English-Hindi Translation – Obtaining Mediocre Results with Bad Data and Fancy Models

Ondřej Bojar, Pavel Straňák, Daniel Zeman, Gaurav Jain, Michal Hrušecký, Michal Richter, Jan Hajič

ÚSTAV FORMÁLNÍ A APLIKOVANÉ LINGVISTIKY

Data Issues and Normalization

Different corpora were processed differently:
- Tides:
  - Sentence ends with a full stop (.)
  - Euro-arabic digits (0123456789)
- Emille:
  - Sentence ends with a danda (।)
  - Devanagari digits (०१२३४५६७८९)
- What else can be written in more ways:
  - Characters with nukta (क़ख़ग़ज़ड़ढ़फ़}): फ़ vs. फ+़ vs. फ
  - Combined diacritics ordering: प+ा+ॅ vs. प+ॅ+ा
  - Candrabindu replaced by anusvara: पाँच vs. पांच
  - Control characters, zero-width joiners etc.
  - Non-ASCII punctuation, e.g. "—" vs. "-"
- We try to normalize all of this
- In addition we re-tokenize (Anglo-American)

Impossible to normalize:
e.g. varying transcription of English words:
स्टैंडर्डज (*sṭaiṁḍarḍaja*)
स्टैंडर्डस (*sṭaiṁḍarḍasa*)
स्टैंडर्ड्स (*sṭaiṁḍarḍsa*)

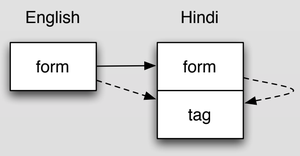
More problems in the data:
- During encoding conversions a parenthesis in English is wrongly considered to be romanized Hindi:
  - Information Commis(s)ioner => ईन्फ़ोर्मेटिओन् छ्ोम्मिसिओनेर् (*īnṅormaṭion chommisioner*), real transcription might be इन्फ़ोर्मेशन कोमिशनेर (*informeśana komiśanera*)
- More than 200 Hindi sentences in Tides start in devanagari but then switch into unreadable latin text:
  - प्रादेशिक - जनसंख्या बंगाली बंग्लादेश ह्यपूर्वी बंगालह् से आए अधिकांश विस्थापित दक्षिण अंडमान , नेल , हैवलाक , मध्य अंडमान , उ <arI AMDmaana tqaa ilaiTla AMDmaana maoM basaae gae .
- Character danda (end-of-sentence) changed to a vertical bar, then encoded as |BAR;, and then latin letters re-encoded into Devanagari: |भाष्;
- Recurring mysterious sequence ऋ-ऊण्ण्छष्- Q-UNSCR-; appears anywhere – even in the middle of a Hindi word.

Related Results

System	BLEU
Mumbai (Damani et al., 2008)	8.53
Kharagpur (Goswami et al., 2008)	9.76
Prague (Bojar et al., 2008)	10.17
Dublin (Srivastava et al., 2008)	10.49
present Joshua	**11.10**

Morphology in Moses

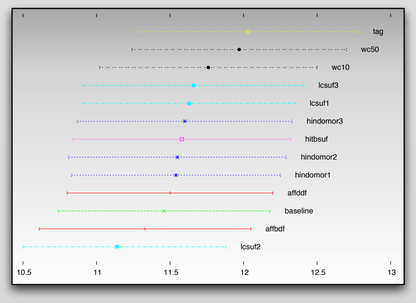
- Moses supports explicit modeling of morphology on the target side.
- An additional language model is applied on the stream of target-side tags.



We experiment with several formalizations of Hindi morphology:

			10 Word			Affisix		
Form	Tag	Textbook	2 Letters	Classes	Hindomor	bbf	bdf	ddf
उन्हें	PRP	उन्हें	◌ें	2	◌	—	—	—
वहां	PRP	वहां	◌ां	2	◌	—	—	—
कलकत्ता	NNP	आ	ता	3	◌ा	ता	ता	—
शहर	NN	शहर	हर	3	र	र	—	—
दिखाया	VM	आ	या	7	◌ा	◌ा	—	—
गया	VAUX	गया	या	11	◌ा	—	—	—
.	SYM	.	.	6	—	—	—	—

... with no significant improvement.



English: They were shown Calcutta City .
Hindi: उन्हें वहां कलकत्ता शहर दिखाया गया .
Gloss: them there Calcutta City shown was .

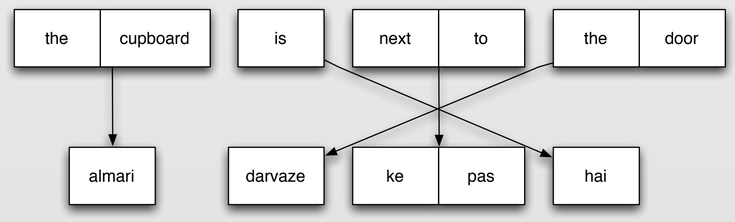
- Tags too coarse-grained for Hindi morphology.
- Automatic word classes seem to match the tags.
- Different configurations of Affisix provide different granularity.

Conclusion

- Best published BLEU score for TIDES test set achieved.
  - In general, the English-to-Hindi MT comparison in problematic due to different datasets used by various research groups.
- Hierarchical models (Joshua) lead to better BLEU than Moses with morphological factors.
  - Manual evaluation less conclusive about the improvement.
- Lessons learned about the data:
  - Obtaining data is easier than cleaning them up.
  - Two different corpora from different sources may overlap!

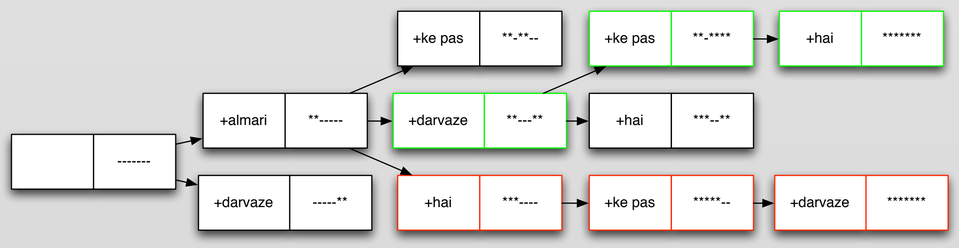
Moses vs. Joshua

- English-to-Hindi translation requires significant amount of reordering:



Phrase-based decoding (Moses)

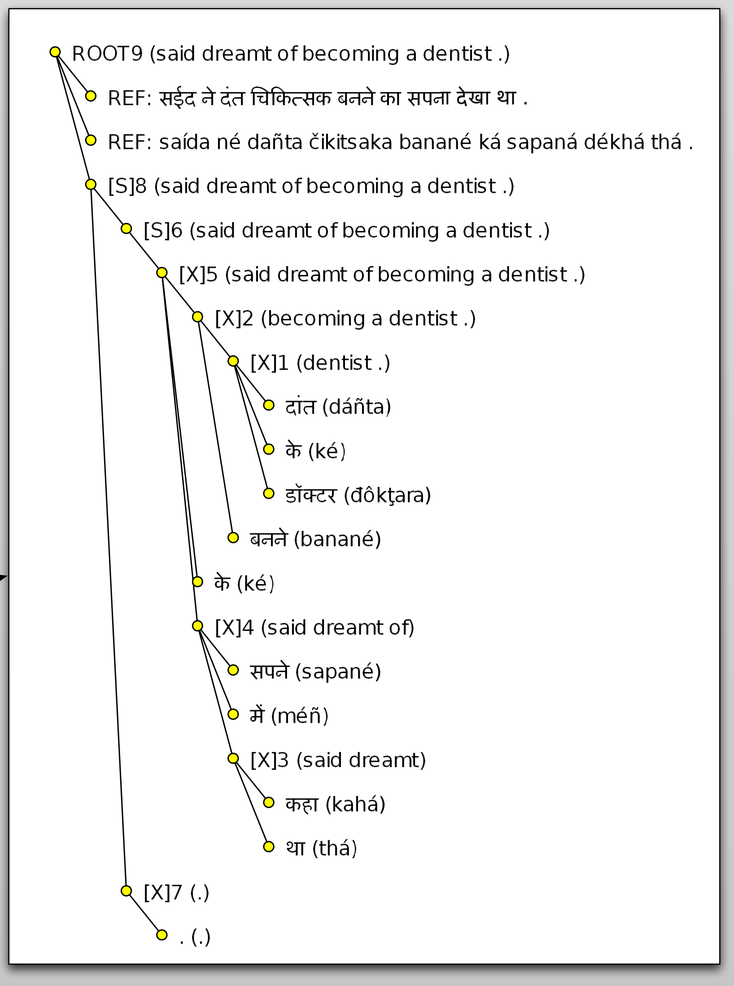
- Phrase-based models (Moses) explore the space of hypotheses from left to right.
- The default reordering model promotes monotone translation.
- The best available alternative is lexicalized reordering: swap phrases given the words in them.



Hierarchical decoding (Joshua)

- Hierarchical models (Joshua) parse the input sentence and reorder nonterminals as required.
- Grammar extracted automatically from the parallel training corpus

SRC said dreamt of becoming a dentist .
REF सईद ने दंत चिकित्सक बनने का सपना देखा था .
OUT दांत के डॉक्टर बनने के सपने में कहा था .



Training Data	Joshua	Moses
Tides	12.27±0.83	11.46±0.72
Tides+DP	**12.58±0.77**	11.93±0.75
Tides+DP+Emille	11.32±0.74	10.06±0.72
Tides+DP+Dict	12.43±0.79	11.90±0.78

Manual Quality Judgements

- Three independent samples (100 sents., 100 sents., 53 sents.).
- The annotator sees the input sentence and several hypotheses.
- The reference translation is shuffled among the hypotheses.

SRC the private sector units are thirty to forty years old .
तीस हैं निजी आधार पर क्षेत्र यूनिट्स को को 40 वर्ष की आयु से ऊपर हैं |
* निजी क्षेत्र के स्थान पर हैं , 30 से 40 वर्ष से अधिक आयु के थे .
** निजी क्षेत्र की इकाइयां 30 से 40 वर्ष तक पुरानी हैं .
** निजी क्षेत्र की इकाइयों में 30 से 40 साल पुरानी हैं .

- Flags used:
empty incomprehensible,
* related to input, partial translation of phrases,
** acceptable and preserving most of the meaning, possibly still with many errors.

- Different from Ramanathan et al. (2009) who claim to have improved on average:
  - from little meaning conveyed, dysfluent Hindi, most phrases correct, ungrammatical overall
  - to much of meaning conveyed, non-native Hindi, few minor grammatical errors

Morphology and more data for Moses

System	0	*	**	BLEU
REF	6	11	83	—
OOD	80	17	3	1.85±0.24
TIDP	26	44	30	11.93±0.75
WC10	38	46	16	11.76±0.74

OOD out of domain: trained on all except for Tides
TIDP Tides + Daniel Pipes, no morphology
WC10 Tides + 3-gram LM for automatic word classes (10 classes)

- Six (percent) of reference translations were not acceptable!
- Text domain very important, OOD training poor in terms of BLEU and manual evaluation.
- More data more important than treating morphology (TIDP>WC10).
- However, BLEU does not discriminate between TIDP and WC10.

Moses vs. Joshua

System	0	*	**	BLEU
REF	6	10	84	—
Joshua	32	37	31	12.58±0.77
Moses	35	35	30	11.93±0.75
Moses-DPipes+POStags	32	42	26	12.03±0.75

- Identical training conditions:
  - Data: Tides + Daniel Pipes
  - No morphology.
- Joshua insignificantly outperforms Moses (both BLEU and manual judgments).

- The second probe also indicates that more data are more important than morphology:
  - This time, automatic POS tags used, not word classes
  - The result is somewhat ambiguous: the number of both ** and empty decrease.

Impact of Emille training data on Moses

System	0	*	**	BLEU
REF	0	8	45	—
TI DP	20	14	19	11.89±0.76
TI DP EM	22	19	12	9.61±0.75
TI DP EM oth	17	25	11	10.97±0.79
TI DP EM oth DICTFilt	23	17	13	10.96±0.75
TI DP EM oth DICTFull	22	16	15	10.89±0.69

- BLEU almost matches manual judgements this time.
- The addition of Emille significantly decreases the quality.
- Other data slowly compensate for the loss.

A later analysis revealed that Emille overlaps with Tides development dataset => model overfitting.

The research has been supported by the grants GAAV ČR 1ET201120505 (Grant Agency of the Academy of Sciences of the Czech Republic), MSM0021620838 (Czech Ministry of Education), FP7-ICT-2007-3-231720 (EuroMatrix Plus), and GAUK 4307/2009 (Grant agency of the Charles University).