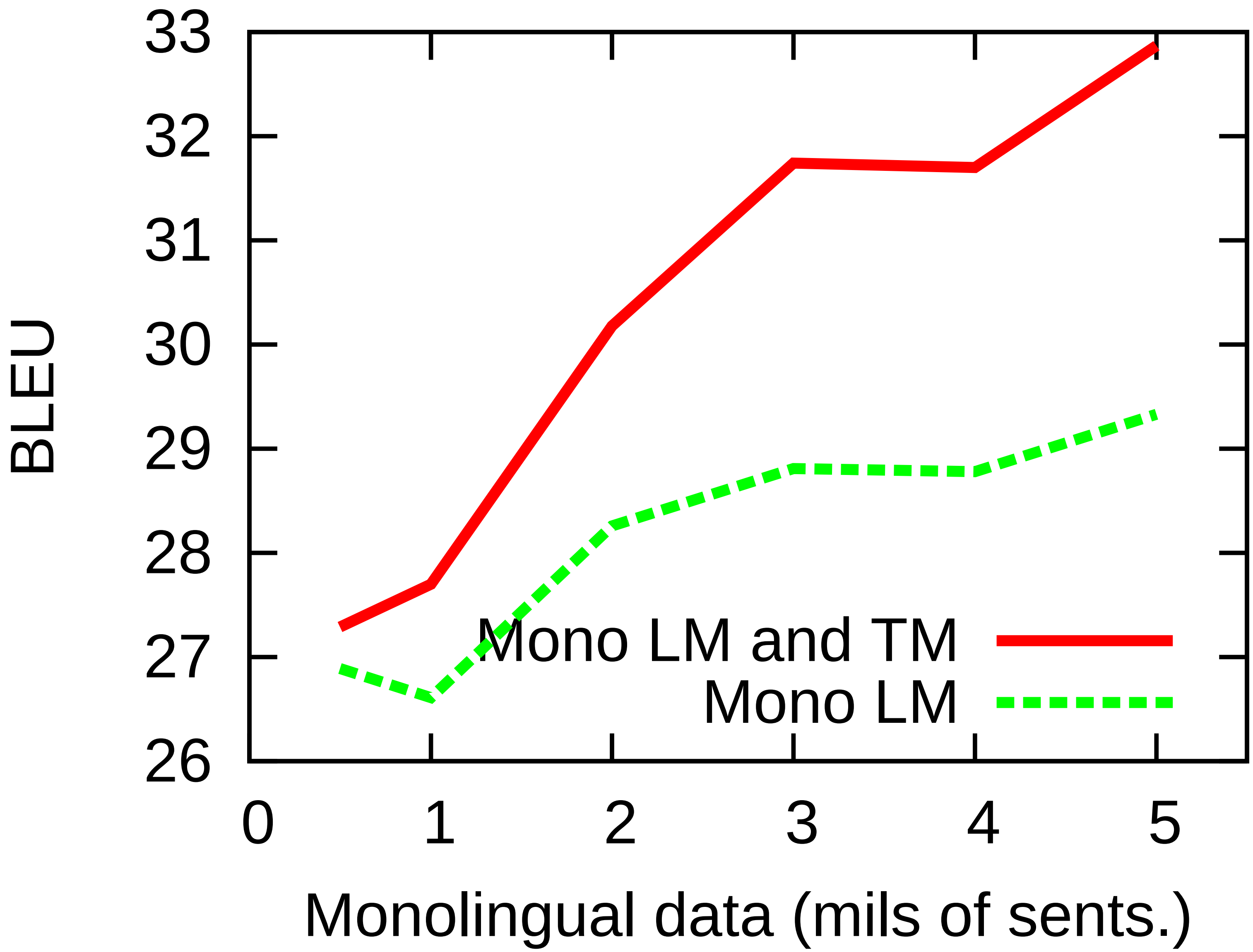
Czech	English	
form →	form	+LM

or

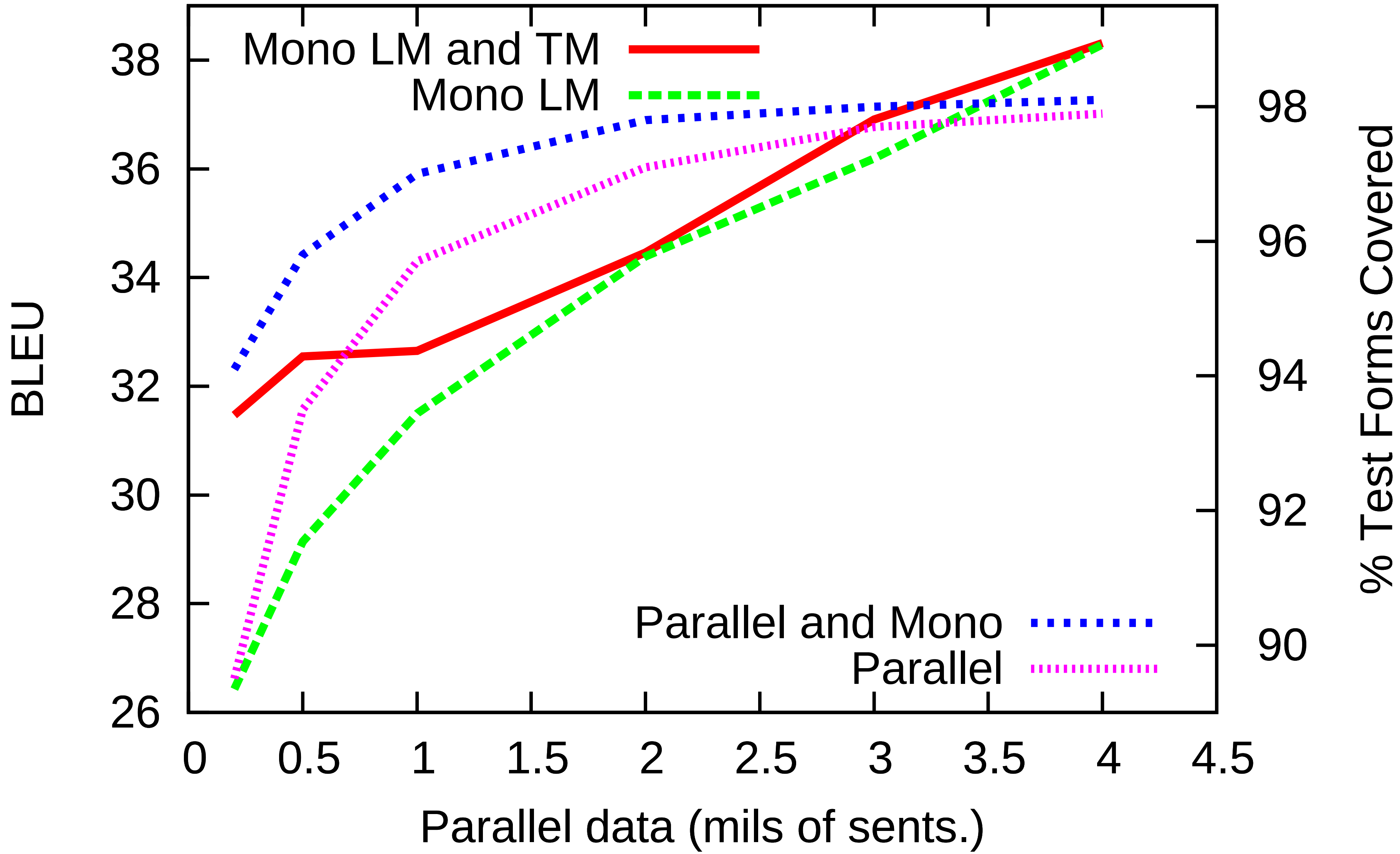
Czech	English	
lemma →	form	+LM

- Other languages (e.g. Turkish, German) need different back-off techniques:
  – Split German compounds.
  – Separate and allow to ignore Turkish morphology.

Small Para, Increasing Mono



Increasing Para, Fixed Mono



Summary So Far

- Target-side rich morphology causes data sparseness.
- Factored setups compact the sparseness.

  . . . but the search space is likely to explode at runtime.
- Explosion contained thanks to pruning.

  . . . but the pruning happens without linear context

  $\Rightarrow$ high risk of search errors.

One of possible promising techniques for handling sparseness and avoiding the explosion:

- Reverse self-training (Bojar and Tamchyna, 2011).

. . . so that was morphology, how about syntax?

Constituency vs. Dependency Trees

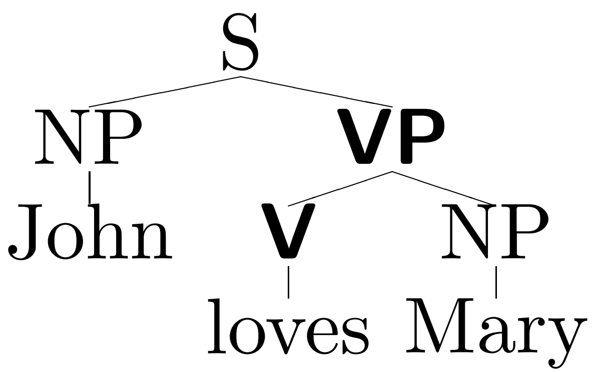
Constituency trees (CFG) represent only bracketing:
= which adjacent constituents are glued tighter to each other.

Dependency trees represent which words depend on which.
+ usually, some agreement/conditioning happens along the edge.

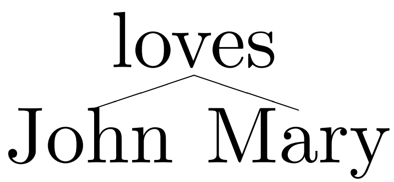
Constituency

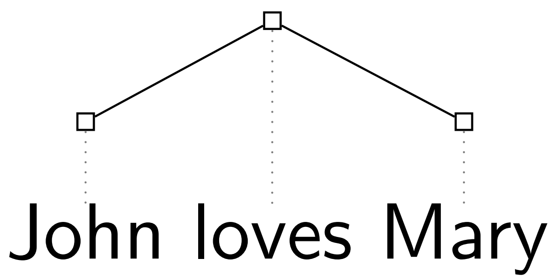
John (loves Mary)

John $_{\text{VP}}$(loves Mary)



Dependency





What Dependency Trees Tell Us

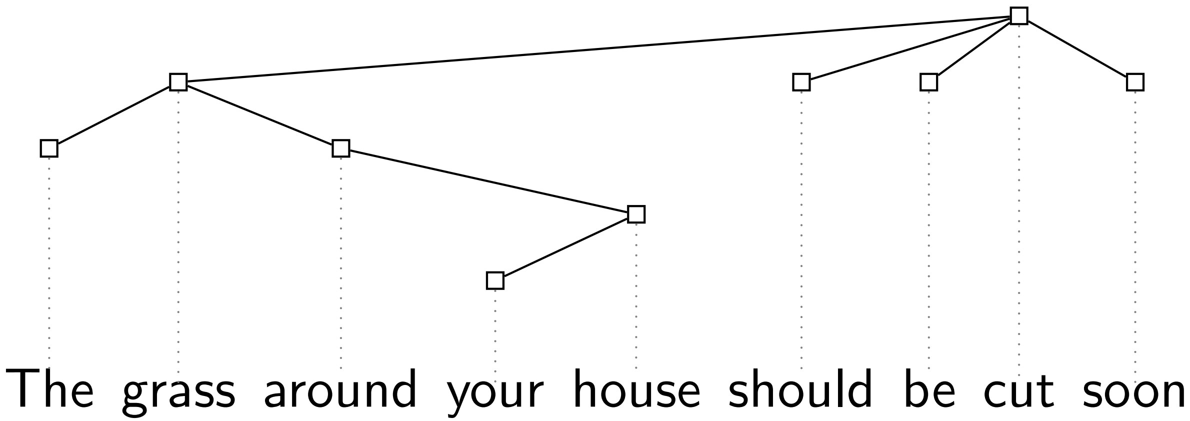
Input:	The **grass** around your house should be **cut** soon.
Google SMT:	**Trávu** kolem vašeho domu by se měl **snížit** brzy.
(Google NMT:	Tráva kolem vašeho domu by měla být brzy zkrácena.)

- Bad lexical choice for cut = sekat/snížit/krájet/řezat/. . .
  - Due to long-distance dependency with grass.
  - One can "pump" many words in between.
  - Could be handled by full source-context (e.g. maxent) model.
- Bad case of tráva.
  - Depends on the chosen active/passive form:

active⇒accusative		passive⇒nominative	
trávu . . .	by**ste** ~~se~~ měl posekat	tráva . . .	by **se** měl**a** posekat
		tráva . . .	by měl**a** **být** poseká**na**

Examples by Zdeněk Žabokrtský, Karel Oliva and others.

Tree vs. Linear Context

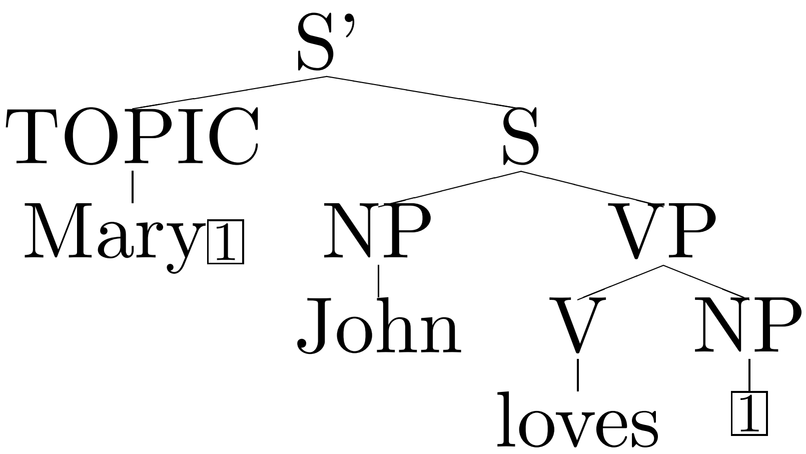


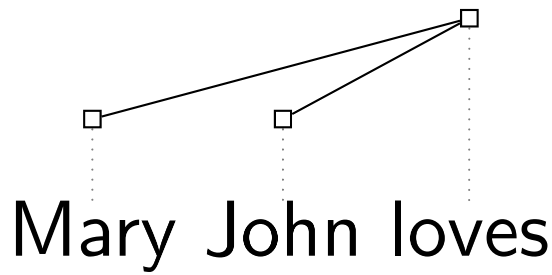
- Tree context (neighbours in the dependency tree):
  - is better at predicting lexical choice than n-grams.
  - often equals linear context:
    Czech manual trees: 50% of edges link neighbours,
    80% of edges fit in a 4-gram.
- Phrase-based MT is a very good approximation.
- Hierarchical MT (phrases with gaps) can even capture the dependency in one phrase:

$$X \rightarrow < \text{the grass } X \text{ should be cut, trávu } X \text{ byste měl posekat} >$$

"Crossing Brackets"

- Constituent outside its father's span causes "crossing brackets."
  - Linguists use "traces" ($\boxed{1}$) to represent this.
- Sometimes, this is not visible in the dependency tree:
  - There is no "history of bracketing".
  - See Holan et al. (1998) for dependency trees including derivation history.



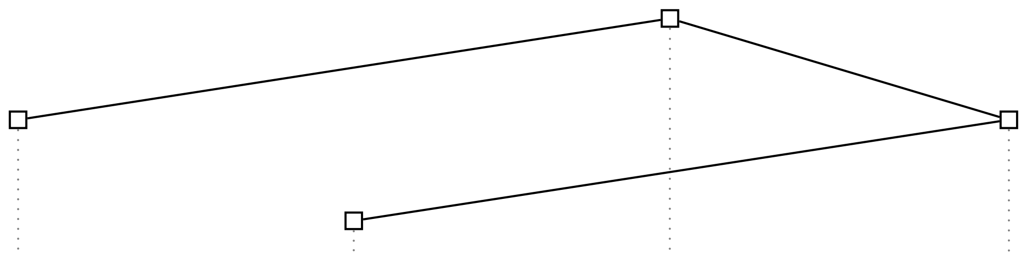


Despite this shortcoming, CFGs are popular and "the" formal grammar for many. Possibly due to the charm of the father of linguistics, or due to the abundance of dependency formalisms with no clear winner (Nivre, 2005).

Non-Projectivity

= a gap in a subtree span, filled by a node higher in the tree.

Ex. Dutch "cross-serial" dependencies, a non-projective tree with one gap caused by saw within the span of swim.



. . . dat Jan kinderen zag zwemmen

. . . that John children saw swim

. . . that John saw children swim.

- 0 gaps $\Rightarrow$ projective tree $\Rightarrow$ can be represented in a CFG.
- ≤ 1 gap & "well-nested" $\Rightarrow$ mildly context sentitive (TAG).

  See Kuhlmann and Möhl (2007) and Holan et al. (1998).

Why Non-Projectivity Matters?

- CFGs cannot handle non-projective constructions:

  Imagine John **grass** saw **being-cut**!

- No way to glue these crossing dependencies together:
  - Lexical choice:

    $$X \rightarrow < \text{grass } X \text{ being-cut, trávu } X \text{ sekat} >$$

  - Agreement in gender:

    $$X \rightarrow < \text{John } X \text{ saw, Jan } X \text{ viděl} >$$

    $$X \rightarrow < \text{Mary } X \text{ saw, Marie } X \text{ viděl}\mathbf{a} >$$

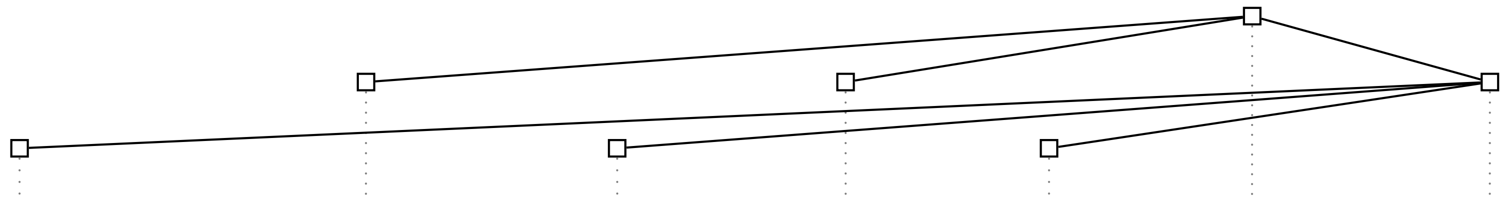
- Phrasal chunks can memorize fixed sequences containing:
  - the non-projective construction
  - and all the words in between! ($\Rightarrow$ extreme sparseness)

Is Non-Projectivity Severe?

Depends on the language.

In principle:

- Czech allows long gaps as well as many gaps in a subtree.



Proti odmítnutí | se | zítra | Petr | v práci | rozhodl | protestovat

Against dismissal | aux-refl | **tomorrow** | Peter | **at work** | decided | **to object**

Peter decided to object against the dismissal at work tomorrow.

In treebank data:

⊖ 23% of Czech sentences contain a non-projectivity.

⊕ 99.5% of Czech sentences are well nested with ≤ 1 gap.

Tectogrammatics: Deep Syntax Culminating

Background: Prague Linguistic Circle (since 1926).

Theory: Sgall (1967), Panevová (1980), Sgall et al. (1986).

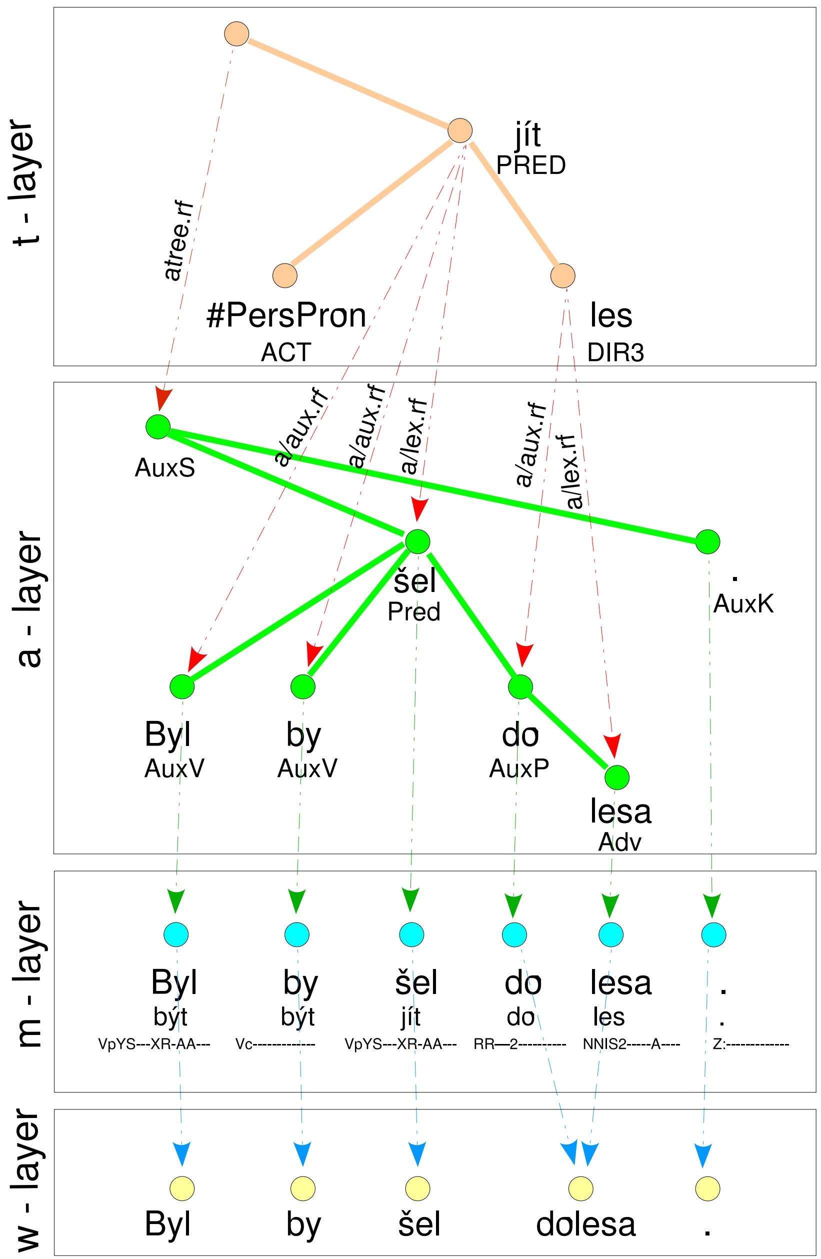
Materialized theory — Treebanks:

- Czech: PDT 1.0 (2001), PDT 2.0 (2006)
- Czech-English: PCEDT 1.0 (2004), PCEDT 2.0 (2012)
- Arabic: PADT (2004)

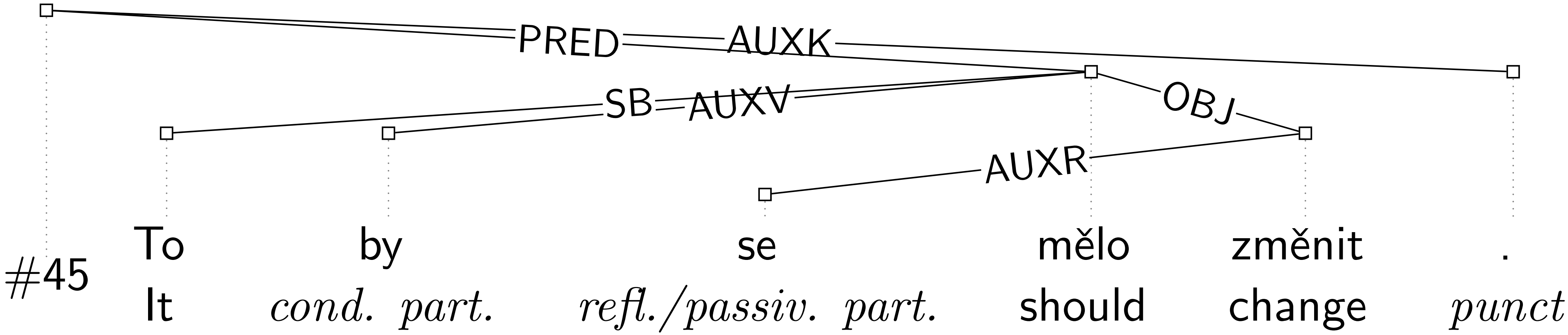
Practice — Tools:

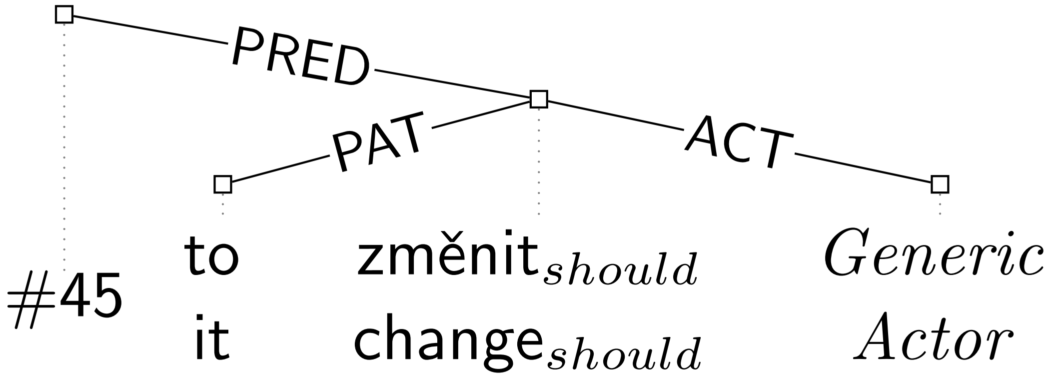
- parsing Czech to surface: McDonald et al. (2005)
- parsing Czech to deep: Klimeš (2006)
- parsing English to surface: well studied (+rules convert to dependency trees)
- parsing English to deep: heuristic rules (manual annotation in progress)
- generating Czech surface from t-layer: Ptáček and Žabokrtský (2006)

Layers in PDT



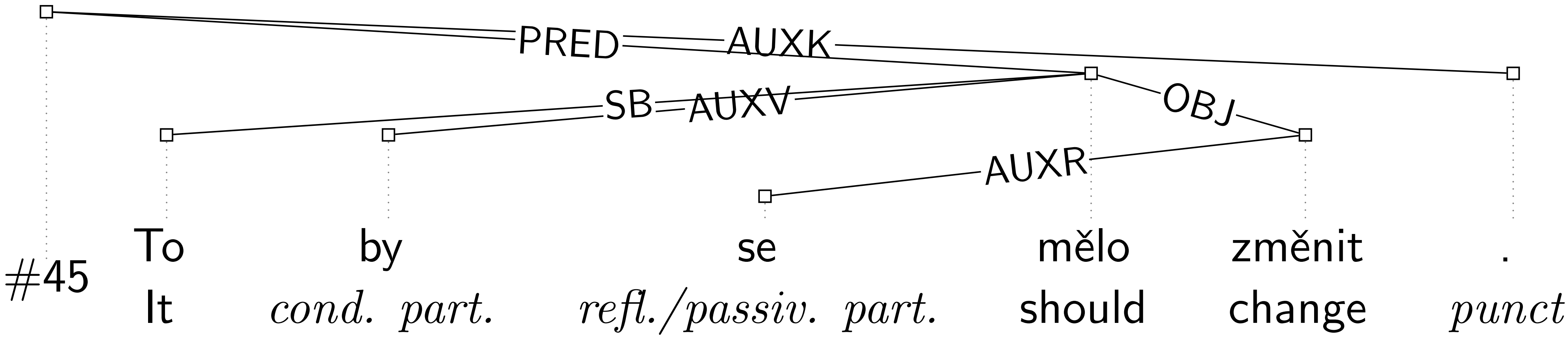
Analytical vs. Tectogrammatical

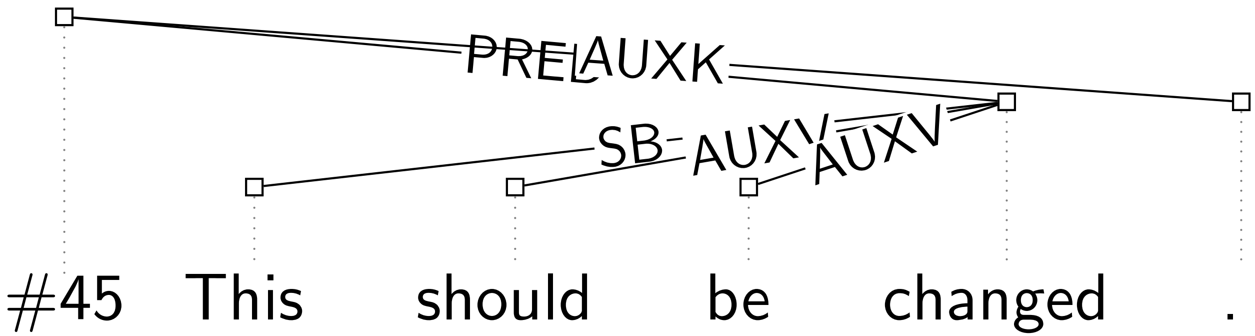




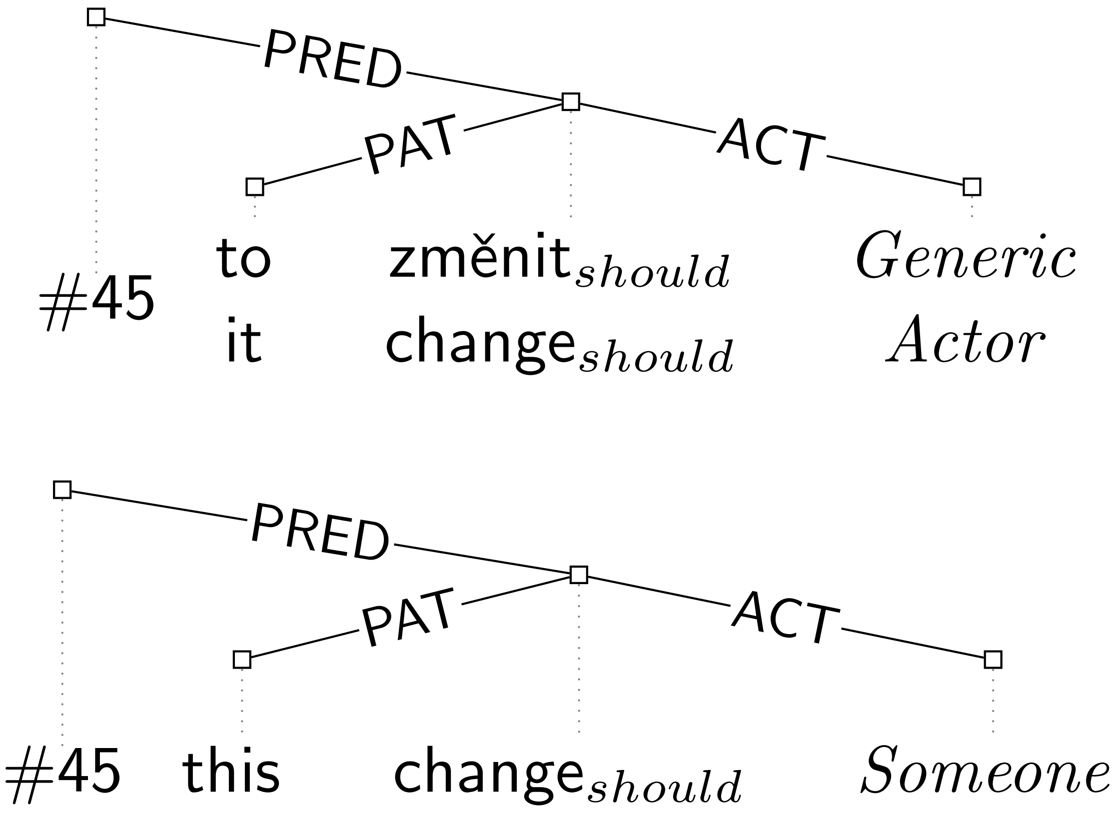
- hide auxiliary words, add nodes for "deleted" participants
- resolve e.g. active/passive voice, analytical verbs etc.
- "full" t-layer resolves much more, e.g. topic-focus articulation or anaphora

Czech and English A-Layer





Czech and English T-Layer



Predicate-argument structure: `change`$_{should}$`(ACT: someone, PAT: it)`

The Tectogrammatical Hope

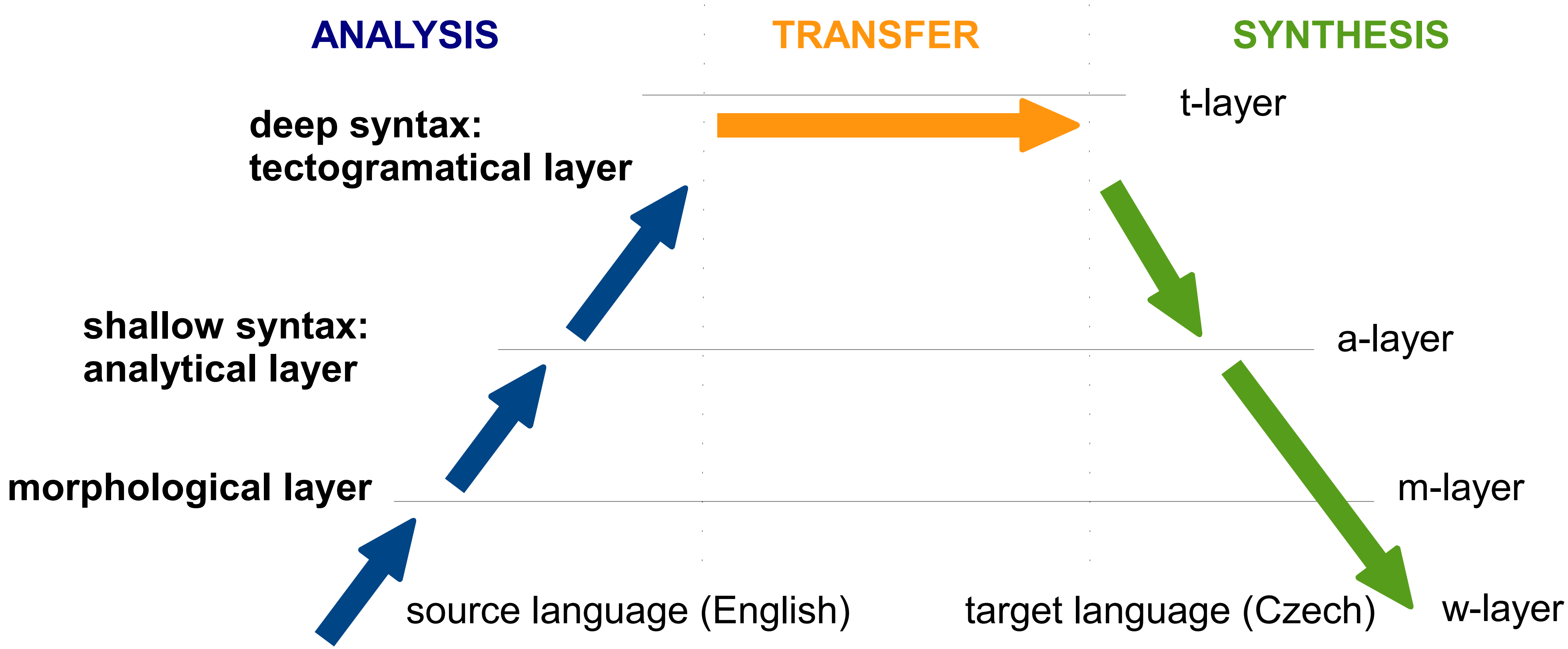
Transfer at t-layer should be easier than direct translation:

- Reduced structure size (auxiliary words disappear).
- Long-distance dependencies (non-projectivites) solved at t-layer.
- Word order ignored / interpreted as information structure (given/new).
- Reduced vocabulary size (Czech morphological complexity).
- Czech and English t-trees structurally more similar
  $\Rightarrow$less parallel data might be sufficient (but more monolingual).
- Ready for fancy t-layer features: co-reference.

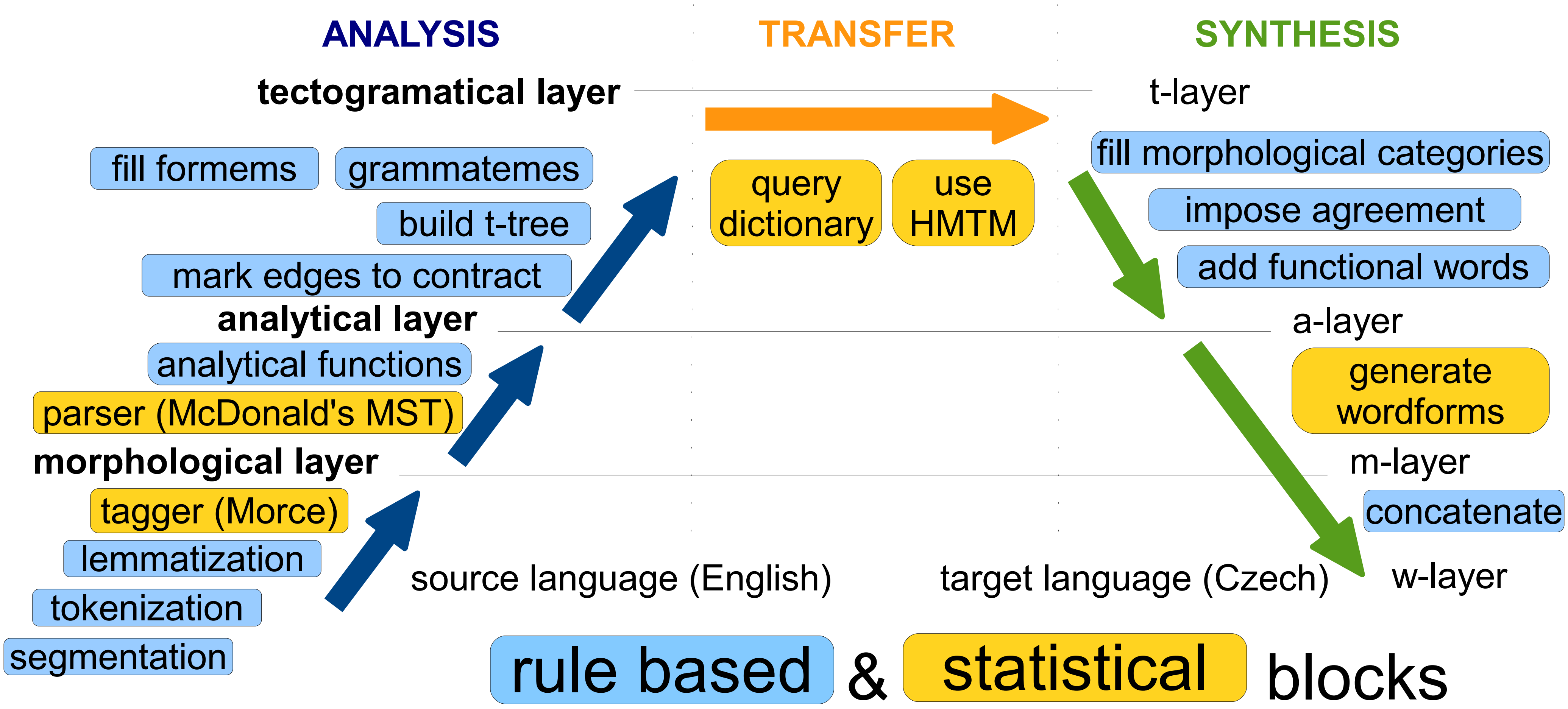
The complications:

- 47 pages documenting data format (PML, XML-based, sort of typed)
- 1200 pages documenting Czech t-structures
  "Not necessary" once you have a t-tree but useful understand or to blame the right people.

“TectoMT Transfer” (1/3)



"TectoMT Transfer" (2/3)



“TectoMT Transfer” (3/3)

To learn more: Slides 6–28 by Martin Popel (2009):

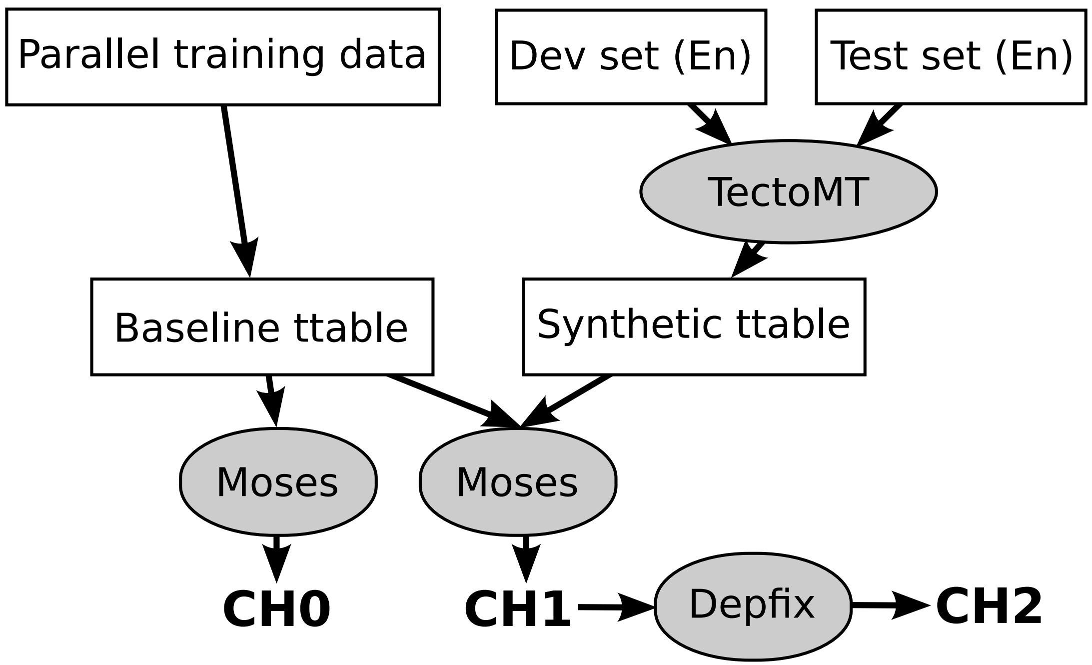
- Illustration of TectoMT transfer.
- Analysis of translation errors.
- Hidden Markov Tree Model (HMTM).

Bad news: TectoMT alone performs poorly.

- Errors cummulate.
- T-layer does bring its independence assumptions.
- No means for plain copy-paste.

Poor Man's System Combination

- Translate input with TectoMT.
- Align translation back to source.
- Extract phrases.
- Add as a separate phrase table.
- MERT to find weights of both phrase tables.



TectoMT Brings Phrases

Input	I saw two green striped cats.
TectoMT Output	Viděl jsem dvě zelené pruhované kočky.

Phrases extracted:

I saw	=	Viděl jsem
I saw two	=	Viděl jsem dvě
...		...
two	=	dvě
two green	=	dvě zelené
two green striped	=	dvě zelené pruhované
two green striped cats	=	dvě zelené pruhované kočky
...		...

TectoMT Brings Phrases

The output of TectoMT covers (most of) the source.

- Long and short phrases, one form only.

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	**kočky**	.
	pily	**dvě**	zelená	pruhovaná	koček	
	. . .	dvou	**zelené**	**pruhované**	kočkám	
	viděl	dvěma	zelení	pruhovaní	kočkách	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	. . .		zelených	pruhovaných		
viděl jsem			zelenými	pruhovanými		
viděla jsem			. . .	. . .		

TectoMT Brings Phrases

The output of TectoMT covers (most of) the source.

- Long and short phrases, one form only.

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	**kočky**	.
	pily	**dvě**	zelená	pruhovaná	**kočky**	
	. . .	**dvě**	**zelené**	**pruhované**	koček	
	viděl	dvou	**zelené**	**pruhované**	kočkám	
	viděla	dvěma	zelení	pruhovaní	kočkách	
	. . .	dvěmi	zeleného	pruhovaného	kočkami	
viděl jsem			zelených	pruhovaných		
viděl jsem			zelenými	pruhovanými		
viděla jsem		**dvě zelené**		**pruhované kočky**		
		dvě zelené pruhované kočky				

Chimera: Complex Combination

Chimera () was beating everyone in 2013–2015.

- Input:
  - Famous cases also relate to graphic elements.
-  TectoMT translates using deep syntax:
  - Slavné případy se **být** týkají grafick**é** prvk**y**.
-  PBMT adds 200M en-cs sents and 3,6G cs words:
  - Slavné případy se týkají také grafick**é** prvk**y**.
-  Automatic error correction for agreement or negation:
  - Slavné případy se týkají také grafických prvků.

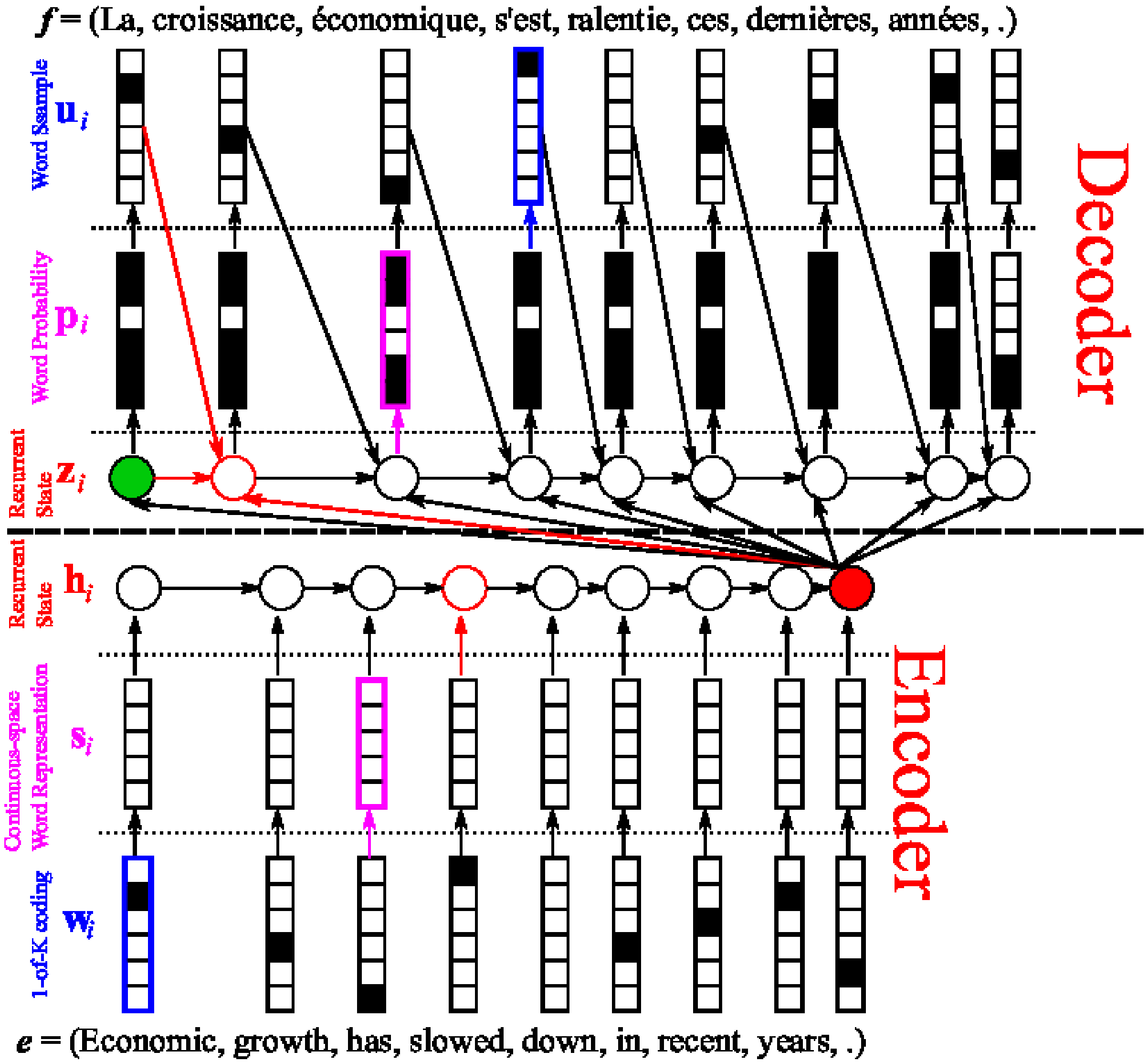
- Google SMT: Slavné případy **týkat** i grafick**é** prvk**y**.
- Google NMT: Slavné případy se také týkají grafických prvků.

Summary So Far

- Meaning of sentences is usually compositional.
- Syntax describes the composition.
  - Expressed with various surface features (e.g. case).
  - Syntactic context more important than linear context.
  - Non-projectivity: composition $\neq$ concatenation.
- Syntax comes at a cost:
  - Theory you have to learn.
  - More complex search space.
  - Cummulation of errors.
- Syntactic SMT did not outperform PBMT in general.
  - We successfully utilized syntax only within PBMT.

And Now for Something...

Remember: $p(e_1^I|f_1^J) = p(e_1|f_1^J) \cdot p(e_2|e_1, f_1^J) \cdot p(e_3|e_2, e_1, f_1^J) \ldots$



Some Advanced Topics in NMT

- Self-Attention

- Linguistic Features in NMT.
- Multi-Task Training.
- Multi-Lingual MT.

These can be done with:

a) dedicated architectures, e.g. Eriguchi et al. (2017)
b) **hacked input/output for seq2seq.**

- Learned Representations.

Self-Attention (Transformer Model)

See slides 46–53 by Jindřich Libovický, Lecture 9, pages 53–60.

Three uses of multi-head attention in Transformer

- Encoder-Decoder Attention:
  - Q: previous decoder layers; K $=$ V: outputs of encoder
  - $\Rightarrow$ Decoder positions attend to all positions of the input.
- Encoder Self-Attention:
  - Q $=$ K $=$ V: outputs of the previous layer of the encoder
  - $\Rightarrow$ Encoder positions attend to all positions of previous layer.
- Decoder Self-Attention:
  - Q $=$ K $=$ V: outputs of the previous decoder layer.
  - Masking used to prevent depending on future outputs.
  - $\Rightarrow$ Decoder attends to all its previous outputs.

Linguistic Features in NMT

- Source word factors easy to incorporate:
  - Concatenate embeddings of the various factors.
  - POS tags, morph. features, source dependency labels help en↔de and en→ro (Sennrich and Haddow, 2016).

- Target word factors:
  - Interleave for morphology: (Tamchyna et al., 2017)

Src	there are a million different kinds of pizza .
Baseline (BPE)	existují miliony druhů piz@@ zy .
Interleave	VB3P existovat NNIP1 milion NNIP2 druh NNFS2 pizza Z: .

  - Interleave for syntax: (Nadejde et al., 2017)

Src BPE	Obama receives Net+ an+ yahu in the capital of USA
Tgt	NP Obama ((S[dcl]\NP)/PP)/NP receives NP Net+ an+ yahu PP/NP in N

Suspicious Results on Multi-Tasking

My students Dan Kondratyuk and Ronald Cardenas retried Nadejde et al. (2017) with:

- sequence-to-sequence model,
- Transformer model.

Predicting target syntax using:

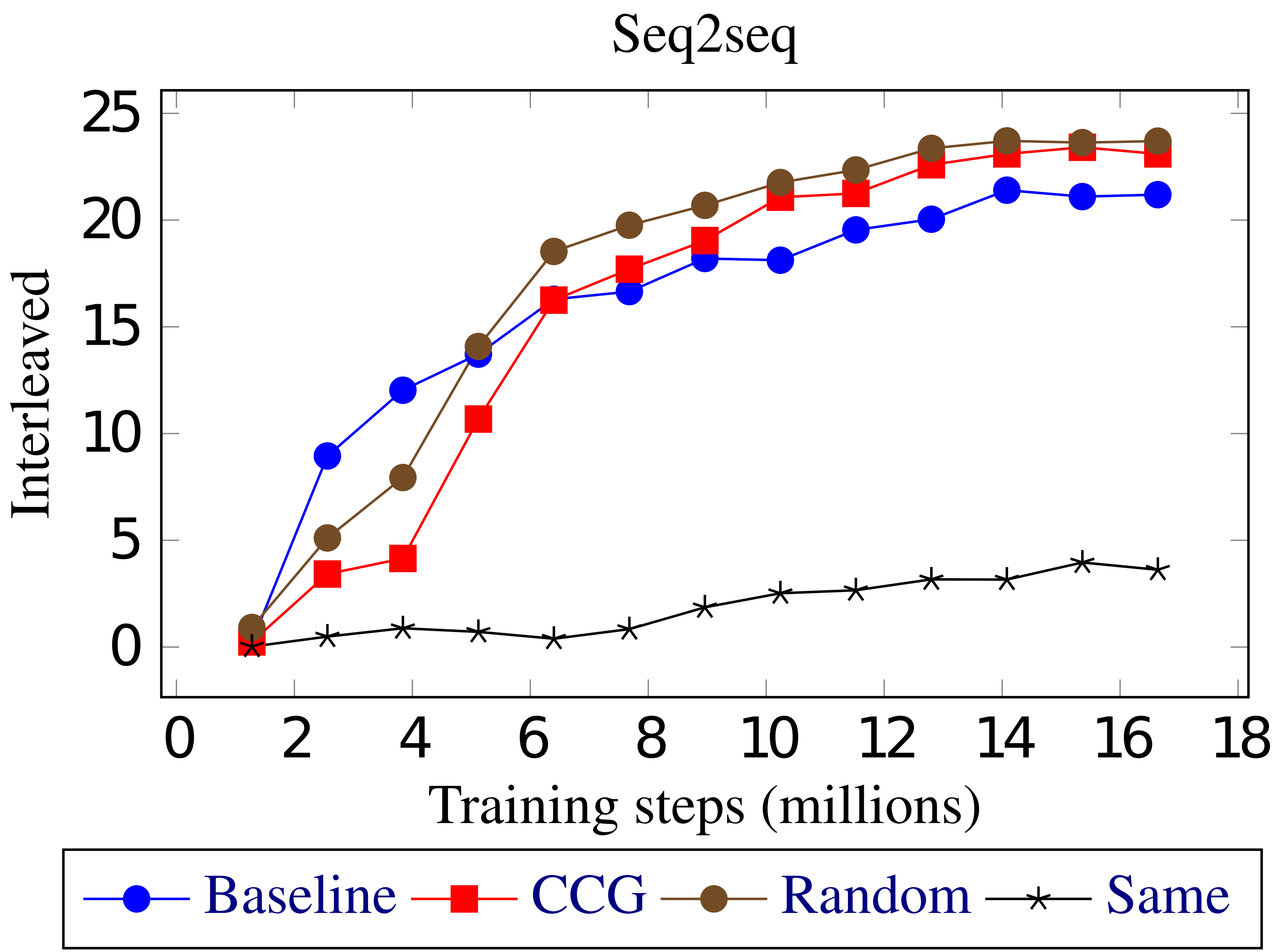
- a secondary decoder

  (The sequence of CCG tags may not match the translated sentence.)
- interleaving.

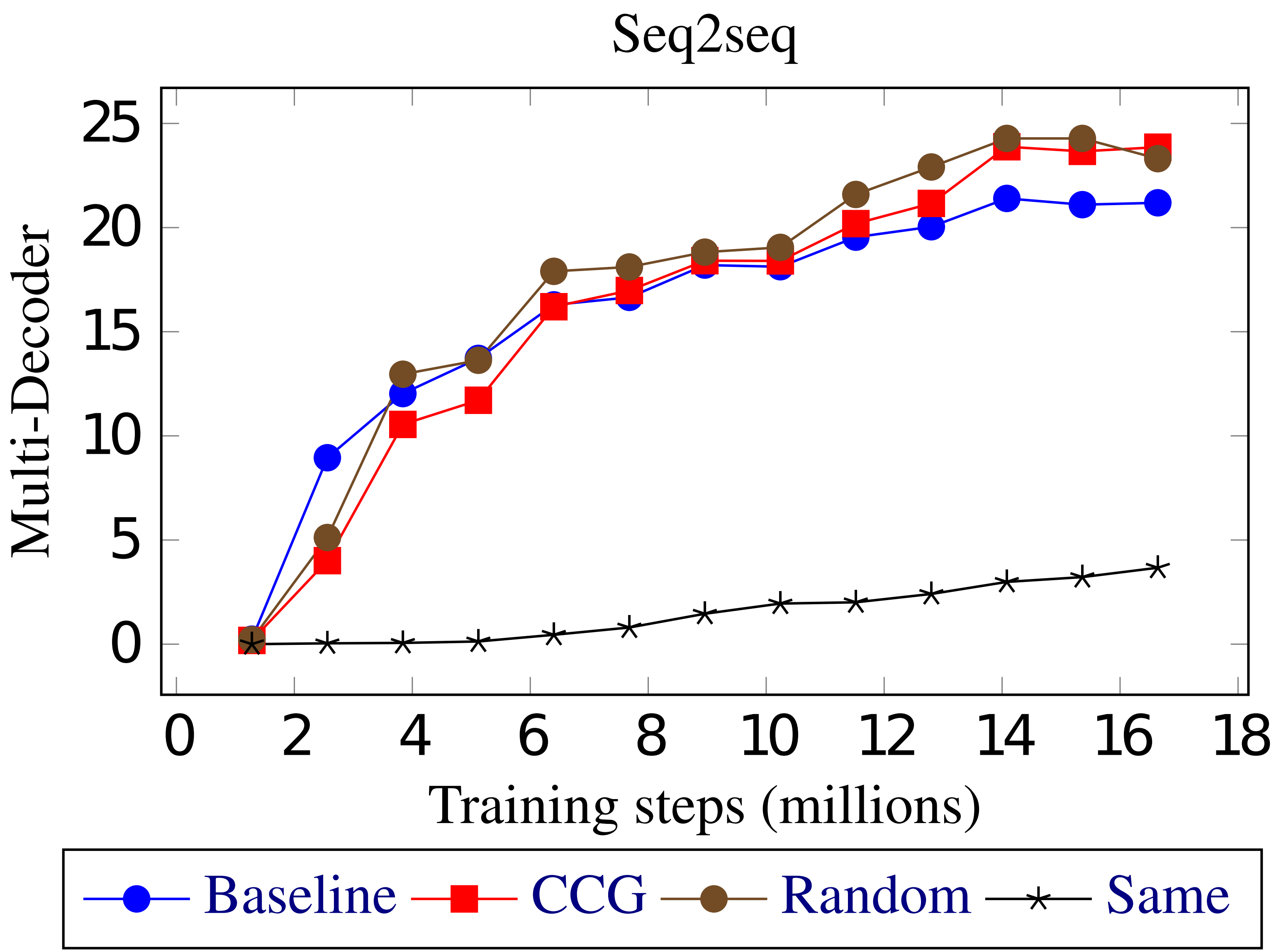
As tags, they used:

- correct CCG tags, • random tags, • a single dummy tag.

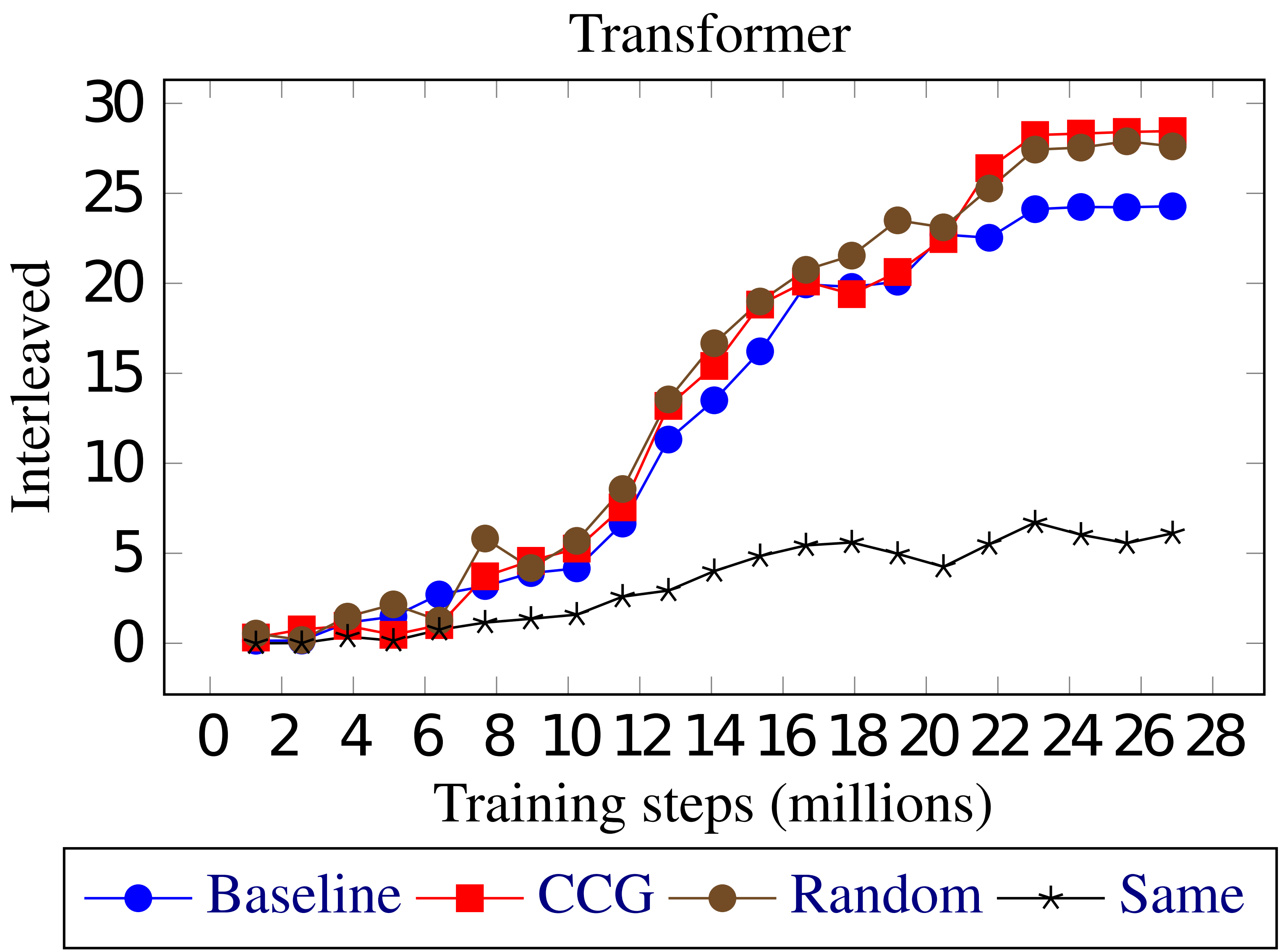
Suspicious Results on Multi-Tasking



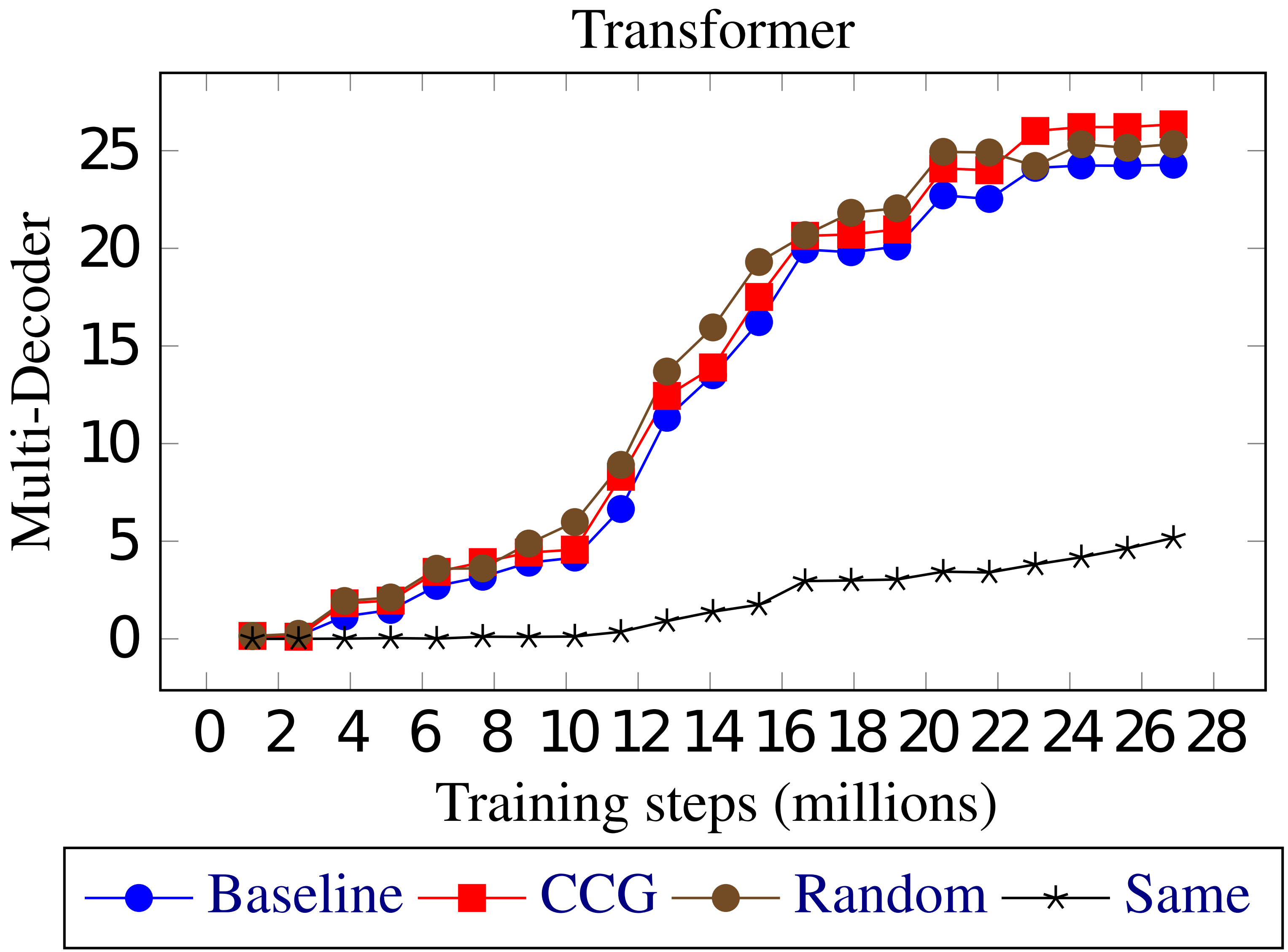
Suspicious Results on Multi-Tasking



Suspicious Results on Multi-Tasking



Suspicious Results on Multi-Tasking



Multi-Lingual MT

. . . simply feed in various language pairs.

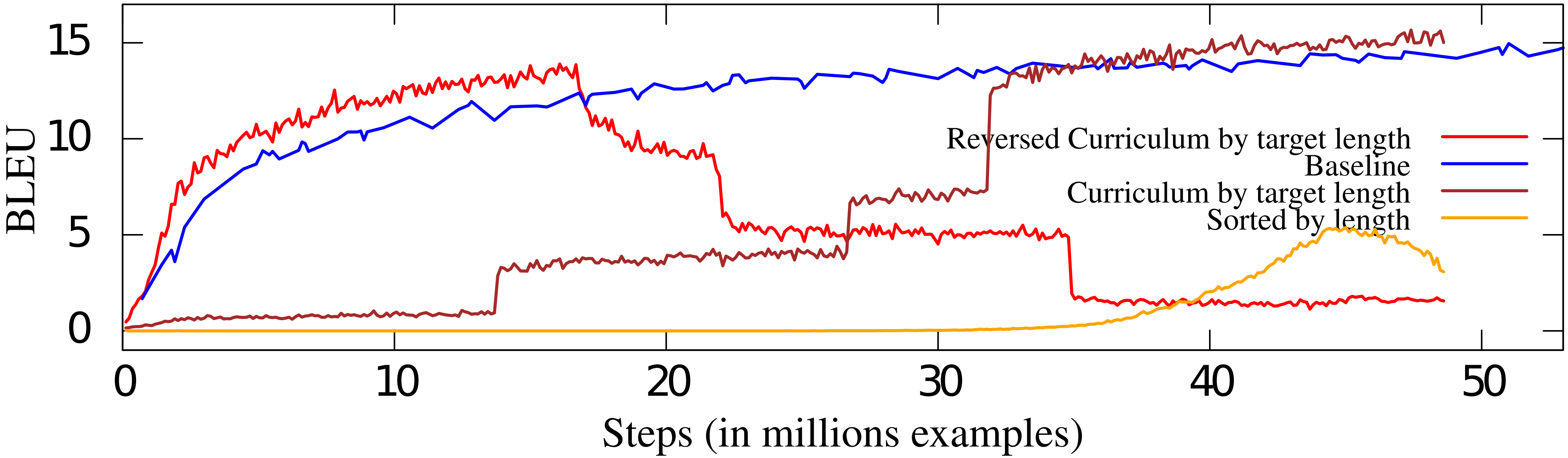
Source Sent 1 (De)	**2en** versetzen Sie sich mal in meine Lage !
Target Sent 1 (En)	put yourselves in my position .
Source Sent 2 (En)	**2nl** I flew on Air Force Two for eight years .
Target Sent 2 (Nl)	ik heb acht jaar lang met de Air Force Two gevlogen .

- The model of the same size will learn both pairs.
- Hopefully benefiting from various similarities.
- Risk of catastrophic forgetting.

See Johnson et al. (2016) or Ha et al. (2017).

Catastrophic Forgetting

- Kocmi and Bojar (2017) explore curriculum learning:
  - Start with simpler sentences first, add complex ones later.
- When “simpler” mean “shorter”:
  - Clear jumps in score as bins of longer sentences are allowed.
  - Reversed curriculum <u>unlearns</u> to produce long sentences.



Surprising Results with Multiling. Transfer

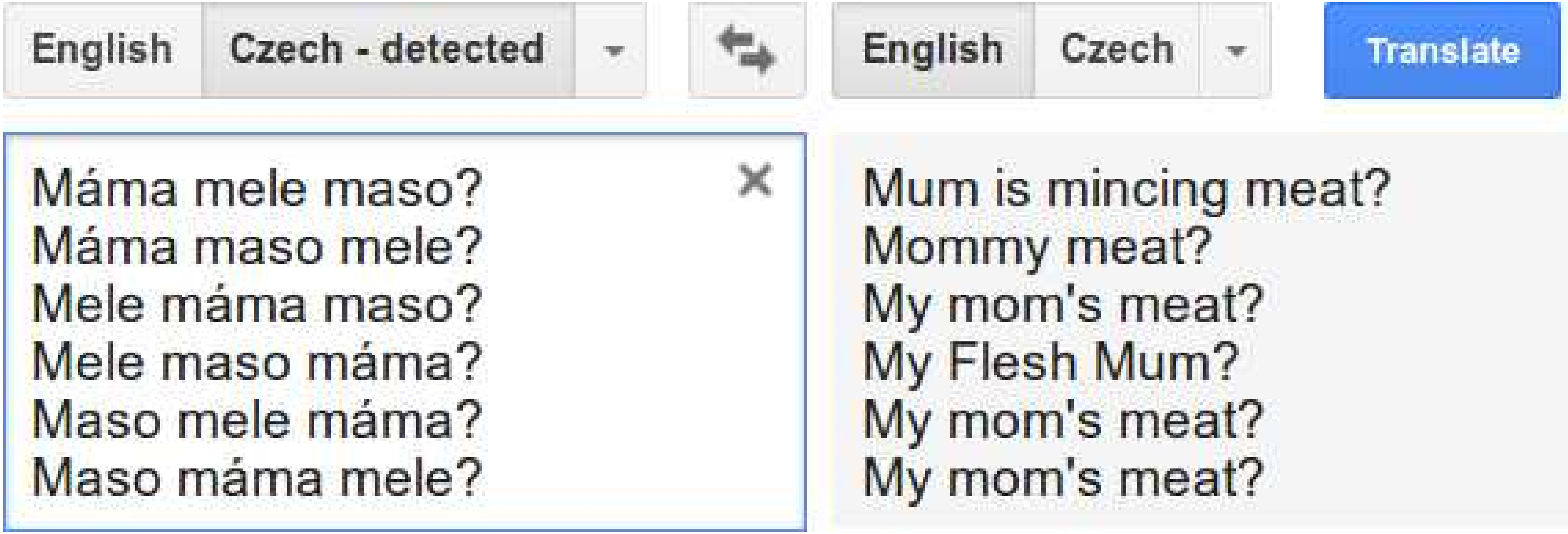
Language pair	Baseline		Start with an English-to-Czech model: Direct transfer		Transformed vocab	
	BLEU	Steps	BLEU	Steps	BLEU	Steps
English-to-Odia	3.54	45k	0.04	47k	**6.38**	**38k**
English-to-Estonian	8.13	**95k**	**14.48**	180k	14.18	175k
English-to-Finnish	14.42	420k	16.12	**255k**	**16.73**	270k
English-to-German	36.72	270k	38.58	190k	**39.28**	**110k**
English-to-Russian	27.81	1090k	25.50	630k	**28.65**	**450k**
English-to-French	33.72	820k	34.41	**660k**	**34.46**	720k
French-to-Spanish	31.10	390k	31.55	435k	**31.67**	**375k**

Best score and lowest training time in each row in bold.

- Reusing the knowledge of English source can help really a lot.
- Pre-training Transformer on fully unrelated language pair can help, too.

Learned Representations

- Deep learning researchers easily claim that NNs learn the <u>meaning</u> of the sentences.
- This is possible, but not achieved in practice, yet:



Translating to Summarize

Input:

legendární slovenská punkrocková kapela extip se letos vrátila na pódia poté, co vyšla v reedici její debutová deska pekný, škaredý deň, kterou přehraje 1. prosince na sedmičce na strahově. soubor nezanikl, i když bratislavskou punkovou scénu v devadesátých letech rozložily drogy. své zkušenosti s tím má kytarista sveto korbel, který odpovídal na otázky novinek.

Human Output:

slovenská punková legenda extip se vrátila

Translating to Summarize

Input:

legendární slovenská punkrocková kapela extip se letos vrátila na pódia poté, co vyšla v reedici její debutová deska pekný, škaredý deň, kterou přehraje 1. prosince na sedmičce na strahově. soubor nezanikl, i když bratislavskou punkovou scénu v devadesátých letech rozložily drogy. své zkušenosti s tím má kytarista sveto korbel, který odpovídal na otázky novinek.

Human Output:

slovenská punková legenda extip se vrátila

"Summarized" by Google Transformer Model:

slovenská kapela extip se vrací do prahy

Meaning Understood?

Input:

legendární slovenská punkrocková kapela extip se letos vrátila na pódia poté, co vyšla v reedici její debutová deska pekný, škaredý deň, kterou přehraje 1. prosince na sedmičce na **strahově**. soubor nezanikl, i když bratislavskou punkovou scénu v devadesátých letech rozložily drogy. své zkušenosti s tím má kytarista sveto korbel, který odpovídal na otázky novinek.

Human Output:

slovenská punková legenda extip se vrátila

“Summarized” by Google Transformer Model:

slovenská kapela extip se vrací do **prahy**

Meaning Understood? Surely Not.

na strahově	slovenská kapela extip se vrací do **prahy**
v o2 aréně	slovenská kapela extip se vrací do **prahy**
na hradecku	slovenská kapela extip se vrací do **čech**
u vajgaru	slovenská kapela extip se vrací do **prahy**

Not Understood.

na strahově	slovenská kapela extip se vrací do **prahy**
v o2 aréně	slovenská kapela extip se vrací do **prahy**
na hradecku	slovenská kapela extip se vrací do **čech**
u vajgaru	slovenská kapela extip se vrací do **prahy**
ve stromovce	slovenská kapela extip se vrací na scénu. tentokrát kvůli drogám v reedici. s. s. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. i. m. . . . . m. . . . m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. m. . m. m. m. m. m. m. m. m. m. m.

Many More Details

. . . see the 234 slides (ACL 2016 tutorial, 58MB):

`https://sites.google.com/site/acl16nmt/`

The basics of NMT are here:

- slides 14-19, 24-25: NMT for one word, overview.
- slides 47-53: Recurrent neural LM.
- slides 84-95: Encoder-decoder, decoding.
- slides 130-140: Encoder-decoder with attention.
- slides 192-204: Multi-task and multi-lingual.
- . . . but also the basics of NN, e.g. GRU (slides 72-79).

Summary

Linguistic features added:

- as factors (word-level annotations) to phrase-based MT
- as deep syntax, organizing the whole process.
- as source factors to NMT.
- as secondary tasks to NMT.

SMT (and transfer-based MT) suffer from unjustified assumptions.

Neural networks:

- get rid of most of the assumptions.
- but are very expensive to train.
- and it is still not clear how much generalization is learned.

References

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 182–189, Athens, Greece, March. Association for Computational Linguistics.

Ondřej Bojar and Jan Hajič. 2008. Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation. In Proceedings of the Third Workshop on Statistical Machine Translation, pages 143–146, Columbus, Ohio, June. Association for Computational Linguistics.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving Translation Model by Monolingual Data. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 330–336, Edinburgh, Scotland, July. Association for Computational Linguistics.

Ondřej Bojar. 2007. English-to-Czech Factored Machine Translation. In Proceedings of the Second Workshop on Statistical Machine Translation, pages 232–239, Prague, Czech Republic, June. Association for Computational Linguistics.

Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. 2017. Learning to parse and translate improves neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 72–78, Vancouver, Canada, July. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2017. Effective strategies in zero-shot neural machine translation. CoRR, abs/1711.07893.

Jan Hajič and Barbora Hladká. 1998. Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In Proceedings of COLING-ACL Conference, pages 483–490, Montreal, Canada.

Tomáš Holan, Vladislav Kuboň, Karel Oliva, and Martin Plátek. 1998. Two Useful Measures of Word Order Complexity. In A. Polguere and S. Kahane, editors, Proceedings of the Coling '98 Workshop: Processing of

References

Dependency-Based Grammars, Montreal. University of Montreal.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. CoRR, abs/1611.04558.

Václav Klimeš. 2006. Analytical and Tectogrammatical Analysis of a Natural Language. Ph.D. thesis, ÚFAL, MFF UK, Prague, Czech Republic.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum Learning and Minibatch Bucketing in Neural Machine Translation. In Proceedings of Recent Advances in NLP (RANLP 2017).

Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In Proc. of EMNLP.

Marco Kuhlmann and Mathias Möhl. 2007. Mildly context-sensitive dependency languages. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 160–167, Prague, Czech Republic, June. Association for Computational Linguistics.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In Proceedings of HLT/EMNLP 2005, October.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. Natural Language Engineering, 7(3):207–223.

Maria Nadejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. 2017. Predicting target language ccg supertags improves neural machine translation. In Proceedings of the Second Conference on Machine Translation, Volume 1: Research Paper, pages 68–79, Copenhagen, Denmark, September. Association for Computational Linguistics.

Joakim Nivre. 2005. Dependency Grammar and Dependency Parsing. Technical Report MSI report 05133, Växjö University: School of Mathematics and Systems Engineering.

References

Jarmila Panevová. 1980. Formy a funkce ve stavbě české věty [Forms and functions in the structure of the Czech s
Academia, Prague, Czech Republic.

Jan Ptáček and Zdeněk Žabokrtský. 2006. Synthesis of Czech Sentences from Tectogrammatical Trees. In Proc. of TSD, pages 221–228.

Adwait Ratnaparkhi. 1996. A Maximum Entropy Part-Of-Speech Tagger. In Proceedings of the Empirical Methods in Natural Language Processing Conference, University of Pennsylvania, May.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In Proceedings of the First Conference on Machine Translation, pages 83–91, Berlin, Germany, August. Association for Computational Linguistics.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. The Meaning of the Sentence and Its Semantic and Pragmat
Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

Petr Sgall. 1967. Generativní popis jazyka a česká deklinace. Academia, Prague, Czech Republic.

Aleš Tamchyna, Marion Weller-Di Marco, and Alexander Fraser. 2017. Modeling target-side inflection in neural machine translation. In Proceedings of the Second Conference on Machine Translation, Volume 1: Research Paper, pages 32–42, Copenhagen, Denmark, September. Association for Computational Linguistics.

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Semi-supervised model adaptation for statistical machine translation. Machine Translation, 21(2):77–94.