Generating Czech Word Forms in MT:

From System Combination to Black Art

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

Outline

- Targeting Czech.
  - Vocabulary sizes.
  - Source of the morphological explosion.
  - OOV rates.
- Failed: Factored attempts to generate forms on the fly.
- Promising: Two-Step Translation.
- Universal: System Combination:
  - Improving alignments, adding weights.
  - Larger LMs, Tag LMs.
- Black Art: Reverse Self-Training.
- Summary.

Vocabulary Sizes for en and cs

WMT10 (Bojar and Kos, 2010)	Large	Small	Dev
Sentences	7.5M	126.1k	2.5k
Czech Tokens	79.2M	2.6M	55.8k
English Tokens	89.1M	2.9M	49.9k
Czech Vocabulary	923.1k	**138.7k**	15.4k
English Vocabulary	646.3k	**64.7k**	9.4k
Czech Lemmas	553.5k	60.3k	9.5k
English Lemmas	611.4k	53.8k	7.7k

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

Morphological Explosion in Czech

(In)flective lang.: many categories expressed in a single suffix:

- Czech nouns and adjectives: 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		. . .	. . .		

Out-of-Vocabulary Rates

Dataset (# Sents)	Language	n-grams Out of: Corpus Voc. 1	Corpus Voc. 2	Phrase-Table Voc. 1	Phrase-Table Voc. 2
7.5M	Czech	2.2%	30.5%	3.9%	44.1%
	English	1.5%	13.7%	2.1%	22.4%
	Czech + English input sent	1.5%	29.4%	3.1%	42.8%
126k	Czech	6.7%	48.1%	12.5%	65.4%
	English	3.6%	28.1%	6.3%	45.4%
	Czech + English input sent	5.2%	46.6%	10.6%	63.7%
126k	Czech lemmas	4.1%	36.3%	5.8%	52.6%
	English lemmas	3.4%	24.6%	6.9%	53.2%
	Czech + English input lemmas	3.1%	35.7%	5.1%	38.1%

- Significant vocabulary loss during phrase extraction:
  - e.g. 2.2%→3.9% for 7.5M Czech.
- OOV of Czech forms ∼twice as bad as in English, cf. the reds.
- OOV of Czech lemmas lower than in English, see the greens.

Overview of MT Systems Discussed

- Phrase-Based:
  - Vanilla Moses.
  - Factored for Morphological Generation on the Fly.
  - Two-Step Translation.
- TectoMT.
- ROVER System Combination.
- Phrase-Based:
  - Reverse Self-Training.

Factored Translation Scenarios

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)

Data	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	–

T+C = form→form (i.e. vanilla), generate tag, use extra tag LM
T+C+C = form→form, generate lemma and tag, use extra lemma LM and tag LM
T+T+G = lemma→lemma, tag→tag, generate form

- T+T+G explodes the search space
  - too many translation options ⇒ stacks overflown
    ⇒ important options pruned before LM context can pick them

Two-Step Attempts (WMT10) 1/2

1. English → lemmatized Czech
   - meaning-bearing morphology preserved
   - max phrase len 10, distortion limit 6
   - large target-side (lemmatized LM)
2. Lemmatized Czech → Czech
   - max phrase len 1, monotone

Src	after a sharp drop		
Mid	po+6	ASA1.prudký	NSA-.pokles
Gloss	after+voc	adj+sg...sharp	noun+sg...drop
Out	po	prudkém	poklesu

- Only 1-best output passed, will try lattice.

Two-Step Attempts (WMT10) 2/2

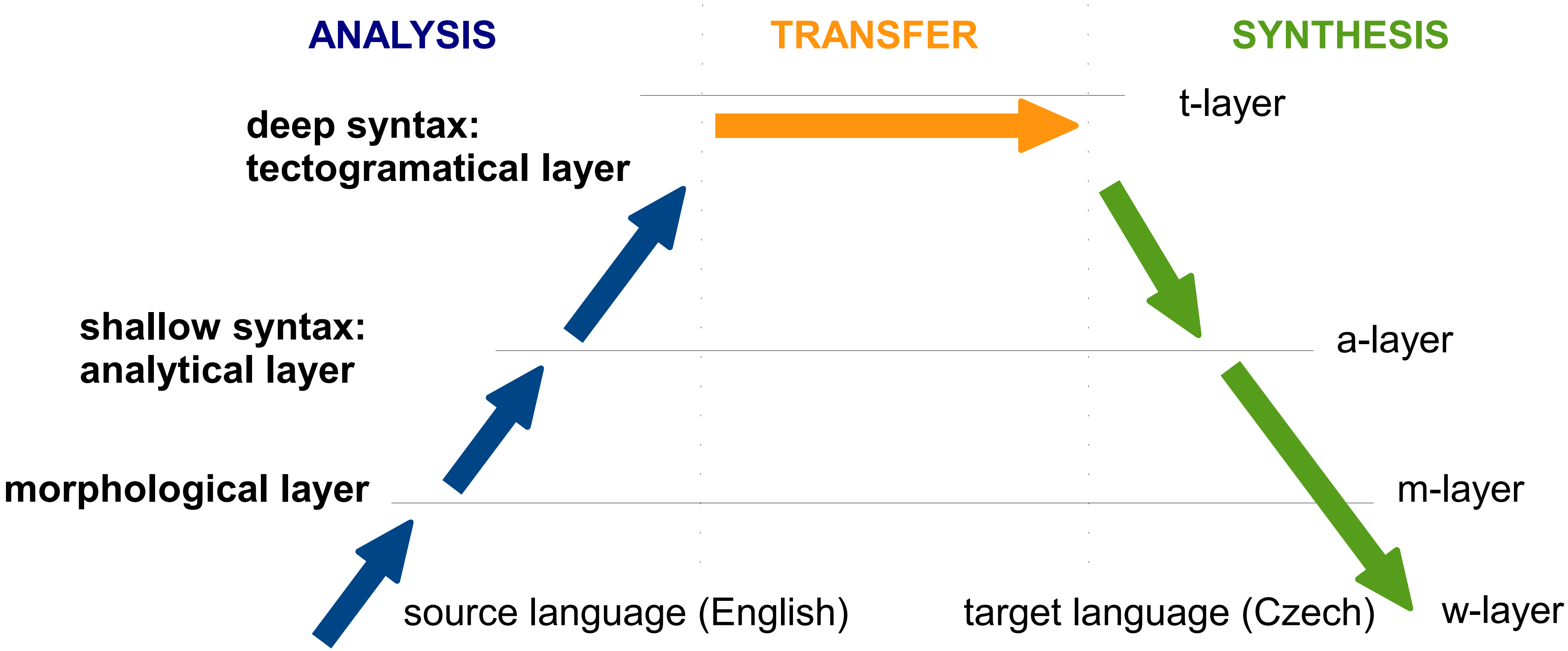
Data Size		Simple		Two-Step		Diff
Parallel	Mono	BLEU	SemPOS	BLEU	SemPOS	B.S.
126k	126k	10.28±0.40	29.92	10.38±0.38	30.01	$\nearrow\nearrow$
126k	13M	12.50±0.44	31.01	12.29±0.47	31.40	$\searrow\nearrow$
7.5M	13M	14.17±0.51	33.07	14.06±0.49	32.57	$\searrow\searrow$

Manual micro-evaluation of $\searrow\nearrow$, i.e. 12.50±0.44 vs. 12.29±0.47:

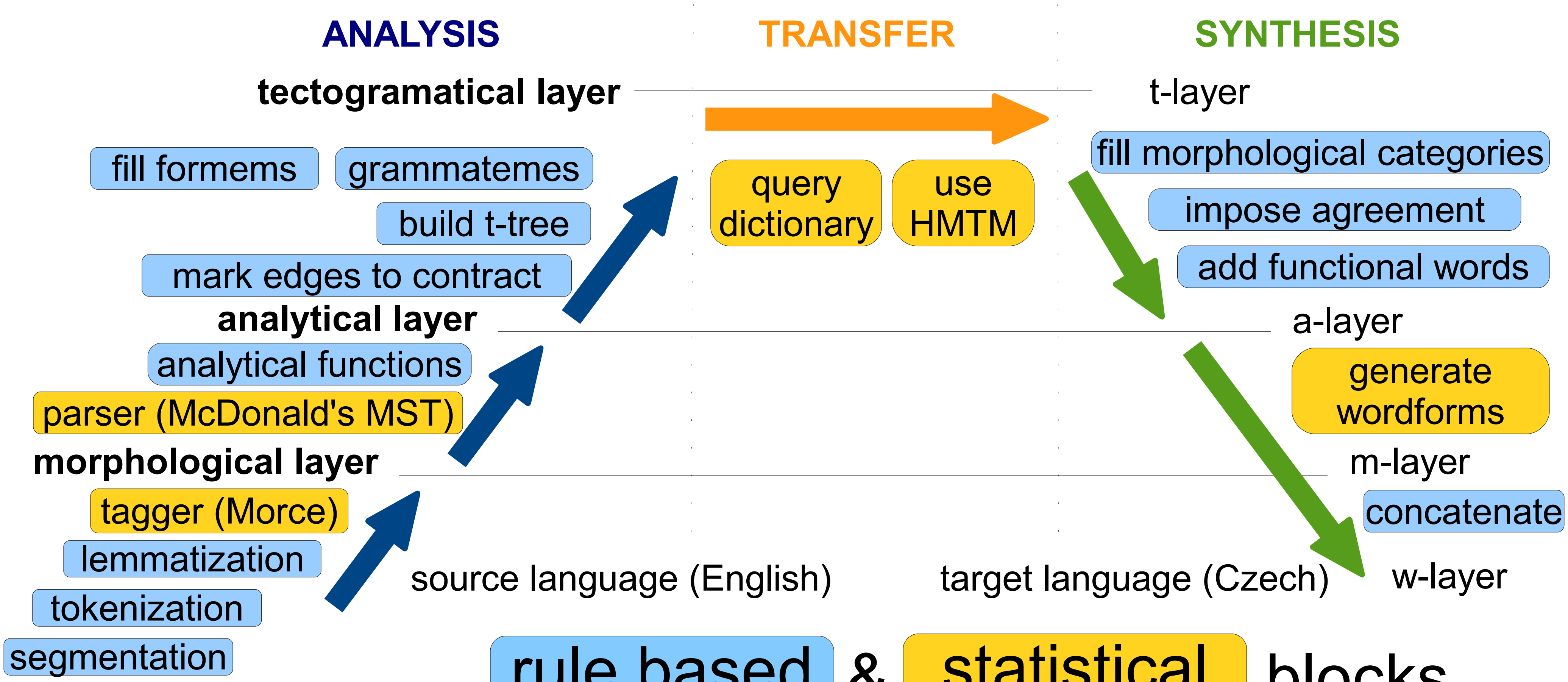
	Two-Step	Both Fine	Both Wrong	Simple	Total
Two-Step	**23**	4	8	-	**35**
Both Fine	7	14	17	5	43
Both Wrong	8	1	28	2	39
Simple	-	3	7	**23**	33
Total	**38**	22	60	30	150

- Each annotator weakly prefers Two-step
  – but they don't agree on individual sentences.

“TectoMT Transfer” (1/2)



“TectoMT Transfer” (2/2)



rule based & statistical blocks

TectoMT vs. Others for en→cs

Metric	Google	**CU-Bojar**	PC Translator	**TectoMT**
$\geq$ others (official)	**70.4**	65.6	62.1	*60.1*
$>$ others	49.1	45.0	**49.4**	*44.1*
Edits acceptable [%]	**55**	40	43	*34*
Quiz-based evaluation [%]	80.3	*75.9*	80.0	**81.5**
BLEU	**0.16**	0.15	*0.10*	0.12
NIST	**5.46**	5.30	*4.44*	5.10

- TectoMT worst (of these 4 sys.) in sentence ranking and editing.
- TectoMT best in quiz-based evaluation (Berka et al., 2011):
  - % of correctly answered Y/N questions given short machine-translated texts.
- TectoMT provides many words needed by the reference. See below.

Even “Bad” Systems Offer Words

Analyzing 44193 toks in the ref of WMT10 syscomb Test set.

- What is the % tokens produced by bojar-primary?
- What is the % tokens produced by one of the secondary systems only?

	bojar-primary (16.90±0.61) vs.			
	bojar-sempos	bojar-2stepsl	tectomt	the 3 other
	16.61±0.59	14.38±0.58	13.19±0.58	-
In Both	48.3	43.8	41.2	50.8
Nowhere	45.4	42.8	41.0	37.0
Primary Only	3.5	8.0	10.6	1.0
Secondary Only	**2.8**	**5.4**	**7.1**	**11.2**

- TectoMT could bring in up to 7.1% tokens, Two-Step 5.4%. . .
- Still 37% tokens of the reference not available.
- Decreasing BLEU: systems less similar to primary score worse.

Rover System Combination (1/2)

Main idea of Fiscus (1997), extended by Matusov et al. (2008):
Systems vote which individual words should appear in the output.

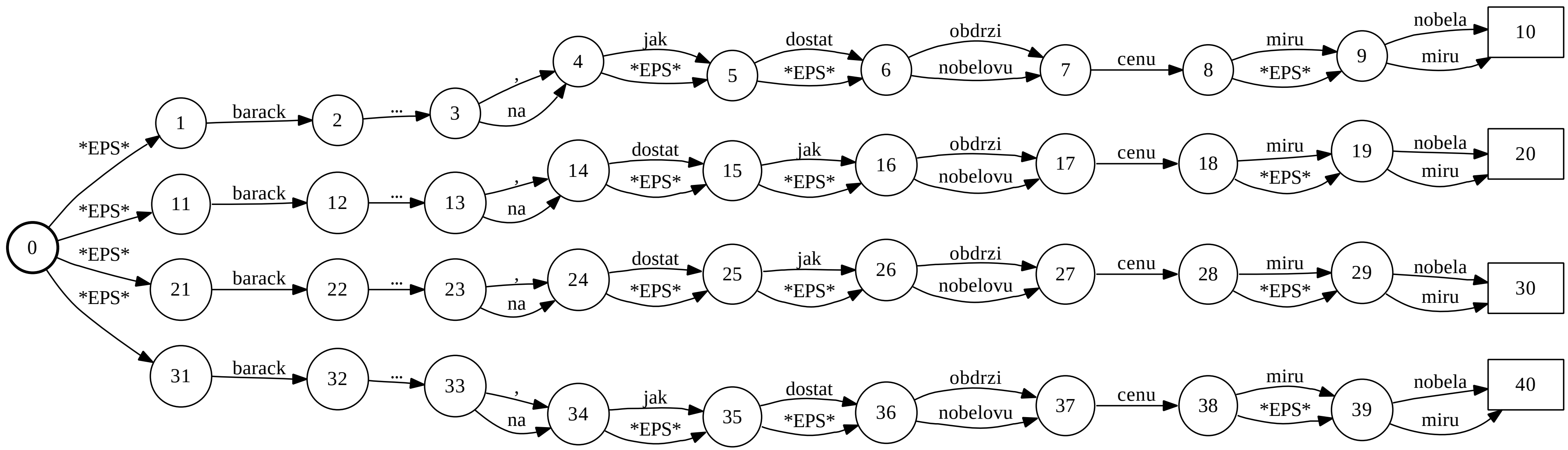
Procedure:

1. Given a "primary system" / "skeleton";
   – Align each system to one skeleton (bold), producing "bitexts":
     barack|**barack** . . . ,|**na** dostat|ϵ jak|ϵ nobelovu|**nobelovu** cenu|**cenu** míru|**míru**
     barack|**barack** . . . na|**na** nobelovu|**nobelovu** cenu|**cenu** míru|**míru**
     barack|**barack** . . . ,|**na** obdrží|**nobelovu** cenu|**cenu** míru|ϵ nobela|**míru**
   – Combine all bitexts to confusion network:

barack	...	na	ϵ	ϵ	nobelovu	cenu	ϵ	míru
barack	...	,	dostat	jak	nobelovu	cenu	ϵ	míru
barack	...	na	ϵ	ϵ	nobelovu	cenu	ϵ	míru
barack	...	,	ϵ	ϵ	obdrží	cenu	míru	nobela

Rover System Combination (2/2)

2. Combine confusion networks of various skeletons to one lattice:



3. Add language model scores.

4. Optimize weights (word penalty, LM, skeleton choice, . . .).

5. Select best path.

Combined Systems

In the following, we:

- Combine only ÚFAL's systems built for the WMT10 shared task.
- Tune and evaluate on WMT10 combination task datasets.

	Dev Set		Test Set	WMT10 Manual Rank
bojar-primary	16.00±1.15	$\nearrow$	16.90±0.61	65.5
bojar-sempos	15.76±1.12	$\nearrow$	16.61±0.59	-
bojar-2step	13.59±1.12	$\nearrow$	14.38±0.58	-
tectomt	11.48±1.04	$\nearrow$	13.19±0.58	60.1
google	17.32±1.25	$\searrow$	16.76±0.60	70.4
eurotran	9.64±0.92	$\nearrow$	11.04±0.48	54.0
pctrans2010	10.24±0.92	$\nearrow$	10.84±0.46	62.1

Note Google discrepancy between Dev and Test $\Rightarrow$ overfitting would be very likely.

Manual System Combination

To check the plausibility of “voting assumption” we manually do the task:

- Myself:
  - English→Czech, WMT10, 4 systems, 52 sents.
  - Reference translation available.
  - Attempted to stick to the original word order.
- Matusov (2009) (p. 140 talks about TC-STAR07 es→en):
  - Chinese(?)→English, IWSLT 2006, 4 systems, 489 sents.
  - Without looking at source or reference.
  - Allowed any reordering.
  - No further analysis beyond BLEU/TER/WER/PER.

Plausibility of Voting Assumption

How many produced tokens actually had the majority support?

	Matusov (2009)		My en→cs WMT10			
	Manual		Manual		Auto	
Supported by	Toks	%	Toks	%	Toks	%
1	978	15.8	160	19.4	30	3.6
2	1117	18.1	110	13.3	183	21.9
≤ 2	2095	**33.9**	270	**32.7**	213	**25.5**
3	1279	20.7	137	16.6	188	22.5
4	2806	45.4	417	50.6	435	52.0
Total	6180	100.0	824	100.0	836	100.0

... about $\frac{1}{3}$ of manually and $\frac{1}{4}$ of automatically combined tokens have no majority support (weights influence this).

Main Examined Directions

No Rover, just Moses, simply "add to training":

- Add the 3 other outputs to training data of bojar-primary.

Within RWTH Rover implementation (minor modifications):

- Improving word alignments.

RWTH alignment + Moses path selection and MERT:

- More detailed lattice arc weights.
- Handling of indicators in log-linear framework.
- Larger LMs.
- LMs for morphological tags.

Baseline Combinations

Dataset	Test	Test	Dev
Weights	Default	Optimized	Default
Baseline RWTH	17.50±0.64	17.42±0.63	16.28±1.20
Add-to-training	-	17.25±0.62	16.58±1.25
Baseline RWTH+Moses	-	17.19±0.61	-
bojar-primary	-	16.90±0.61	16.00±1.15
google	-	16.76±0.60	17.32±1.25

- RWTH marginally better unoptimized (sys. weights equal).
- MERT opt. in Moses worse than JaneOpt in RWTH setup.

  Exceptionally, with milder pruning, Baseline RWTH+Moses got 17.57±0.61.
- Add-to-training works but very inefficient implementation:
  - Need to re-align, re-extract phrases, re-tune in MERT.

Improving Word Alignments

- GIZA++: No use of the fact that words are in the same lang.
  - Baseline:

    $\Rightarrow$ obdrží|nobelovu cenu|cenu **míru**|ϵ nobela|**míru**
  - Align lemmas and include an "equivalence dictionary"[1] in training:

    $\Rightarrow$ obdrží|nobelovu cenu|cenu **míru**|**míru** nobela|ϵ
- Some misalignments fixed, some errors remained.
- Also tried including automatically generated synonym classes.

[1] E.g. *míru*=*míru* as a separate sentence.

Results of Improving Alignments

	RWTH Optimizer		Moses MERT	
	Unoptimized	Optimized	Less Pruning	Dflt Pruning
Average±StdDev	17.52±0.01	17.45±0.05	17.32±0.06	17.25±0.10
eqvoc-lem-syndict	17.52±0.63	**17.51±0.62**	17.30±0.60	17.16±0.60
eqvoc-lem-syndict	17.51±0.62	17.48±0.61	17.33±0.60	17.00±0.58
eqvoc-lem-syndict	17.52±0.63	17.48±0.62	17.21±0.60	17.29±0.59
eqvoc-lem-syndict	17.51±0.64	17.48±0.63	17.27±0.61	17.32±0.61
eqvoc-stem3	17.52±0.63	17.48±0.62	17.41±0.64	17.35±0.62
eqvoc-lem	**17.53±0.63**	17.47±0.61	17.35±0.59	17.29±0.62
eqvoc-lem-syndict	**17.53±0.63**	17.47±0.62	17.26±0.61	17.29±0.60
eqvoc-lem-syndict	17.52±0.63	17.47±0.62	17.25±0.61	17.26±0.60
eqvoc-stem4	17.52±0.63	17.47±0.62	17.36±0.61	17.07±0.60
eqvoc-lem-syndict	17.52±0.64	17.46±0.64	17.36±0.62	17.32±0.61
eqvoc-lem-syndict	17.51±0.63	17.46±0.63	17.26±0.61	17.33±0.60
eqvoc-lem-syndict	17.49±0.63	17.45±0.63	17.34±0.61	17.32±0.58
lem	17.50±0.63	17.45±0.63	17.27±0.60	**17.37±0.61**
eqvoc	17.51±0.64	17.44±0.63	17.27±0.59	17.18±0.59
eqvoc-lem-syndict	**17.53±0.63**	17.44±0.61	17.22±0.59	17.21±0.60
eqvoc-lem-syndict	**17.53±0.63**	17.44±0.63	17.37±0.61	17.33±0.60
baseline	17.50±0.64	17.42±0.63	**17.57±0.61**	17.19±0.61
eqvoc-lem-syndict	17.52±0.64	17.37±0.61	17.41±0.63	17.30±0.63

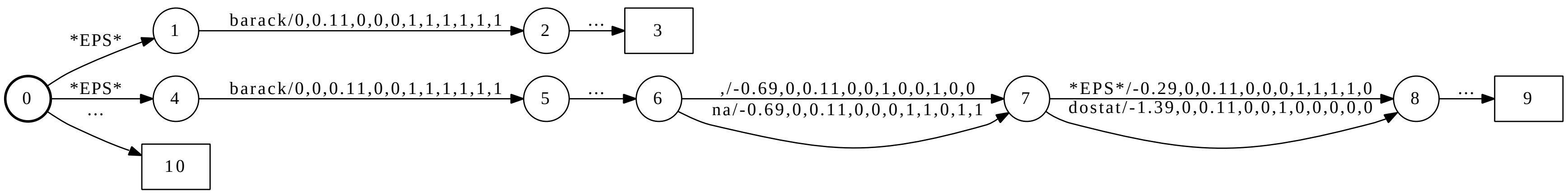
- Many variants of automatic synonym dict.
- Mixed results.
- Moses MERT less stable.

Lattice Arc Weights

- Remember: We need to select highest-scoring path in lattice.
  - Each arc contributes to the overall score of the path.
- The score can be a vector of components:

Apriori-weight. For each system and sentence (e.g. based on outside scores). So far not used.
Voting (RWTH). The percentage of systems producing this arc.
Sentence-level. One for each system, indicating whether the system provided the skeleon. Collected incrementally along the sentence.
Arc-level. One for each system, indicating if the arc was produced by the given system (incl. epsilon). These add up to voting-weight.
Primary-arcs. Indication whether the primary system produced this arc.
Primary-words (RWTH). Zero for eps., else indication whether the primary system produced this word.

- Weights of the components tuned using MERT on a dev set.



- Moses supports multiple weights and lattice input. (Dyer et al., 2008)

Indicators in Log-Linear Model

- Moses operates in log domain:
  - Scores are added along the path and multiplied by weights.
  - Normalization: Divide each weight by $\sum |w_i|$.

$\Rightarrow$ The encoding of indicators influences MERT search.

	Probability Domain		Log Domain		
Indicator Meaning	no	yes	no	yes	
Bad	0	1	$-\infty$	0	
Common	$e^0 = 1$	$e^1 \approx 2.7$	0	1	
Inverted	$e^1 \approx 2.7$	$e^0 = 1$	1	0	cf. tropical semiring
Minus-Plus	$e^{-1} \approx 0.3$	$e^1 \approx 2.7$	-1	1	

- Empirically Common/Inverted/Minus-Plus always differ but always fall within avg±stddev (3*7*18=378 experiments).

Larger LMs

- By default, only 3gr LM based on combined hypotheses is used.
- RWTH saw no gains from using additional LM (G. Leusch, p.c.).
- en→cs and Moses MERT do make use of that.
- Additional data: WMT10mono, 13M sents, 211M tokens.

	Underlying Alignment		
	Baseline	Eqvoc+Lemmas	$\oslash \pm \sigma$ Across All
RWTH Unoptimized	17.50±0.64	17.53±0.63	17.52±0.01
Moses +5grLM	17.36±0.61	17.49±0.61	17.48±0.06
Moses +4grLM	17.63±0.59	17.45±0.62	17.46±0.08
RWTH Optimized	17.42±0.63	17.47±0.61	17.45±0.05
Moses +3grLM	17.46±0.61	17.44±0.63	17.41±0.07
Moses (small LM)	17.32±0.63	17.34±0.61	17.32±0.06

- With the additional LM, Moses can reach RWTH optimizer.
- Higher n-grams marginally better.

LMs for Morphological Tags

- Bojar (2007) gains by using an additional LM over morphological tags in the factored translation (Koehn and Hoang, 2007).

Source	Target	
lowercase →	lowercase	
	morph. tag	**+ 6grLM**

- Hypotheses are "tagged with unigram tagger" on the fly.

	Underlying Alignment		
	Baseline	Eqvoc+Lemmas	$\oslash \pm \sigma$ Across All
Moses +tagLM, no Pruning	**17.88±0.62**	**17.95±0.59**	**17.90±0.12**
RWTH Unoptimized	17.50±0.64	17.53±0.63	17.52±0.01
RWTH Optimized	17.42±0.63	17.47±0.61	17.45±0.05
Moses (small LM)	17.32±0.63	17.34±0.61	17.32±0.06
Moses +tagLM, with Pruning	15.15±0.51	-	-

- Need to switch off beam pruning, tagged hyps wouldn't survive.

TagLM and Large LM

- We can combine TagLM and regular LM.
- This makes 15 weights in MERT optimization:
  - 9 arc weights, 3 LM weights, 2 tagger weights, word penalty.

Source		Target	
lowercase	→	lowercase	**+5grLM**
		morph. tag ←	+ 6grLM

	Baseline	Underlying Alignment Eqvoc+Lemmas	$\oslash \pm \sigma$ Across All
Moses +tagLM +5grLM	**18.01±0.66**	**17.80±0.59**	**17.97±0.09**
Moses +tagLM	17.88±0.62	17.95±0.59	17.90±0.12
RWTH Unoptimized	17.50±0.64	17.53±0.63	17.52±0.01
Moses +5grLM	17.36±0.61	17.49±0.61	17.48±0.06
RWTH Optimized	17.42±0.63	17.47±0.61	17.45±0.05
Moses (small LM)	17.32±0.63	17.34±0.61	17.32±0.06
RWTH Optimized AllSys	18.02±0.65	18.07±0.67	-

- In terms of BLEU score, this approaches the combination of all 7 systems.
- Incidentally, Moses +tagLM +5grLM using Minus-Plus got up to 18.26±0.64.

Manual Evaluation of Sys. Comb.

- Manually ranked 65 sentences.
  - All the hyps get either one of equally-*, or
  - At least one hyp gets 1 and others get lower ranks.

		Equally		Ranked as			
		Poor	Ok	1	2	3	4
Moses +tagLM +5grLM	**18.01±0.66**	11	7	18	16	10	3
RWTH Optimized	17.42±0.63	11	7	**22**	17	7	1
Moses (small LM)	17.32±0.63	11	7	17	14	14	2
bojar-primary	16.90±0.61	11	7	14	20	9	4
google	16.76±0.60						

- Improved over single-best.
- Results unstable, need many more sentences and annotators.

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 $\rightarrow$	form	+LM

or

Czech	English	
lemma $\rightarrow$	form	+LM

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

  $\Rightarrow$ See the talks by Chris Dyer and Marcello Federico.

Mixing Para+Mono

Simple concatenation (denoted ".").

- Just append the baseline parallel and the monolingual texts.

Interpolated in MERT (denoted "+").

- Separate weight for the LM trained on the monolingual data.
- Separate five weights for the phrase table extracted from the monolingual data.

Results

BLEU	TM	LM	Manual
10.56±0.39	para	para	
10.70±0.40	mono	mono	
10.98±0.38	mono	para+mono	
11.06±0.40	mono	para.mono	
12.20±0.40	para	para+mono	
12.24±0.44	para	para.mono	baseline
12.27±0.41	para.mono	para+mono	
12.33±0.43	para.mono	para.mono	29 over 19 better
12.65±0.42	para+mono	para.mono	35 over 27 better

- For LM, interpolation ("+") usually beats concat. (".").
  - Here domains match exactly ⇒ no gain.
- Reverse self-training works (TM "+") for en→cs small data.
- 2M monolingual (alone!) make a reasonable baseline (10.70±0.40).

Summary

- Generating target Czech forms is hard:
  - Failed factored attempts.
  - Promising two-step attempts.
  - Interesting black art of Reverse self-training.
- System combination (voting over words) for en→cs.
  - Moved to MERT optimization in Moses, more weights, LMs.
  - Improvement in BLEU thanks to TagLM.
  - Somewhat less convincing in manual evaluation.
  - Surely better than single-best outputs.

  . . . I would rather vote over "constituents". ⇝ Future.

Help us and combine ÚFAL's systems for WMT (due March 14).

Last chance to beat Google, if not too late already!

References

Jan Berka, Martin Černý, and Ondřej Bojar. 2011. Quiz-Based Evaluation of Machine Translation. Prague Bulletin of Mathematical Linguistics, 95, March.

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.

Ondrej Bojar and Kamil Kos. 2010. 2010 Failures in English-Czech Phrase-Based MT. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 60–66, Uppsala, Sweden, 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.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In Proceedings of ACL-08: HLT, pages 1012–1020, Columbus, Ohio, June. Association for Computational Linguistics.

J.G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, pages 347–354. IEEE.

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

Evgeny Matusov, Gregor Leusch, Rafael E. Banchs, Nicola Bertoldi, Daniel Dechelotte, Marcello Federico, Muntsin Kolss, Young-Suk Lee, Jose B. Marino, Matthias Paulik, Salim Roukos, Holger Schwenk, and Hermann Ney. 2008. System Combination for Machine Translation of Spoken and Written Language. IEEE Transactions on Audio, Speech and Language Processing, 16(7):1222–1237, September.

References

Evgeny Matusov. 2009. Combining Natural Language Processing Systems to Improve Machine Translation of Spee
Ph.D. thesis, RWTH Aachen University, Aachen, Germany, December.

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