

Forms Wanted: Training SMT on Monolingual Data



Ondřej Bojar, Aleš Tamchyna
bojar@ufal.mff.cuni.cz, a.tamchyna@gmail.com
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.
 - Caveat: BLEU much less reliable.
- Failed: Factored attempts to generate forms on the fly.
- Promising: Two-Step Translation.
- 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 $\geq 2,300$ tags seen	50 used
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			

Wide margin for improvement: Standard BLEU ~12% vs. lemmatized BLEU ~21%

OOV Rates



Dataset (# Sents)	Language	<i>n</i> -grams Out of: Corpus Voc.		Phrase-Table Voc.	
		1	2	1	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%

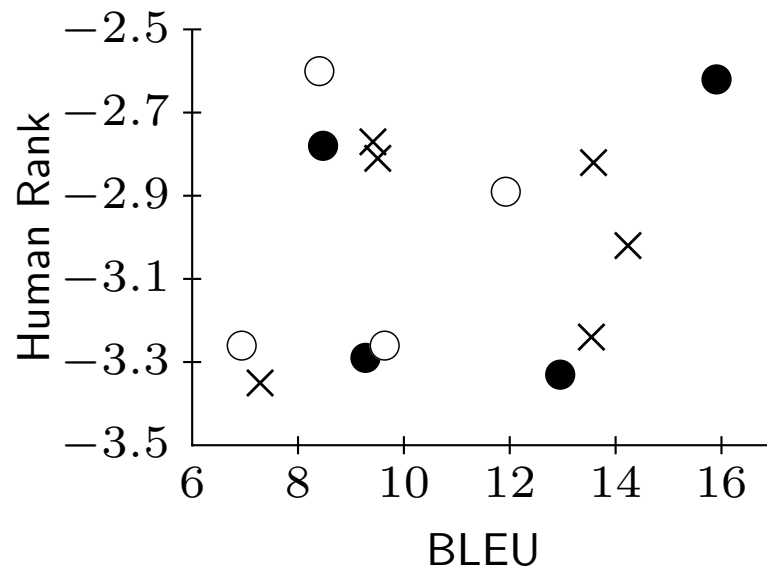
- Significant vocabulary loss during phrase extraction:
 - e.g. 2.2% → 3.9% for 7.5M Czech.
- OOV of Czech forms ~twice as bad as in English, cf. the reds.
- OOV of Czech lemmas lower than in English, see the greens.

Side Note: BLEU vs. Human Rank

- Large vocabulary impedes the performance of BLEU.

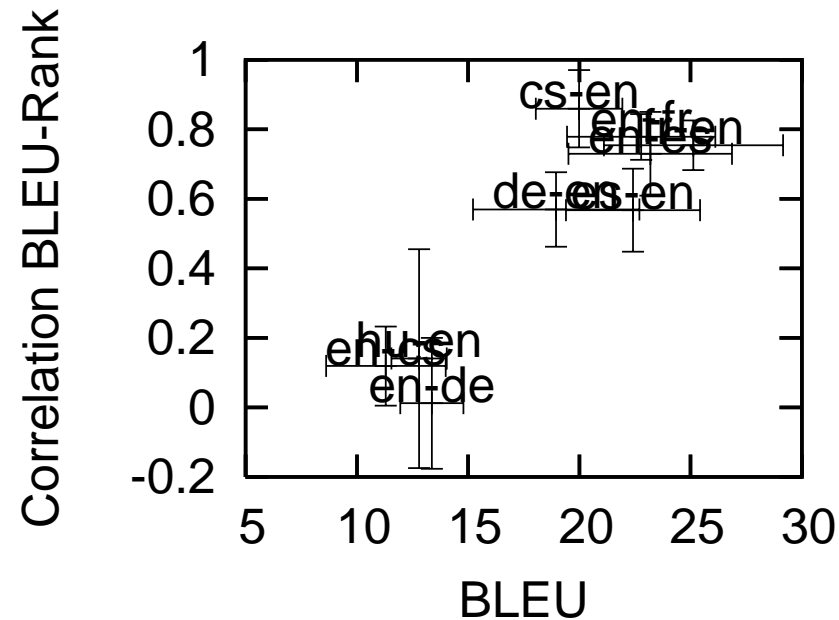
En→Cs Systems

WMT08, WMT09



Various Language Pairs

WMT08, WMT09, MetricsMATR



⇒ BLEU does not correlate with human rank if below ~ 20 .

Reason 1: Focus on Forms



SRC	Prague Stock Market falls to minus by the end of the trading day
REF	pražská burza se ke konci obchodování propadla do minusu
cu-bojar	praha stock market klesne k minus na <u>konci</u> obchodního dne
pctrans	praha trh cenných papírů padá minus <u>do</u> konce obchodního dne

- Only a single unigram in each hyp. confirmed by the reference.
- Large chunks of hypotheses are not compared at all.

Confirmed by Reference	Yes	Yes	No	No
Contains Errors	Yes	No	Yes	No
Running words	6.34%	36.93%	22.33%	34.40%

Reason 2: Sequences Overvalued



BLEU overly sensitive to sequences:

- Gives credit for 1, 3, 5 and 8 four-, three-, bi- and unigrams,
- Two of three serious errors not noticed,
⇒ Quality of cu-bojar overestimated.

SRC	Congress yields: US government can pump 700 billion dollars into banks					
REF	kongres ustoupil : vláda usa může do bank napumpovat 700 miliard dolarů					
cu-bojar	<u>kongres</u>	výnosy	<u>: vláda usa může</u>	čerpadlo	<u>700 miliard dolarů</u>	v bankách
pctrans	<u>kongres</u>	<u>vynáší</u>	<u>: us vláda může</u>	<u>čerpat</u>	<u>700 miliardu dolarů</u>	<u>do bank</u>

⇒ Bojar et al. (2010) use SemPOS, a coarse metric that correlates better with humans for Czech and English.

Factored Phrase-Based MT

- Both input and output words can have more factors.

Mapping 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 To ensure “vertical” coherence.

Target-side language models (+LM)

Applicable to various target-side factors.

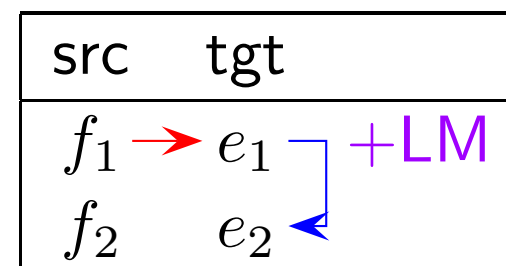
$p(\text{dvě kočkách}) < p(\text{dvě kočky})$; $p(\text{fem-nom masc-nom}) < p(\text{fem-nom fem-nom})$

\Rightarrow To ensure “horizontal” coherence.

(Koehn and Hoang, 2007)

src	tgt
f_1	e_1
f_2	e_2

+LM



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)



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 WMT08, T+C was still 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 before the linear context is available.
- Hieu Hoang wasted a year on trying asynchronous factors.
 - Pruning hard to design (no clear comparison for partial translation options).
- In a completely different decoder Bojar et al. (2009) use “delayed factors”.
 - The final value generated only after the full hypothesis is ready.

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
 - trained on much more data
 - max phrase len 1, monotone

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

Two-Step Attempts (WMT10) 2/2



Training Sents		Vanilla		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	↗ ↗
126k	13M	12.50±0.44	31.01	12.29±0.47	31.40	↘ ↗
7.5M	13M	14.17±0.51	33.07	14.06±0.49	32.57	↘ ↘

Manual micro-evaluation of ↘ ↗, i.e. 12.50±0.44 vs. 12.29±0.47:

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

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

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. . . <i>kočka honit. . . (lem.)</i>
	I saw a cat	=	viděl jsem kočku <i>vidět být kočka (lem.)</i>
Mono 2M	?		četl jsem o kočce <i>číst být o kočka (lem.)</i>
	I read about a cat	←	Use reverse translation 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

- We even don’t correct the bug that phrases available only in one of the tables score better than phrases scored by both paths.
- Other languages (e.g. Turkish, German) need different back-off techniques:
 - Split German compounds.
 - Separate and allow to ignore Turkish morphology.
 - ⇒ 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 \Rightarrow no gain.
- Reverse self-training works (TM “+”) for en \rightarrow cs small data.
- 2M monolingual (alone!) make a reasonable baseline (10.70±0.40).

- Czech is an interesting target language for MT.
 - `http://ufal.mff.cuni.cz/czeng`
 - 8 million parallel sents (~90 milion words per lang.)
- BLEU unreliable if below 20.
- Naive factored setup to generate unseen forms fails.
 - Search space explodes.
- Two-step approach (translate to lemmas first) promising.
 - Future: pass lattice between step 1 and 2.
 - Future: better model for step 2 (see Alex Fraser's talk).
- Reverse-self training works on small datasets.
 - Future: test on larger dataset.
 - Future: back-off for the reverse step in other languages.

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.

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, Miroslav Janíček, and Miroslav Týnovský. 2009. Evaluation of Tree Transfer System. Project Euromatrix - Deliverable 3.4, ÚFAL, Charles University, March.

Ondřej Bojar, Kamil Kos, and David Mareček. 2010. Tackling Sparse Data Issue in Machine Translation Evaluation. In Proceedings of the ACL 2010 Conference Short Papers, pages 86–91, Uppsala, Sweden, July. Association for Computational Linguistics.

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

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