What a Transfer-Based System Brings to the Combination with PBMT

Aleš Tamchyna and Ondřej Bojar

surname@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics

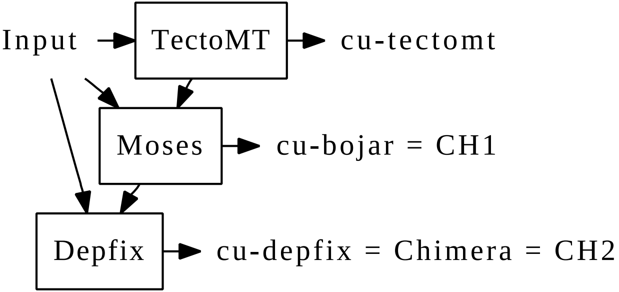
Charles University in Prague

July 31, 2015

Outline

- Chimera: Our WMT system.
- Targetting Czech with phrase-based MT.
- TectoMT: Deep syntactic transfer.
- Poor man's Combination.
- What TectoMT brings to the combination:
  - Phrases otherwise unreachable.
  - Linguistic phenomena improved.
  - Easier search.
- Summary.

Our WMT System: Chimera



Input

TectoMT

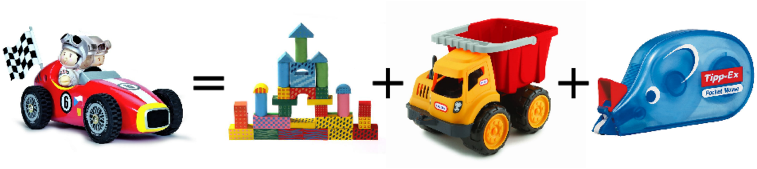
cu-tectomt

Moses

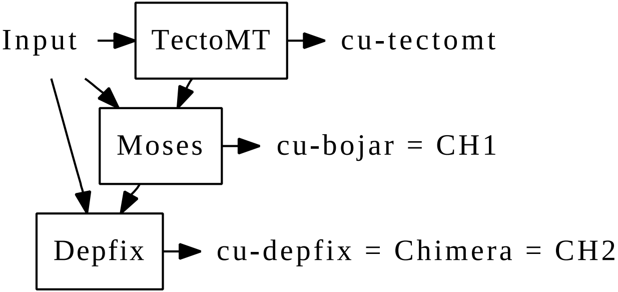
cu-bojar = CH1

Depfix

cu-depfix = Chimera = CH2



Our WMT System: Chimera



Input

TectoMT

cu-tectomt

Moses

cu-bojar = CH1

Depfix

cu-depfix = Chimera = CH2

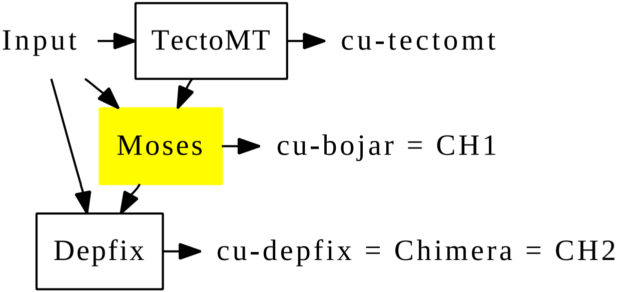
Chimera is a hybrid system of three components:

- TectoMT: Deep-syntactic transfer-based system.
- Moses: Factored phrase-based system.
- Depfix: Automatic post-correction (grammar, negation).

Performance in WMT

	System	BLEU	TER	Manual
WMT13	CH2: 🧱+🚚+🩹	20.0	**0.693**	**0.664**
	CH1: 🧱+🚚	**20.1**	0.696	0.637
	CH0: 🚚	19.5	0.713	–
	GOOGLE TR.	18.9	0.720	0.618
	🧱 CU-TECTOMT	14.7	0.741	0.455
WMT14	CH2: 🧱+🚚+🩹	21.1	0.670	**0.373**
	UEDIN-UNCONSTR.	**21.6**	**0.667**	0.357
	CH1: 🧱+🚚	20.9	0.674	0.333
	GOOGLE TR.	20.2	0.687	0.168
	🧱 CU-TECTOMT	15.2	0.716	-0.177
WMT15	CH2: 🧱+🚚+🩹	**18.8**	**0.715**	**0.686**
	CH1: 🧱+🚚	18.7	0.717	–
	CH0: 🚚	17.6	0.730	–
	GOOGLE TR.	16.4	0.750	0.515
	🧱 CU-TECTOMT	13.4	0.763	0.209

Chimera Overview



Targetting Czech with PBMT

First phase of phrase-based MT:

- Construction of translation options.

I	saw	two	green	striped	cats	.
I	saw	two	green	striped	cats	.
I have seen		a pair of	lime	strappy	kittens	!
we	had seen	2	fresh	banded	...	
...	...	...	gaily	stripy		
			free	striper		
			...	tigers		
				...		

Targetting Czech with PBMT

To reduce noise and size of search space:

- Limit number of options per span to e.g. 20.

I	saw	two	green	striped	cats	.
I	saw	two	green	striped	cats	.
I have seen		a pair of	lime	strappy	kittens	!
we	had seen	2	~~*fresh*~~	banded	...	
...	...	...	~~*gaily*~~	~~*stripy*~~		
			~~*free*~~	~~*striper*~~		
			...	~~*tigers*~~		
				...		

Targetting Czech with PBMT

Czech is fusional: suffix encodes many categories:

- Nouns and Adjs: 7 cases, 4 genders, 3 nums, …

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			…	…		

Targetting Czech with PBMT

Grammatical agreement:

- Elements of NPs must agree in case, num and gend.

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			…	…		

Targetting Czech with PBMT

A different verb may select for a different case.

- … different choice of forms needed.

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**		
zrak mi utkvěl na			zelenému	pruhovanému		
			zeleným	pruhovaným		
	…		zelenou	pruhovanou		
viděl jsem			zelenými	pruhovanými		
viděla jsem			…	…		

Our Moses Setup

- Phrase-based (not hierarchical, not OSM).
- Tuned with MERT (not MIRA, …).
- Tuned towards BLEU (sadly best anyway).
- Factored, <u>in the simplest form</u>:

$$\text{word form} \rightarrow \left\{ \begin{array}{c} \text{word form} \\ \text{morphological tag} \end{array} \right\}$$

Our Moses Setup

- Phrase-based (not hierarchical, not OSM).
- Tuned with MERT (not MIRA, …).
- Tuned towards BLEU (sadly best anyway).
- Factored, in the simplest form:

$$\text{word form} \rightarrow \left\{ \begin{array}{c} \text{word form} \\ \text{morphological tag} \end{array} \right\}$$

green	striped
zelený	pruhovaný
zelené	**pruhované**
zelení	pruhovaní
zelených	**pruhovaných**
zeleným	pruhovaným

Our Moses Setup

- Phrase-based (not hierarchical, not OSM).
- Tuned with MERT (not MIRA, ...).
- Tuned towards BLEU (sadly best anyway).
- Factored, in the simplest form:

$$\text{word form} \rightarrow \left\{ \begin{array}{c} \text{word form} \\ \text{morphological tag} \end{array} \right\}$$

green	striped
zelený$_{\text{sg,masc,nom}}$	pruhovaný$_{\text{sg,masc,nom}}$
zelené$_{\text{sg,fem,gen}}$	**pruhované**$_{\text{sg,fem,gen}}$
zelené$_{\text{sg,fem,dat}}$	**pruhované**$_{\text{sg,fem,dat}}$
zelené$_{\text{pl,fem,nom}}$	**pruhované**$_{\text{pl,fem,nom}}$
zelení$_{\text{pl,masc,nom}}$	pruhovaní$_{\text{pl,masc,nom}}$
zelených$_{\text{pl,masc,loc}}$	**pruhovaných**$_{\text{pl,masc,loc}}$
zeleným	pruhovaným

Our Moses Setup

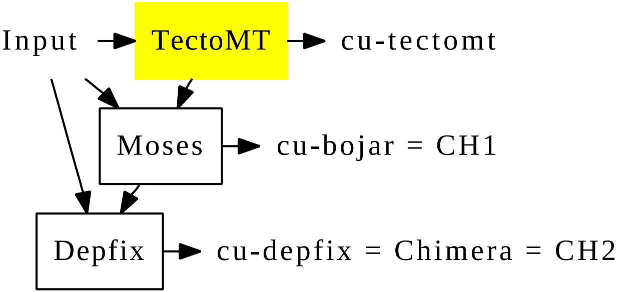
- Phrase-based (not hierarchical, not OSM).
- Tuned with MERT (not MIRA, ...).
- Tuned towards BLEU (sadly best anyway).
- Factored, in the simplest form:

$$\text{word form} \rightarrow \left\{ \begin{array}{c} \text{word form} \\ \text{morphological tag} \end{array} \right\}$$

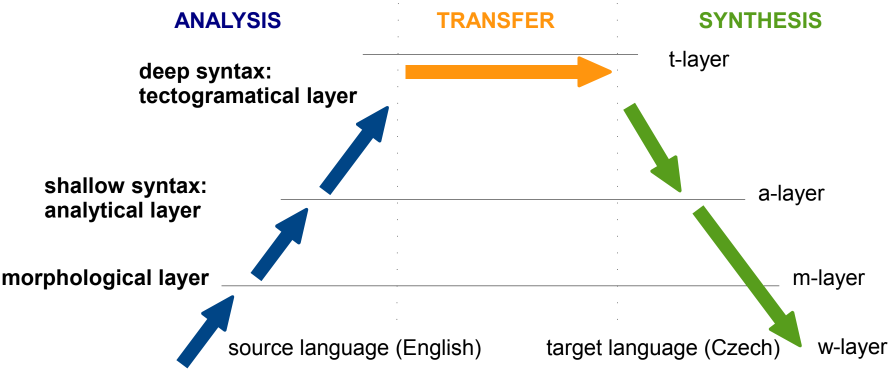
- Large Data, multiple language models.

LM ID	**factor**	**order**	**# tokens**
long	stc	7	685M
big	stc	4	3903M
morph	tag	10	817M
longm	tag	15	817M

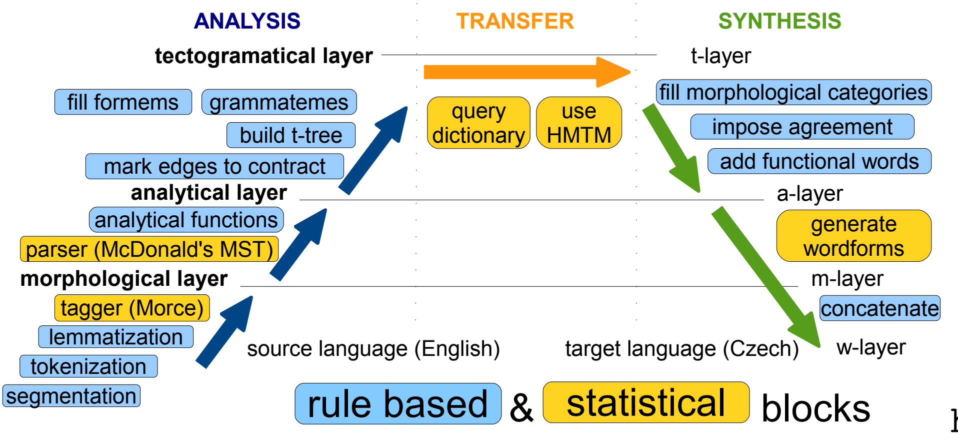
Chimera Overview



TectoMT: Deep MT



TectoMT: Deep MT



ANALYSIS

TRANSFER

SYNTHESIS

tectogramatical layer

t-layer

fill formems

grammatemes

build t-tree

mark edges to contract

analytical layer

analytical functions

parser (McDonald's MST)

morphological layer

tagger (Morce)

lemmatization

tokenization

segmentation

query dictionary

use HMTM

fill morphological categories

impose agreement

add functional words

a-layer

generate wordforms

m-layer

concatenate

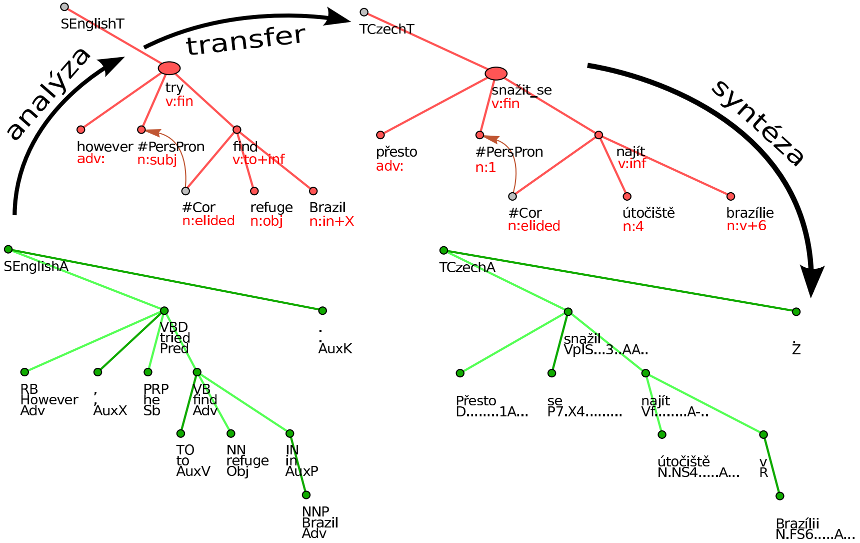
w-layer

source language (English)

target language (Czech)

rule based & statistical blocks

In a Nutshell: Tree-to-Tree Translation



- T-layer abstraction ⇒ tree structure kept unchanged.

TectoMT Key Features

A Typical Deep Syntactic System

- Only content words have nodes.
- Words represented as lemmas.

Edge Labels: Formemes

- Compact string (atomic) for syntactic and morphological properties and neighbourhood.

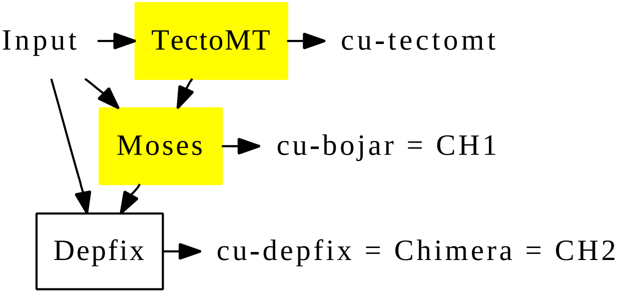
Hidden Markov Tree Model

- For globally best choice of t-lemmas and formemes.
- Source and target trees assumed **isomorphic**.

Maximum-Entropy Translation Model

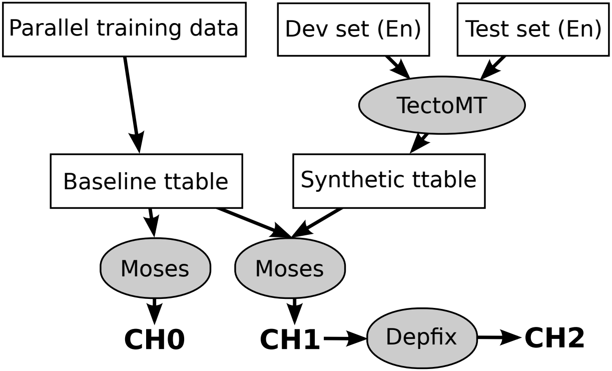
- One classifier per source lemma.
- Features: lemmas and formemes of immediate neighbours (both tree and linear).

Chimera Overview



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 Phrases as Translation Options

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 Phrases as Translation Options

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 Phrases as Translation Options

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				

What are TectoMT Phrases Like (1/3)

- On average, they seem worse than corpus phrases.

Manual annotation of 100 phrases (2 annotators):

- Can you imagine a context where the phrase would be a good translation (OK)?

		OK	**Bad**	**Unsure**	**IAA**
ttable	from corpus	76.0%	17.5%	6.5%	78.0
	by TectoMT	66.3%	26.3%	7.4%	83.0
used	from corpus	89.0%	7.5%	3.5%	94.0
	by TectoMT	87.5%	9.0%	3.5%	87.0

- 9–26% of phrases by TectoMT introduce an error.
- 8–18% of phrases from corpus introduce an error.

What are TectoMT Phrases Like (1/3)

- On average, they seem worse than corpus phrases.

Manual annotation of 100 phrases (2 annotators):

- Can you imagine a context where the phrase would be a good translation (OK)?

		OK	**Bad**	**Unsure**	**IAA**
ttable	from corpus	76.0%	17.5%	6.5%	78.0
	by TectoMT	66.3%	26.3%	7.4%	83.0
used	from corpus	89.0%	7.5%	3.5%	94.0
	by TectoMT	87.5%	9.0%	3.5%	87.0

Note the high agreement

- 9–26% of phrases by TectoMT introdu
- 8–18% of phrases from corpus introduc

What are TectoMT Phrases Like (2/3)

- Longer ones are used, compared to corpus phrases.

		by TectoMT	**from corpus**	**both**	**total**
phrase	count	3606	10033	18322	31961
tokens	avg. len.	**3.68**	2.47	1.56	2.08
phrase	count	3503	9400	8203	21106
types	avg. len.	**3.73**	2.52	2.07	2.54

- Used phrases by TectoMT are 1.2 word longer that used phrases from corpus.

⇒ Search simplified.

⇒ MERT more stable (StdDev of 0.07 compared to 0.15).

What are TectoMT Phrases Like (3/3)

- ∼10% of TectoMT phrases <u>cannot be reached</u> using corpus phrases.
- Corresponds to 32% of sentences:

Constraint decoding: attempt of CH0 to reach translations by CH1:

all	**different?**	**reachable?**	**score diff (CH1 · CH0)**	
3003	2665	1741	1601 ($<$)	modelling errors
			140 ($>$)	search errors
		924	(unreachable)	
	338	(identical)		

What are TectoMT Phrases Like (3/3)

- ∼10% of TectoMT phrases <u>cannot be reached</u> using corpus phrases.
- Corresponds to 32% of sentences:

Constraint decoding: attempt of CH0 to reach translations by CH1:

all	different?	reachable?	score diff (CH1 · CH0)	
3003	2665	1741	1601 (<)	modelling errors
			140 (>)	search errors
		924	(unreachable)	
	338	(identical)		

Modelling errors:	CH1		CH0
BLEU on these 1601 sentences	24.78	>	23.03

Towards the Reference

TectoMT	CH0	CH1	Tokens	Types			
			1gr	1gr	2gr	3gr	4gr
✓	✓	✓	44.7%	41.6%	15.1%	6.5%	3.0%
-	-	-	32.9%	35.0%	63.0%	77.5%	85.8%
-	✓	✓	8.6%	8.8%	9.3%	7.2%	5.1%
✓	-	✓	4.5%	4.8%	3.8%	2.5%	1.5%
-	✓	-	3.6%	3.8%	3.5%	2.5%	1.8%
✓	-	-	3.5%	3.7%	2.9%	1.9%	1.2%
-	-	✓	1.4%	1.4%	1.9%	1.8%	1.5%
✓	✓	-	0.8%	0.8%	0.4%	0.2%	0.1%
Total (100 %)			60.6k	56.3k	57.3k	54.5k	51.6k

Towards the Reference

TectoMT	CH0	CH1					gr
✓	✓	✓					%
-	-	-	32.9%	35.0%	63.0%	77.5%	85.8%
-	✓	✓	8.6%	8.8%	9.3%	7.2%	5.1%
✓	-	✓	4.5%	4.8%	3.8%	2.5%	1.5%
-	✓	-	3.6%	3.8%	3.5%	2.5%	1.8%
✓	-	-	3.5%	3.7%	2.9%	1.9%	1.2%
-	-	✓	1.4%	1.4%	1.9%	1.8%	1.5%
✓	✓	-	0.8%	0.8%	0.4%	0.2%	0.1%
Total (100 %)			60.6k	56.3k	57.3k	54.5k	51.6k

1/3 of reference usually not reached in morphologically rich languages

Towards the Reference

TectoMT	CH0	CH1	Tokens	Types			
			1gr	1gr	2gr	3gr	4gr
✓	✓	✓	44.7%	41.6%	15.1%	6.5%	3.0%
-	-	-	[illegible]	[illegible]	[illegible]	[illegible]	35.8%
-	✓	✓	[illegible]	[illegible]	[illegible]	[illegible]	5.1%
✓	-	✓	4.5%	4.8%	3.8%	2.5%	1.5%
-	✓	-	3.6%	3.8%	3.5%	2.5%	1.8%
✓	-	-	3.5%	3.7%	2.9%	1.9%	1.2%
-	-	✓	1.4%	1.4%	1.9%	1.8%	1.5%
✓	✓	-	0.8%	0.8%	0.4%	0.2%	0.1%
Total (100 %)			60.6k	56.3k	57.3k	54.5k	51.6k

Words we produced thanks to TectoMT

Towards the Reference

TectoMT	CH0	CH1	Tokens	Types			
			1gr	1gr	2gr	3gr	4gr
✓	✓	✓	44.7%	41.6%	15.1%	6.5%	3.0%
-	-	-	32.9%	35.0%	63.0%	77.5%	85.8%
-	✓	✓	8.6%	8.8%	9.3%	7.2%	5.1%
✓	-	✓	4.5%	4.8%	3.8%	2.5%	1.5%
-	✓	-	3.6%	3.8%	3.5%	2.5%	1.8%
✓	-	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1.2%
-	-	✓	1.4%	1.4%	1.9%	1.8%	1.5%
✓	✓	-	0.8%	0.8%	0.4%	0.2%	0.1%
Total (100 %)			60.6k	56.3k	57.3k	54.5k	51.6k

Positive side-effect

Towards the Reference

TectoMT	CH0	CH1	Tokens	Types			
			1gr	1gr	2gr	3gr	4gr
✓	✓	✓	44.7%				
-	-	-	32.9%				
-	✓	✓	8.6%				
✓	-	✓	4.5%	4.8%	3.8%	2.5%	1.5%
-	✓	-	3.6%	3.8%	3.5%	2.5%	1.8%
✓	-	-	3.5%	3.7%	2.9%	1.9%	1.2%
-	-	✓	1.4%	1.4%	1.9%	1.8%	1.5%
✓	✓	-	0.8%	0.8%	0.4%	0.2%	0.1%
Total (100 %)			60.6k	56.3k	57.3k	54.5k	51.6k

Manually checked for linguistic phenomena

4-grams Won Thanks to TectoMT

OK Anyway	42 (31.1 %)
Worsened	4 (3.0 %)
Bad Anyway	2 (1.5 %)
Word Order esp. Syntax of Complex NPs	13 (9.6 %)
Valency of Verbs and Nouns	12 (8.9 %)
Agreements in NPs or Subj-Verb	10 (7.4 %)
Clause Structure (Conjunctions etc.)	8 (5.9 %)
Lexical Choice	7 (5.2 %)
Avoided Superfluous Comma	5 (3.7 %)
Possessive ('s or of)	5 (3.7 %)
Properties of Verbs (number, tense, …)	4 (3.0 %)
Reflexive Particle	3 (2.2 %)
Other	20 (14.8%)
Total	135 4-grams

4-grams Won Thanks to TectoMT

OK Anyway	42 (31.1 %)
Worsened	4 (3.0 %)
Bad Anyway	[illegible]
Word Order esp. Syntax of Complex NPs	13 (9.6 %)
Valency of Verbs and Nouns	12 (8.9 %)
Agreements in NPs or Subj-Verb	10 (7.4 %)
Clause Structure (Conjunctions etc.)	8 (5.9 %)
Lexical Choice	7 (5.2 %)
Avoided Superfluous Comma	5 (3.7 %)
Possessive ('s or of)	5 (3.7 %)
Properties of Verbs (number, tense, ...)	4 (3.0 %)
Reflexive Particle	3 (2.2 %)
Other	20 (14.8%)
Total	135 4-grams

No real win

4-grams Won Thanks to TectoMT

OK Anyway	42 (31.1 %)
Worsened	4 (3.0 %)
Bad Anyway	2 (1.5 %)
Word Order esp. Syntax of Complex	
Valency of Verbs and Nouns	12 (8.9 %)
Agreements in NPs or Subj-Verb	10 (7.4 %)
Clause Structure (Conjunctions etc.)	8 (5.9 %)
Lexical Choice	7 (5.2 %)
Avoided Superfluous Comma	5 (3.7 %)
Possessive ('s or of)	5 (3.7 %)
Properties of Verbs (number, tense, ...)	4 (3.0 %)
Reflexive Particle	3 (2.2 %)
Other	20 (14.8%)
Total	135 4-grams

Small loss

4-grams Won Thanks to TectoMT

OK Anyway	42 (31.1 %)
Worsened	4 (3.0 %)
Bad Anyway	2 (1.5 %)
Word Order esp. Syntax of Complex NPs	13 (9.6 %)
Valency of Verbs and Nouns	12 (8.9 %)
Agreements in NPs or Subj-Verb	10 (7.4 %)
Clause Structur c.)	8 (5.9 %)
Lexical Choice	7 (5.2 %)
Avoided Superf	5 (3.7 %)
Possessive ('s or of)	5 (3.7 %)
Properties of Verbs (number, tense, ...)	4 (3.0 %)
Reflexive Particle	3 (2.2 %)
Other	20 (14.8%)
Total	135 4-grams

Wide range of small improvements

4-grams Won Thanks to TectoMT

OK Anyway	42 (31.1 %)
Worsened	4 (3.0 %)
Bad Anyway	2 (1.5 %)
Word Order esp. Syntax of Complex NPs	13 (9.6 %)
Valency of Verbs and Nouns	12 (8.9 %)
Agreements in NPs or Subj-Verb	10 (7.4 %)
Clause Structure (Conjunctions etc.)	8 (5.9 %)
Lexical Choice	7 (5.2 %)
Avoided Superfluous Comma	5 (3.7 %)
Possessive ('s or of)	5 (3.7 %)
Properties of Verbs (number, tense, ...)	4 (3.0 %)
Reflexive Particle	3 (2.2 %)
Other	20 (14.8%)
Total	135 4-grams

TectoMT Complementary to LMs

LMs	**-TectoMT**	**+TectoMT**	Δ
long	21.32	22.93	+1.61
big	22.00	23.19	+1.19
long morph	22.01	23.48	+1.47
big long	22.26	23.84	+1.58
big morph	22.21	23.89	+1.68
big long morph	22.48	24.10	+1.62
all + longm	22.59	**24.24**	+1.65

- TectoMT in 2015 brought ∼1.5 BLEU across various subsets of LMs.

Summary

The state of the art is hybrid:

- PBMT to fully benefit from huge data.
- Transfer-based MT for a wide range of things.
    - Complex NPs, valency, agreement, clause structure.
    - Some of these suggestions would not be reachable otherwise.

Adding tailored phrases to PBMT helps:

- Phrases are longer ⇒ search simplified.
- Some words won by side-effects.
- Lower variance of MERT.