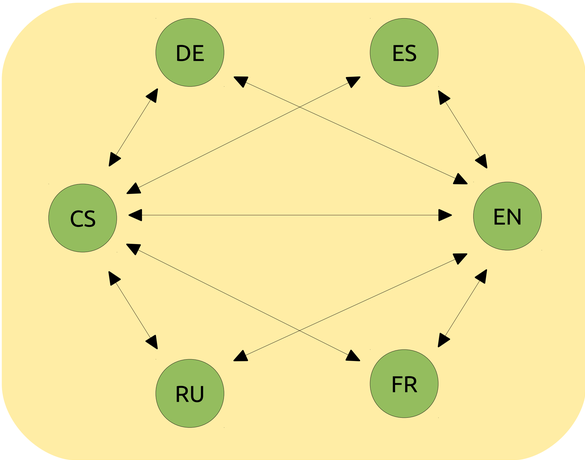
CUni Multilingual Matrix in the WMT2013 Shared Task

Daniel Zeman, Karel Bílek
Charles University in Prague, Institute of Formal and Applied Linguistics
Univerzita Karlova v Praze, Ústav formální a aplikované lingvistiky
Malostranské náměstí 25, CZ-11800 Praha
zeman@ufal.mff.cuni.cz, kb@karelbilek.com

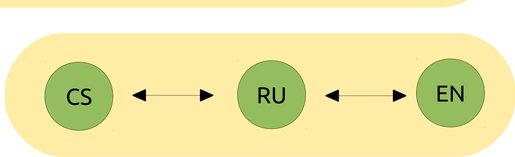
LANGUAGES



Goal of first set:

One system with similar configuration on more language pairs, try several corpora, compare results

(constrained)



Goal of second set, with Rusian:

Take as much data as you can, "throw" it on it, see if it helps

(unconstrained)

PARALLEL DATA

Corpus	SentPairs	Tokens lng1	Tokens lng2
cs-en	786,929	18,196,080	21,184,881
de-en	2,098,430	55,791,641	58,304,756
es-en	2,140,175	62,444,507	59,811,355
fr-en	2,164,891	70,363,304	60,583,967
ru-en	150,271	3,889,251	4,100,148
de-cs	657,539	18,160,857	17,788,600
es-cs	697,898	19,577,329	18,926,839
fr-cs	693,093	19,717,885	18,849,244
ru-cs	103,931	2,642,772	2,319,611

Number of sentence pairs and tokens for every language pair in the parallel training corpus (EuroParl v7 + News Commentary v8) for the basic corpora. The xx-cs corpora were obtained as intersections of xx-en and cs-en.

Corpus	SentPairs	Tokens lng1	Tokens lng2
CzEng cs-en	14,833,358	204,837,216	235,177,231
UN es-en	11,196,913	368,154,702	328,840,003
fr-en	12,886,831	449,279,647	372,627,886
Gigaword fr-en	22,250,400	854,353,231	694,394,577

Number of sentence pairs and tokens for every language pair in the parallel training corpus for the additional corpora.

Corpus	SentPairs	Tokens lng1	Tokens lng2
UMC ru-cs	93,395	2,073,102	2,019,683
Subs ru-cs	2,324,373	16,019,077	15,956,553
Intercorp ru-cs	148,847	1,564,967	1,613,830
UMC ru-en	92,775	2,046,753	2,253,070
Subs ru-en	1,790,209	10,220,121	8,831,100
Yandex ru-en	1,000,000	23,975,583	26,205,200
wikihead ru-en	514,859	2,642,772	2,319,611
CCrawl ru-en	878,386	17,399,366	18,772,065

Number of sentence pairs and tokens for additional Russian-Czech data for the unconstrained experiments

MONOL. DATA

Corpus	Segments	Tokens
newsc+euro.cs	830,904	18,862,626
newsc+euro.de	2,380,813	59,530,113
newsc+euro.en	2,466,167	67,033,745
newsc+euro.es	2,330,369	66,928,157
newsc+euro.fr	2,384,293	74,962,162
newsc.ru	183,083	4,340,275
news.all.cs	27,540,827	460,356,173
news.all.de	54,619,789	1,020,852,354
news.all.en	68,341,615	1,673,187,787
news.all.es	13,384,314	388,614,890
news.all.fr	21,195,475	557,413,929
news.all.ru	19,912,911	361,026,791
gigaword.en	117,905,755	4,418,360,239
gigaword.es	31,304,148	1,064,660,498
gigaword.fr	21,674,453	963,571,174

Number of segments (paragraphs or sentences) and tokens for every monolingual corpus in the constrained experiments. Newsc = News Commentary; news.all = crawled news from all the years.

Corpus	Segments	Tokens
zpravostroj.cs	1,531,403	23,424,109
webcoll.cs	4,053,223	78,688,001
pdt.cs	115,844	1,957,246
wikipedia.cs	3,695,172	58,068,659
okoun.cs	580,249	10,382,272

Number of segments (paragraphs or sentences) and tokens for every monolingual corpus in the unconstrained experiments.

THE SYSTEMS

- Tokenization
- Constrained system: in-house quotation marks normalization, supervised truecasing (based on lemmas ← TreeTagger or Morče)
- Unconstrained system: very quick stemmer
- Word alignment: Giza++ operating on lemmas
- Hexagram language model
- Moses decoder, no lexical reordering model, no factored model
- All BLEU scores computed by the system on tokenized test data, truecased (except for the sentence-initial letter)

Supervised truecasing
Training data with lemma factor:
A token is lowercased unless its lemma is uppercased.

Changes sentence-inital tokens, words in English headings, highlighting uppercase etc. *Errors on common-proper ambiguities?*

WMT RESULTS

Dir.	Score	Range
cs-en	0.543	6-7/11
de-en	0.396	14/17
fr-en	0.420	10-11/13
sp-en	0.462	10/12
en-de	0.460	10-12/15
en-fr	0.427	12/16
en-sp	0.446	10-11/13
en-cs	0.505	5-7/12
en-ru	0.331	14/14
en-ru unc.	0.498	8/14
ru-en	0.215	19/19
ru-en unc.	0.476	13-15/19

Official results for WMT translation tasks, from *Findings of the 2013 Workshop on Statistical Machine Translation*

BLEU

LM	news	newsall+czeng	**newsall+czeng**		
TM	news	czeng	**news+czeng**		
en-cs	16.32	17.79	**17.86**		
LM	**news**				
TM	**news**				
en-de	**18.33**				
LM	news	news+gigaw	**news+gigaw**	un+gigaw	
TM	news	news	**news+un**	un+gigaw	
en-es	28.08	28.56	**28.44**	20.16	
LM	news	news+gigaw	news+gigaw	newsall	un	**news+gigaw**
TM	news	news	giga	news+giga	un	**news+un**
en-fr	29.87	29.88	31.06	30.37	29.33	**30.10**
LM	news	news+gigaw	newsall	**newsall**	newsall	
TM	news	news	news+czeng	**news+czeng**	news	
cs-en	23.28	23.67	24.56	**25.27**	22.73	
LM	news	news+gigaw	news	news+un	**newsall+giga**	
TM	news	news	news+un	un+gigaw	**news**	
es-en	29.16	29.75	28.46	20.67	**28.56**	
LM	**news**	news+gigaw	news+un			
TM	**news**	news+un	un			
fr-en	**28.87**	29.14	27.37			

LM	**news**	news+gigaw
TM	**news**	news
de-en	**23.89**	24.36
LM	**news**	*unconstrained*
TM	**news**	*unconstrained*
en-ru	**15.82**	*16.30*

LM	*unconstrained*	*unconstrained*	*unconstrained*
TM	**news**	news+gigaw	news
ru-en	**19.75**	20.03	*22.40*

LM	**news**	
TM	**news**	
cs-de	**15.95**	
LM	news	**news+gigaw**
TM	news	**news**
cs-es	21.70	**22.20**
LM	**news**	news+gigaw
TM	**news**	news
cs-fr	**22.20**	21.96
LM	**news**	*unconstrained*
TM	**news**	*unconstrained*
cs-ru	**15.00**	16.50
LM	**news**	
TM	**news**	
de-cs	**14.88**	
LM	**news**	
TM	**news**	
es-cs	**15.80**	
LM	**news**	*unconstrained*
TM	**news**	*unconstrained*
ru-cs	**15.06**	15.80

LM	**news**
TM	**news**
fr-cs	**15.06**

The research has been supported by the grants P406/11/1499 and GAUK 639012.