Experiments with Czech→English Phrase-Based Translation

Ondřej Bojar, Evgeny Matusov, Hermann Ney
obo@cuni.cz, {matusov,ney}@cs.rwth-aachen.de

August 23, 2006

Outline

- Properties of Czech language.
- Phrase-based MT on one slide.
- Data overview: Small data but large vocabulary task.
- Improving MT quality.
  - Reliable alignment.
  - Simple rule-based handling of numbers.
  - Dependency-based corpus expansion.
  - More monolingual and parallel data.
  - Fixing clear BLEU errors.
- Comparison with other systems.
- Summary.

Properties of Czech Language

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

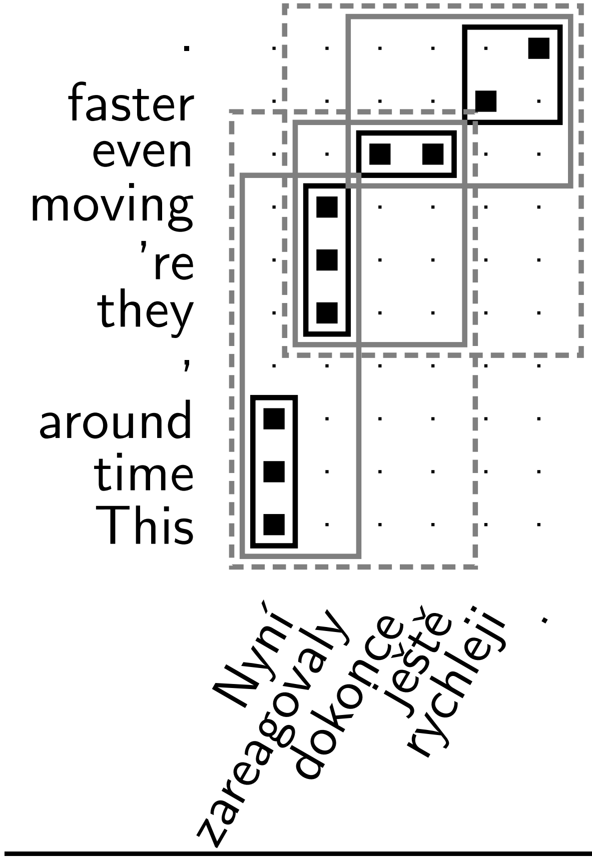
- rigid global word order phenomena: clitics
- rigid local word order phenomena: coordination, clitics mutual order

Nonprojective sentences	16,920	23.3%
Nonprojective edges	23,691	1.9%

Known parsing results	Czech	English
Edge accuracy	69.2–82.5%	91%
Sentence correctness	15.0–30.9%	43%

Data by (Collins et al., 1999), (Holan, 2003), Zeman (http://ckl.mff.cuni.cz/˜zeman/ /projekty/neproj/index.html) and (Bojar, 2003). Consult (Kruijff, 2003) for measuring word order freeness.

Alignments, Phrases and Phrase-Based MT



This	=	nyní
time	=	nyní
around	=	nyní
they	=	zareagovaly
. . .	=	. . .
This time around	=	Nyní
they 're moving	=	zareagovaly
even	=	dokonce ještě
. . .	=	. . .
This time around, they 're moving	=	Nyní zareagovaly
even faster	=	dokonce ještě rychleji
. . .	=	. . .

Phrase-based MT: choose such segmentation of input string and such phrase "replacements" to make the output sequence "coherent" (3-grams most probable).

Data Overview: Small Data but Large Vocabulary

Prague Czech-English Dependency Treebank (PCEDT) 1.0:

- Wall Street Journal section of Penn Treebank translated to Czech.

		Czech	English
Train:	Sentences	21,106	
	Tokens	474,452	493,462
	Vocabulary	56,970	30,739
	Singletons	55.1%	47.6%
Dev:	Sentences	259	
	Tokens	6,386	6,522
	Tokens out of vocabulary	7.3%	3.6%
Test:	Sentences	256	
	Tokens	5815	6175
	Tokens out of vocabulary	8.2%	4.3%

Obtaining Reliable Alignment

	BLEU (ETest)		Alignment Error Rate	
	Intersection	Union	Intersection	Union
Baseline (word forms)	28.2	29.8	27.4	25.5
Stemming	-	30.6	-	-
Lemmas	29.8	32.0*	15.0	17.2
Lemmas + singletons	30.8*	31.9*	14.6	17.4

⇒ Use full lemmatization, if possible.

⇒ Alignment Error Rate does not correlate with performance of MT.

		Vocab		Singl/Vocab	
Type of input for the alignment		CZ	EN	CZ	EN
Forms	Produkce malých vozů se více než ztrojnásobila .	57k	31k	55.1%	47.6%
Stem4	Prod malý vozů se více než ztro .	17k	14k	36.5%	35.8%
Lem+Sing	produkce malý vůz se hodně než-2 UNK-verb .	15k	13k	0.1%	0.0%
Lemmas	produkce malý vůz se hodně než-2 ztrojnásobit .	28k	25k	46.4%	47.5%

Handling Numbers

- Numbers cause severe data sparseness.
- Should be handled uniformly, but some "translation" of them is needed.

	Sample input	Input to PBT	Output
Baseline	na 57,375 dolarech	na 57,375 dolarech	at 57,375 $
Numbers	na 57,375 dolarech	na _NUM dolarech	at $ 57,375
Numbers + Correction	na 57,375 dolarech	na _NUM dolarech	**at $ 57.375**

	Devtest	Etest
Baseline	34.6	32.0
Numbers	34.1	30.9
Numbers + Correction	34.7	32.9*

Dependency-based Corpus Expansion

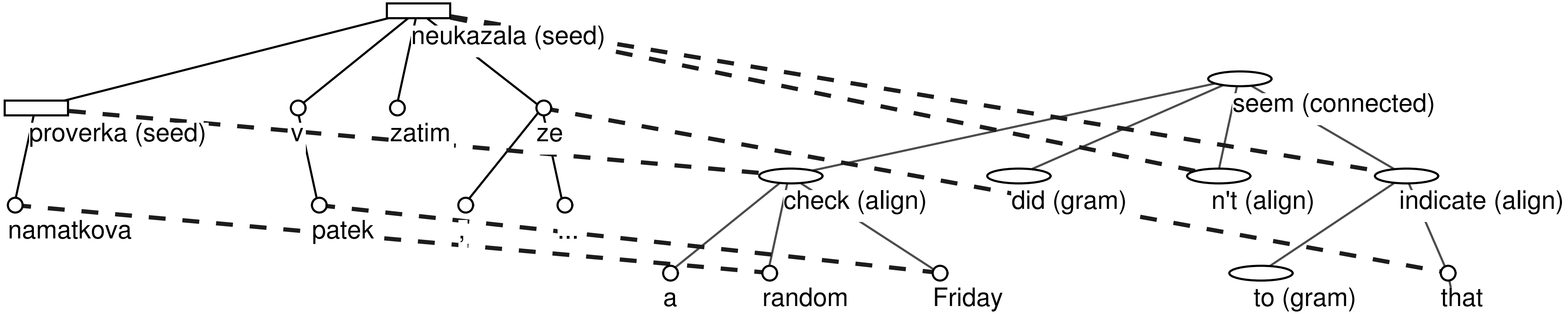
Create new training sentences (with new n-grams) by deleting aligned leaves of dependency structures ("reducing sentences").

- Off-line: print all possible reduced sentences given the training corpus ⇒unbearable, explosion of data.

- On-line: given the test source data (the set of "needed" n-grams)
  - Scan training corpus for sentences with sample *non-contiguous* occurrences of the needed n-grams.
  - Mark the source nodes, aligned nodes and all dependency neighbours needed for a certain level of grammaticality.
  - Print marked nodes.

Dependency-Based Corpus Expansion: Example

- Test data need to translate *prověrka neukázala*.
- Training data provide this bigram, *non-contiguous* in linear order but close in dependency tree.
- After marking aligned nodes and nodes for grammaticality, we obtain: *prověrka neukázala = check did n't seem to indicate*



A opravdu , namátková prověrka v pátek zatím neukázala , že by stávka měla dopad na ostatní letecké operace .

Indeed , a random check Friday did n't seem to indicate that the strike was having much of an effect on other airline operations .

Corpus Expansion Does Not Help

	Devtest			Etest		
Training sentences	5k	10k	20k	5k	10k	20k
Baseline	27.5	31.6	34.6	25.4	28.4	32.0
Expanded Corpus	27.4	31.9	34.5	25.0	28.0	32.3

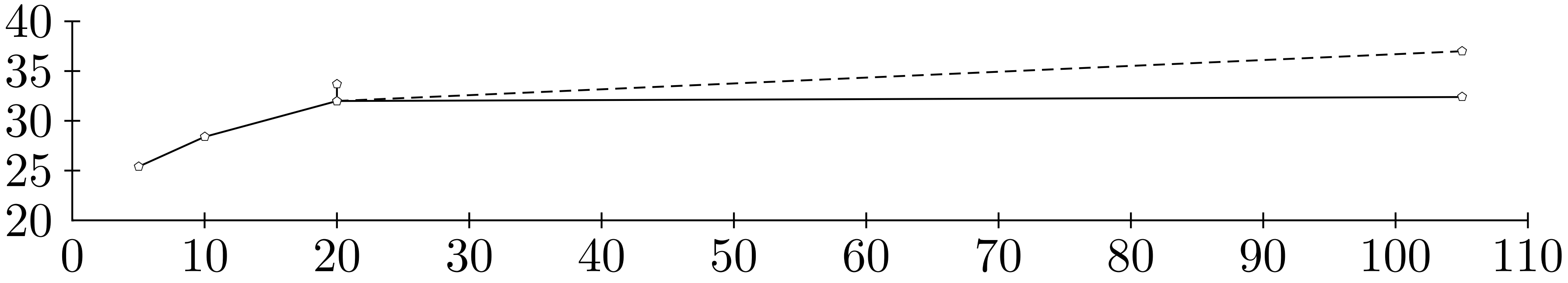
Reasons:

- Small margin for improvement.
  Inherent distributional properties of languages (words close in dep.tree ⇒ likely to occur adjacently anyway).
  Phrase-based system can back-off to translate using shorter phrases.

- Random errors ⇒ low quality of generated phrases.
  Bad translation, bad alignment, bad automatic tree, bad selection of nodes.

Impact of Additional Data

	Devtest	Etest
Mini: 5k sentences	27.5	25.4
Half: 10k sentences	31.6	28.4
Baseline: 20k sentences	34.6	32.0
20k + 85k out-of-domain sentences	36.6*	32.4
20k sentences, bigger in-domain LM	37.9*	33.7*
20k + 85k out-of-domain sentences, bigger in-domain LM	40.9*	37.0*



A Method for Finding Clear BLEU Errors

Top missing bigrams:

19	, "	12	” said
12	of the	10	Free Europe
10	Radio Free	7	. "
6	L.J. Hooker	6	United States
6	in the	6	the United
6	the strike		. . .

Top superfluous bigrams:

26	, ’’	18	’’ .
14	” said	12	, which
11	Svobodná Evropa	8	, when
8	the state	7	, who
7	J. Hooker	7	L. J.
7	company GM		. . .

Missing bigram = all references contained it but not the hypothesis.
Superfluous bigram = the hypothesis contained it but none of the references.

Four simple rules to improve BLEU by +0.2 to +0.5 on Etest:

’’ .	→	. "	L. J. Hooker	→	L.J. Hooker
’’	→	"	the U.S.	→	the United States

Comparison with Other Systems

	Average over 5 refs.		4 refs. only	
	Devtest	Etest	Devtest	Etest
DBMT with parser I, no LM	18.6	16.3	-	-
DBMT with parser II, no LM	19.2	17.1	-	-
GIZA++ & ReWrite, bigger LM	22.2	20.2	-	-
PBT, no additional LM	38.7	34.8	36.3	32.5
PBT, bigger LM	41.3	36.4	39.7	34.2
PBT, more parallel texts, bigger LM	42.3	38.1	41.0	36.8

⇒ Phrase-based system twice as good as dependency-based system.

(BLEU is not fair, LM-based systems score better.)

(Various implementations of BLEU and various evaluation settings give non-comparable results.)

Summary of Phrase-Based MT Impressions

lemmatization for alignment	+2.0*
handling numbers	+0.9*
fixing clear BLEU errors	+0.5
dependency-based corpus expansion	+0.3
more out-of-domain parallel texts, also in LM	+0.4
bigged in-domain LM	+1.7*
more out-of-domain parallel texts, bigger in-domain LM	+5.0*

Given BLEU as “the” MT metric:

- Phrase-based system performs surprisingly well (in this easy setting).
- With small data, focus on alignments, corpus specifics and clear errors.
- With more data, in-domain language model is vital.

Sample Output

System Output:

We 'll see whether the campaigns work .

Immediately after Friday 's 190 14-point stock market and a consequent uncertainty excretes several big brokerage firms new ads UNKNOWN_vytrubující usual message : Go on in investing , the market is in order .

Their business is persuade clients from escaping from the market , which individual investors masse fact , after plunging in October .

Source:

Uvidíme , zda reklama funguje .

Okamžitě po pátečním 190 bodovém propadu akciového trhu a následné nejistotě vypouští několik velkých brokerských firem nové inzeráty vytrubující obvyklé poselství : Pokračujte v investování , trh je v pořádku .

Jejich úkolem je odradit klienty od útěku z trhu , což jednotliví investoři hromadně činili po propadu v říjnu .

References

Bojar, Ondřej. 2003. Towards Automatic Extraction of Verb Frames. *Prague Bulletin of Mathematical Linguistics*, 79–80:101–120.

Collins, Michael, Jan Hajič, Eric Brill, Lance Ramshaw, and Christoph Tillmann. 1999. A Statistical Parser of Czech. In *Proceedings of 37th ACL Conference*, pages 505–512, University of Maryland, College Park, USA.

Holan, Tomáš. 2003. K syntaktické analýze českých(!) vět. In *MIS 2003*. MATFYZPRESS, January 18–25, 2003.

Kruijff, Geert-Jan M. 2003. 3-Phase Grammar Learning. In *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development*.