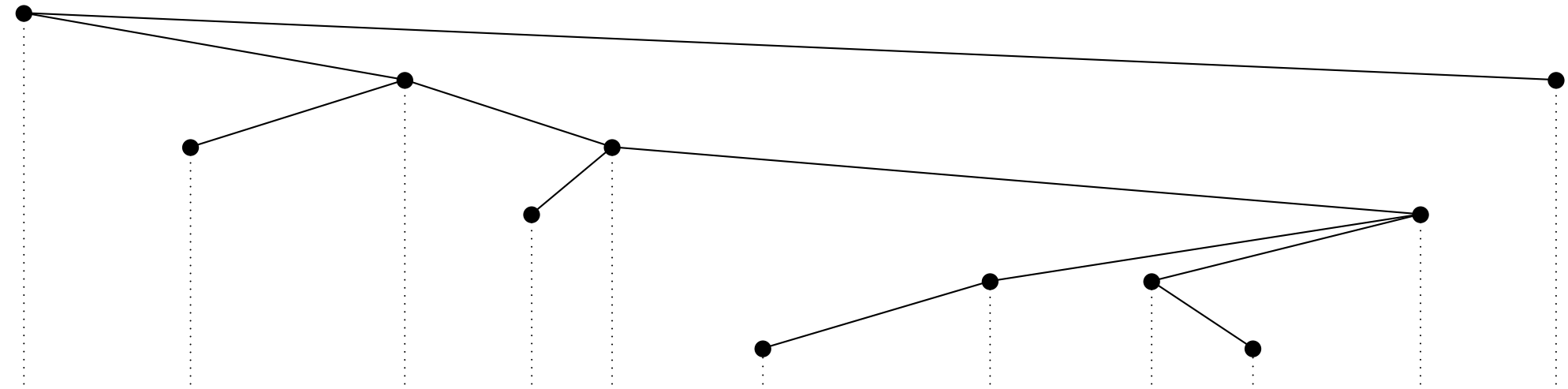
Bad News, NLP Hacking and Feature Fishing

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

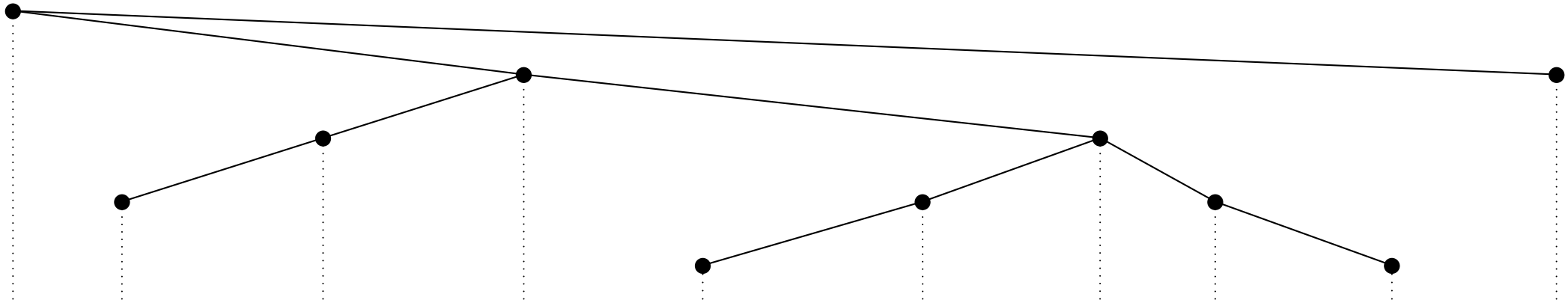
Outline

- Bad news: Syntax-based transfer is hard.
- NLP hacking:
  - Hinglish.
  - Source valency information.
- Proper feature fishing (near future experiments):
  - Phrase table marking, not filtering.
  - Source context features.

Idea: 1: Observe a Pair of Trees. . .

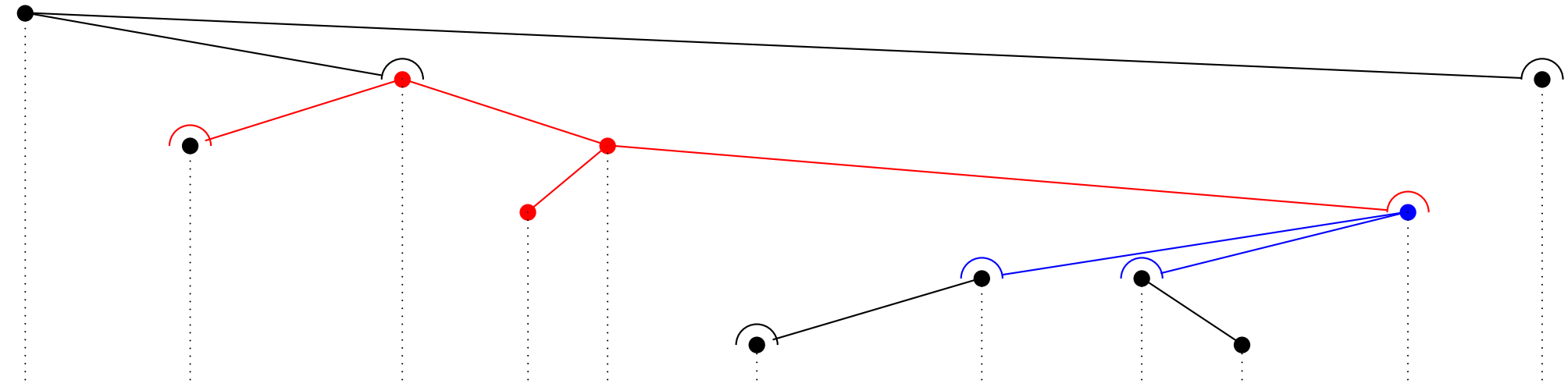


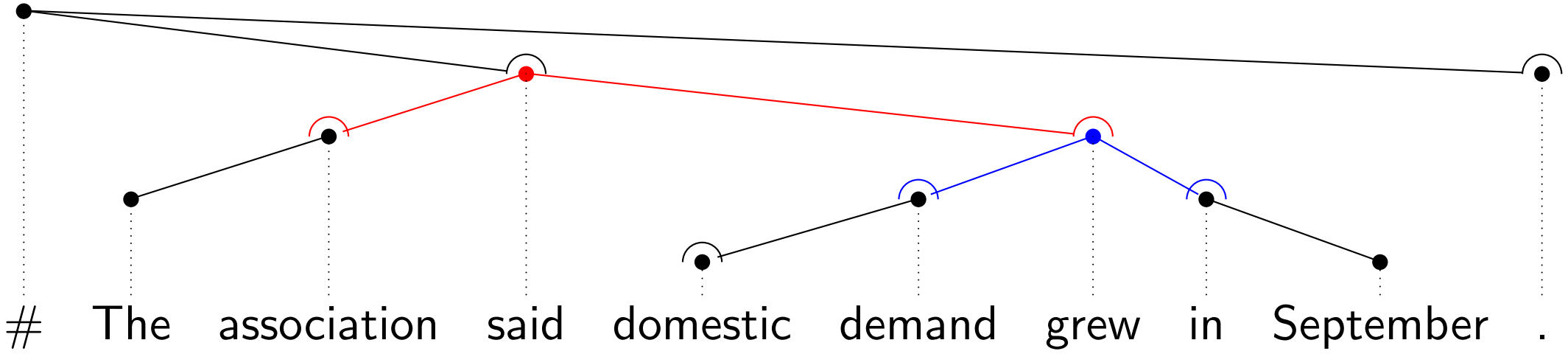
Asociace uvedla , že domácí poptávka v září stoupla .



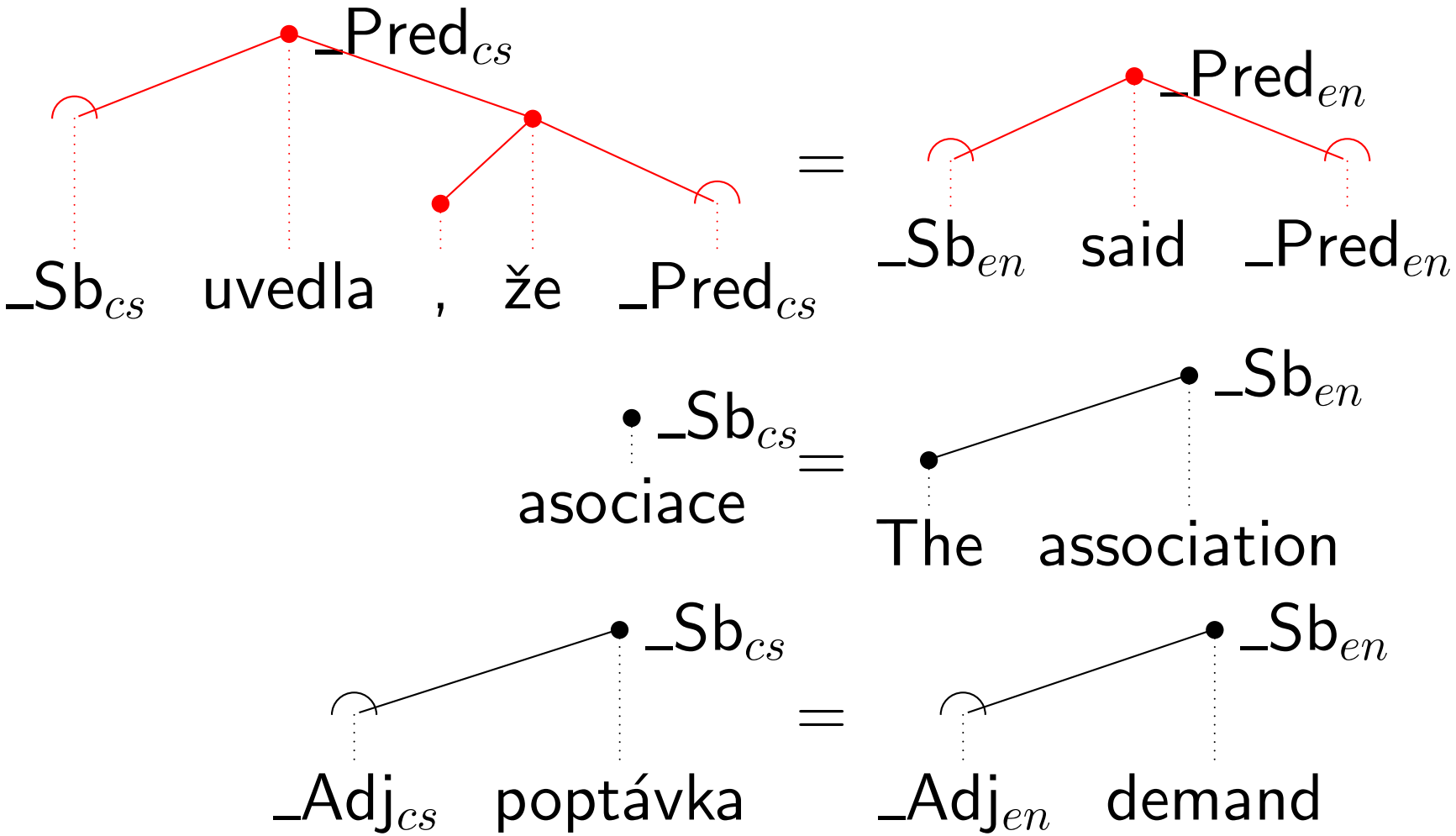
The association said domestic demand grew in September .

2: . . . Decompose into Treelets. . .





3: . . . Collect Dictionary of Treelets



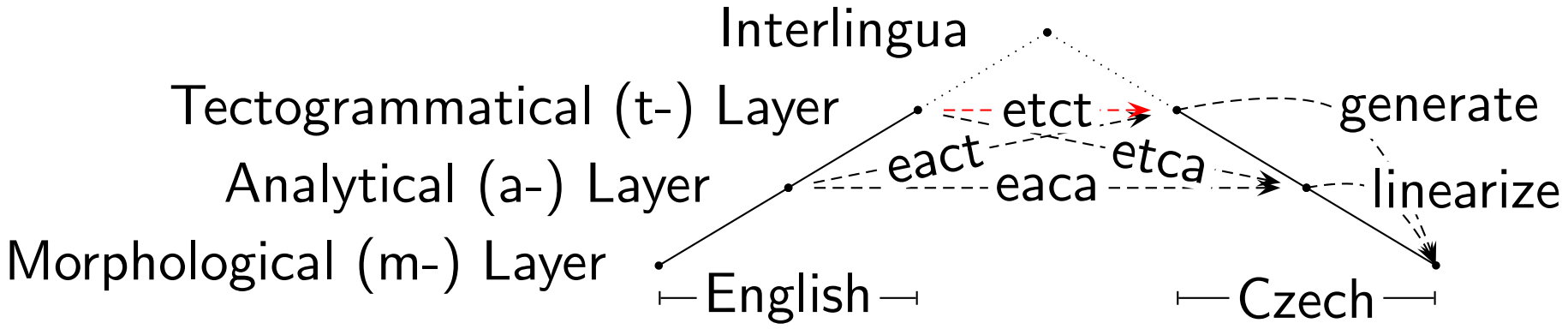
Synchronous Tree Substitution Grammar, e.g. Čmejrek (2006).
More details in Bojar and Čmejrek (2007).

Moses-like Decoding STSG

Given an input dependency tree:

- decompose it into known treelets,
- replace treelets by their treelet translations,
- join output treelets and produce output final tree; linearize or generate plaintext.

Applicable at or across layers:

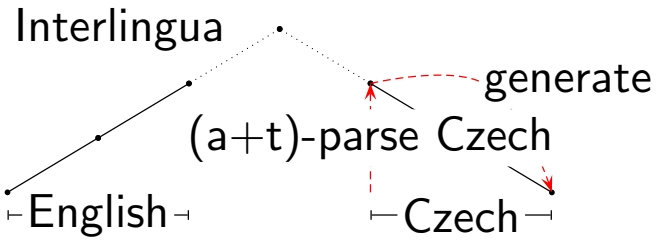


In Reality, t-nodes are not Atomic!

t-nodes have ~25 attributes: t-lemma, functor, gender, person, tense, iterativeness, dispositional modality, . . .

Upper Bound on MT Quality via t-layer:

- Analyse Czech sentences to t-layer.
- Optionally ignore some node attributes.
- Generate Czech surface.
- Evaluate BLEU against input Czech sentences.



	BLEU
Full automatic t-layer, no attributes ignored	36.6±1.2
Ignore sentence mood (assume indicative)	36.6±1.2
Ignore verbal fine-grained info (resultativeness, . . .)	36.6±1.2
Ignore verbal tense, aspect, . . .	24.9±1.1
Ignore all grammatemes	5.3±0.5

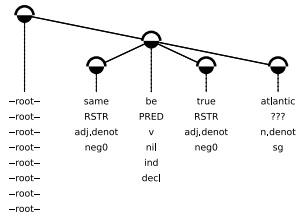
⇒ Node attributes obviously very important.

BLEU Scores for STSG Transfer

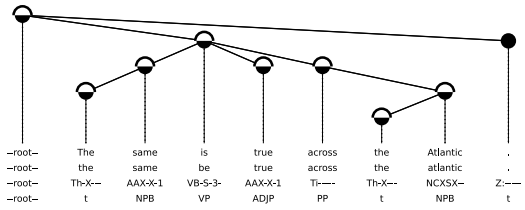
- Identical decoder, only the structure + node labels differ.

Layers \ Language Models	no LM	with LM
epcp, atomic nodes	8.65±0.55	10.90±0.63
eaca, atomic nodes	6.59±0.52	8.75±0.61
etct, generated attrs, fixed structure	5.31±0.53	5.61±0.50
etct, atomic nodes, all attributes	1.61±0.33	2.56±0.35
etct, atomic nodes, just t-lemmas	0.67±0.19	-

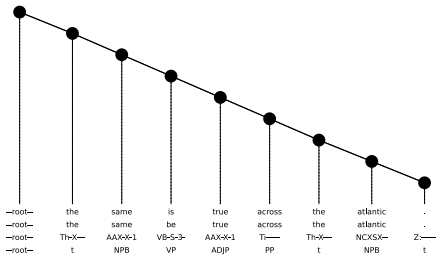
t:



a:



p:

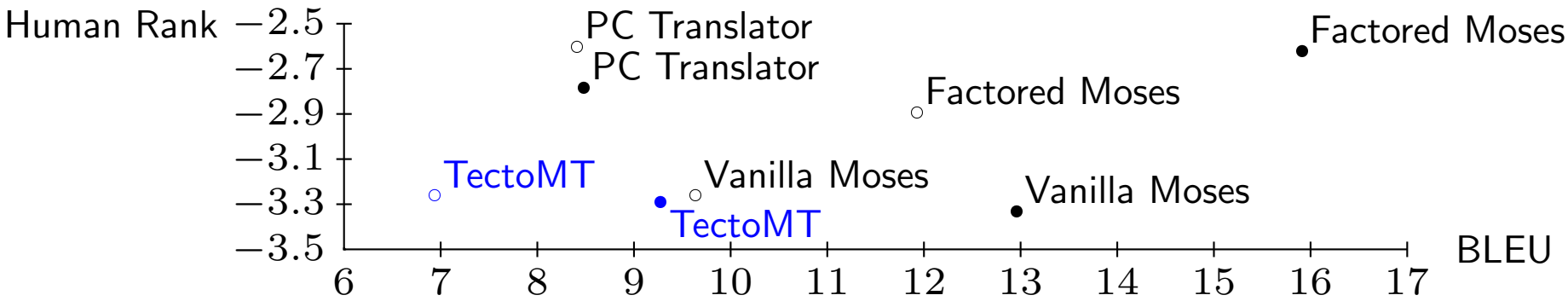


Why Is the t-layer So Poor?

- **Cumulation of Errors**:
  - e.g. 93% tagging * 85% parsing * 93% tagging * 92% parsing = 67%
  - We were using ancient tools: (Ratnaparkhi, 1996), (Collins, 1996), . . .
- **Data Loss** due to incompatible structures:
  - Any error in either of the parses and/or the word-alignment prevents treelet pair extraction.
- **Data Sparseness** when attributes or treelet structure atomic:
  - E.g. different case requires a new treelet pair.
  - There is no adjunction in STSG, new modifier needs a new treelet pair.
- **Combinatorial Explosion** when generating attributes dynamically:
  - Target treelets are first fully built, before combination is attempted.
  - Abundance of t-node attribute combinations
    $\Rightarrow$ e.g. lexically different translation options pushed off the stack
    $\Rightarrow$ n-bestlist varies in unimportant attributes.

Don't Dump Deep Syntax Yet

WMT08 Results	In-domain •		Out-of-domain ○	
	BLEU	Rank	BLEU	Rank
Factored Moses	15.91	-2.62	11.93	-2.89
PC Translator	8.48	-2.78	8.41	-2.60
TectoMT	9.28	-3.29	6.94	-3.26
Vanilla Moses	12.96	-3.33	9.64	-3.26
etct	4.98	-	3.36	-



- TectoMT ranked comparably to vanilla Moses (BLEU is wrong anyway).
- TectoMT great for preparing rich data.

NLP Hacking vs. Feature Fishing

NLP Hacking:

= Hardcoded behaviour based on some (rich/deep) feature.

- Well motivated but not well built into general search.
- Usually equivalent to deterministic modification of the source language.

Feature Fishing:

= Search properly considers additional features.

- Each feature softly steers the search.
- Data (training/optimization) decide which feature is important.
- The research goal is to have a few most informative features.

Feature Fishing $\sim$ Discriminative Training; also tomorrow.

NLP Hacking: Hinglish

Bojar et al. (2008) use TectoMT for rule-based reordering:

1. Parse English using MST parser (McDonald et al., 2005),
2. Move finite verbs to the end of the clause,
3. Transform prepositions to postpositions.

Hinglish→Hindi translation using Moses:

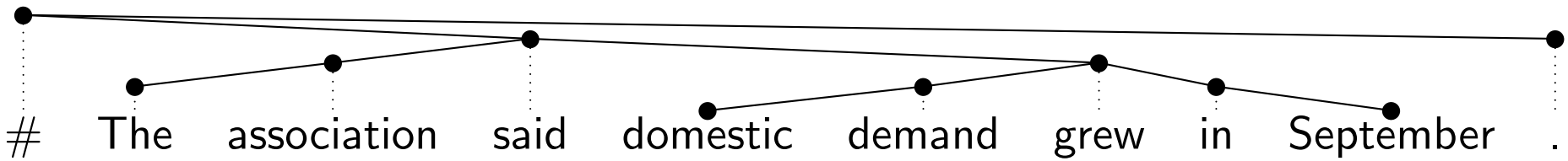
- Baselines: Distance-based or lexicalized reordering,
- Improved: (Rule-base Reord. and) Suffix LM with + Optional

	EILMT	TIDES
Baseline Moses, Distance Reordering	18.88±2.05	10.06±0.76
Baseline Moses, Reordering Using en+hi Forms	19.77±2.03	**10.95±0.75**
Suffix LM+Reord	20.09±2.18	10.18±0.74
Rule-based Reordering + Suffix LM+Reord	**21.01±2.18**	10.29±0.69

Join TectoMT tutorial lab session for SVO→SOV in 12 lines of Perl.

NLP Hacking: Valency Information

Bring non-local information closer based on dependency edges:



To produce "verbose tokens":

the|said assoc.|said said|- domestic|grew demand|grew grew|said in|grew September|in

Remember to back-off with regular tokens:

the assoc. said domestic demand grew in September

Details and further explanation: "Alternative decoding paths" in Friday lecture.

- Should help lexical choice under verbs (verb revealed).
- Should help case choice under prepositions.

en→cs preliminary BLEU scores	80k	2.2M sents.
Baseline	9.77±0.69	**14.57±0.83**
With source valency	**9.98±0.67**	14.52±0.85

Fishing: Phrase Table Marking

- Hard constraints always hurt. Also e.g. Ambati and Lavie (2008).
- Instead of dropping phrase/treelet table entries, *mark* them with an additional score/feature.
- MERT (see Friday class) will decide how much should the marked entries be penalized.

```
in europa ||| in europe ||| 0.829007 0.207955 0.801493 0.492402 2.718 1
europas ||| in europe ||| 0.0251019 0.066211 0.0342506 0.0079563 2.718 1
in europa , ||| in europe ||| 0.011371 0.207955 0.207843 0.492402 2.718 0
```

E.g. mark phrases in phrase table:

- confirmed by a printed/on-line dictionary,
- consistent with surface syntax,
- consistent with deep syntax and t-alignment

Currently me and Václav Novák, happy to join others.

Fishing: Source-Context Features

Some scores phrase translations could be computed on-line:

1. Create translation options for a span as usual.
2. Feed them to an external scorer.
3. Obtain an additional score for each translation option.

Such "dynamic scores" can condition on source sentence context:

- syntactic structure,
- detailed attributes (e.g. case), *without causing data sparseness*.

Consider "John loves Mary":

- Translation options for Mary: $\text{Marie}_{\text{nom}}$ $\text{Marii}_{\text{acc,dat}}$, . . .
- Given "Mary" is object, "$\text{Marii}_{\text{acc,dat}}$" should be promoted.
- Better than relying on the presence of 2-word phrase "loves Mary" in the phrase table.

Me and Kamil Kos are looking for collaborators.

The "backdoor" from Moses to arbitrary external scorer implemented, we need to train the scorer.
Inspired by Carpuat and Wu (2007) and Trevor Cohn (pers.comm.).

Summary

- Syntax as a hard constraint is bad.
  - More so, if your tagger+parser+. . . are not perfect.
- Rich annotation is dangerous when not treated carefully.
  Occam's razor: think twice before adding an attribute.
  - Avoid data sparseness, always provide a back-off.
  - Avoid complex models, they are hard to tune (set parameters).

TectoMT is great for rich annotation and NLP hacking.

Feature fishing for Moses proposed:

- Marking phrases compatible/confirmed by an additional source.
- Dynamic source-context features.

References

Vamshi Ambati and Alon Lavie. 2008. Improving Syntax-Driven Translation Models by Re-structuring Divergent and Nonisomorphic Parse Tree Structures. In *Proc. of AMTA*, pages 235–244.

Ondřej Bojar and Martin Čmejrek. 2007. Mathematical Model of Tree Transformations. Project Euromatrix - Deliverable 3.2, ÚFAL, Charles University.

Ondřej Bojar, Pavel Straňák, and Daniel Zeman. 2008. English-hindi translation in 21 days. In *Proceedings of the 6th International Conference On Natural Language Processing (ICON-2008) NLP Tools Contest*, Pune, India. NLP Association of India.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic.

Martin Čmejrek. 2006. *Using Dependency Tree Structure for Czech-English Machine Translation*. Ph.D. thesis, ÚFAL, MFF UK, Prague, Czech Republic.

Michael Collins. 1996. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191.

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.

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.