English-to-Czech Factored Phrase-based Machine Translation

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

Overview

- Properties of Czech, Data Used
- Motivation and Brief Overview of Factored Phrase-Based MT
- Translation Scenarios
- Granularity of Part-of-Speech Tags
- More Data
- Untreated Morphological Errors
- Summary

Properties of Czech, Data Used

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

News Commentary Corpus - Training Data	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, Carroll, and Pearce, 2001).

Morphological Explosion in Czech, Margin

Apart from lexical ambiguity, MT system has to choose the correct word form:

- 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		. . .	. . .		

Margin for improvement: Standard BLEU ∼12% vs. lemmatized BLEU ∼21%

LM over Forms Insufficient

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!

Add Explicit Morphological Target Factor

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/fem-nom*	*masc/fem-nom*	*fem-nom*	
dvě	zelené	pruhované	kočky	← correct nominative/accusative
fem-nom	*fem-nom*	*fem-nom*	*fem-nom*	
dvěma	zeleným	pruhovaným	kočkám	← correct dative
fem-dat	*fem-dat*	*fem-dat*	*fem-dat*	

- LM over morphological tags generalizes better.
- Tagset size smaller than vocabulary ⇒ can afford e.g. 7-grams.

Factored Phrase-Based MT

More generic than the previous example (Koehn and Hoang, 2007):

- both input and output words can have more factors
- arbitrary number and order of:

Mapping steps ($\rightarrow$)

Translate (phrases of) source factors to target factors.

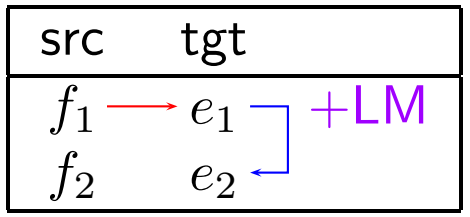
two green $\rightarrow$ dvě zelené

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(dvě kočkách) $<$ p(dvě kočky); p(*fem*-*nom* *masc*-*nom*) $<$ p(*fem*-*nom* *fem*-*nom*)

$\Rightarrow$To ensure "horizontal" coherence.

Translation Scenarios

Translate only (T)

English	Czech	
lowercase →	lowercase	+LM
lemma	lemma	
morphology	morphology	

Translate+Check (T+C)

English	Czech	
lowercase →	lowercase	+LM
lemma	lemma	
morphology	morphology ←	+LM

2·Translate+Check (T+T+C)

English	Czech	
lowercase →	lowercase	+LM
lemma	lemma	
morphology →	morphology ←	+LM

2·Translate+Generate (T+T+G)

English	Czech	
lowercase	lowercase ←	+LM
lemma →	lemma	+LM
morphology →	morphology	+LM

Results of Various Scenarios

	BLEU
Baseline: T	12.9±0.6
T+C	13.6±0.6
T+T+C	13.9±0.6
T+T+G	13.9±0.7

⇒ Multi-factor better than single-factored.

⇒ More complex scenarios do not significantly outperform T+C.

Granularity of POS in T+T+C

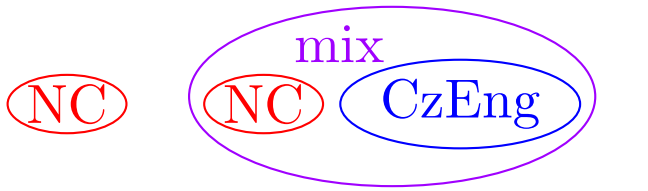
Tagset	Size	Description
full tags	1200	Full Czech positional tag, 15 positions, about 4000 tags defined.
POS+case	184	POS and SUBPOS, for nouns, adjs, pronouns and preps also case.
CNG01	621	Like POS+case, with case, number and gender for {N, A, P, R}.
CNG02	791	Case, number and gender for {N,A,P,R,C,V}, punctuation tag encodes the lemma.
CNG03	1017	Case, number, gender; highlighted reflexive *se/si*. Verbs distinguish number, gender, tense, passivization; highlited *to be*. Preps encode case+lemma, . . . , numbers encode "shape".

Results when Varying Granularity

	BLEU
Baseline: T (single-factor)	12.9±0.6
T+T+C, POS+case	13.2±0.6
T+T+C, CNG01	13.4±0.6
T+T+C, CNG02	13.5±0.7
T+T+C, CNG03	14.2±0.7
T+T+C, full tags	13.9±0.6

⇒ CNG03, the highly optimized tagset, works best.

More Out-of-Domain Data in T+C



$\text{weighted} = \alpha\text{NC} + \beta\text{mix}$

NC . . . 56k sents
mix . . . 1071k sents

Scenario	Phrases from	LMs	BLEU
T	NC	NC	12.9±0.6
T	mix	mix	11.8±0.6
T	mix	weighted	11.8±0.6
T+C CNG03	NC	NC	13.7±0.7
T+C CNG03	mix	mix	13.1±0.7
T+C CNG03	mix	weighted	13.7±0.7
T+C full tags	NC	NC	13.6±0.6
T+C full tags	mix	mix	13.1±0.7
T+C full tags	mix	weighted	13.8±0.7

⇒ Ignoring domain difference usually worse than tuning separate LMs.

⇒ Full tags as good as CNG03 or better in large data setting.

Untreated Morphological Errors

Micro-study: 77 Verb-Modifier pairs in 15 sample *source* sentences:

Translation of	**Verb**	**Modifier**
. . . preserves meaning	56%	79%
. . . is disrupted	14%	12%
. . . is missing	27%	1%
. . . is unknown (not translated)	0%	5%

Even when Verb&Mod correct, 56% of cases are non-grammatical or meaning-disrupted relations.

Sample Errors

Input:	Keep on investing.
MT output:	Pokračovalo investování. (grammar correct here!)
Gloss:	*Continued investing. (Meaning: The investing continued.)*
Correct:	Pokračujte v investování.

⇒language model misled us ⇒ need to include source valency information.

Input:	brokerage firms rushed out ads . . .			
MT Output:	brokerské	firmy	vyběhl	reklamy$_{pl.nom/pl.acc/pl.voc/sg.gen}$
Gloss:	*brokerage*	*firms*	*ran*	*ads*
Correct option 1:	brokerské	firmy	vyběhly	s reklamami$_{pl.instr}$
Correct option 2:	brokerské	firmy	vydaly	reklamy$_{pl.acc}$

Target-side data may be rich enough to learn: vyběhnout–s–$instr$

Not rich enough to learn all morphological and lexical variants:

vyběhl–s–reklamou, vyběhla–s–reklamami, vyběhl–s–prohlášením, vyběhli–s–oznámením, . . .

Summary

- Explicit modelling of target-side morphology improves translation.
- More complex scenarios not significantly better than T+C.
- Fine-tuning of tagset useful in small-data setting only.
- Verb-modifier relations still quite poor.

ufal.mff.cuni.cz → Research → Online Demo

References

Hajič, Jan 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.

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

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

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