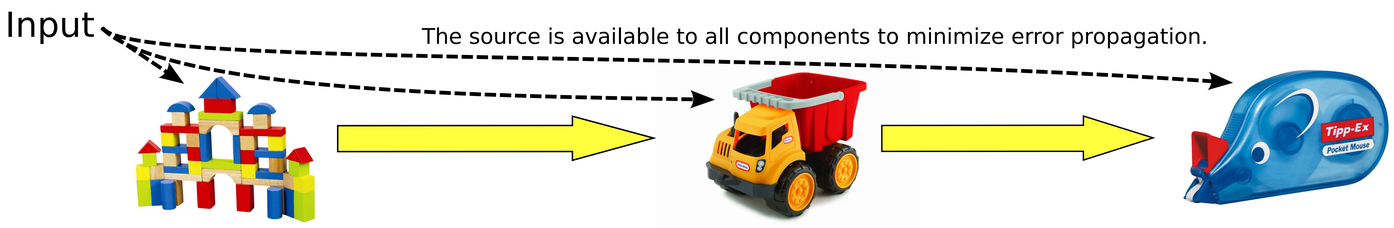
CUNI in WMT14: Chimera Still Awaits Bellerophon

Ondřej Bojar, Martin Popel, Rudolf Rosa, Aleš Tamchyna {bojar,popel,rosa,tamchyna}@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague http://ufal.mff.cuni.cz

English → Czech



TectoMT

- hybrid (rule-based/statistical) MT system
- transfer at a deep syntactic layer (t-layer)
- our combination: get an extra phrase table for Moses from TectoMT output

Sentence Structure and Unseen Forms

Moses

- phrase-based SMT
- large-scale data
- morphological tags as factors for a better grammatical coherence

0.2 GWord Parallel 3.6 GWord Czech

Depfix

- rule-based error correction in (S)MT output
- output parse corrected based on source

not
We're FC Barcelona!

Final Results

System	BLEU	TER	Manual	TectoMT	Factored Moses	Adapted LM	Document-specific LMs	Depfix
CU-DEPFIX	21.1	0.670	0.373	✓	✓	✓		✓
UEDIN-UNCONSTRAINED	21.6	0.667	0.357					
CU-BOJAR	20.9	0.674	0.333	✓	✓	✓		
CU-FUNKY	21.2	0.675	0.287	✓	✓	✓	✓	✓
GOOGLE TRANSLATE	20.2	0.687	0.168					
CU-TECTOMT	15.2	0.716	-0.177	✓				
CU-BOJAR +full 2013 news	20.7	0.677	–					

English → Hindi

Essentially baseline Moses.

	Sentences	Tokens Hindi	Tokens English
HindEnCorp	276k	4.09M	3.95M
NewsCrawl	1.27M	27.27M	
HindMonoCorp	43.38M	945.43M	
Total	44.93M	976.80M	

Morphological Processing of Hindi

Back-off	% of vocab. size
stem4	30
lemma4	32
lemma	90
form	100

- Siva Reddy's POS tagger

Impact of Word Alignment on BLEU

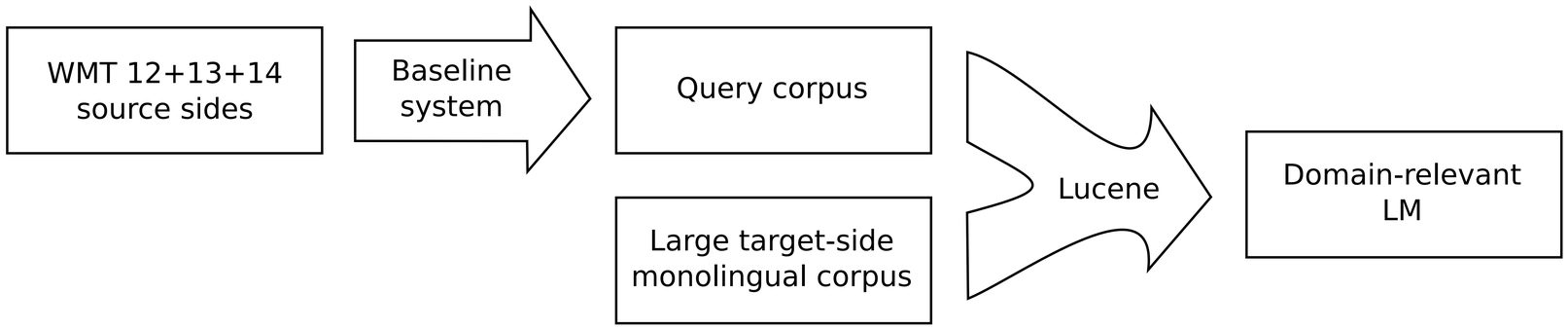
English	Hindi	BLEU
stem4	stem4	22.96+-1.17
lemma	lemma4	22.59+-1.17
lemma	lemma	22.41+-1.20

Other Failed Attempts

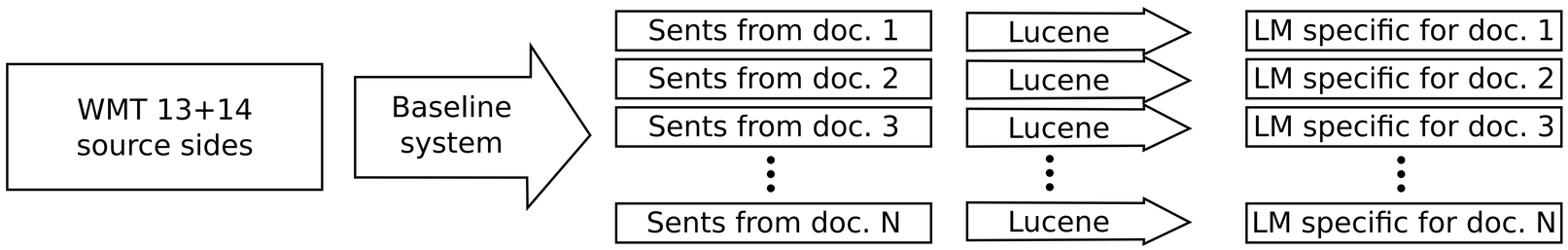
- Factored setup (form→form+POS) with 10-gram POS-LM
- Reverse Self-Training

NEW this year! Proved approach from WMT13

Adapted Language Model



Document-specific Language Models



Presented at WMT 2014, Baltimore, USA.
This research was supported by the grants FP7-ICT-2013-10-610516 (QTLeap), FP7-ICT-2011-7-288487 (MosesCore), SVV 260 104 and GAUK 1572314.
This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).