

Developing Language Resources For English/বাংলা Machine Translation

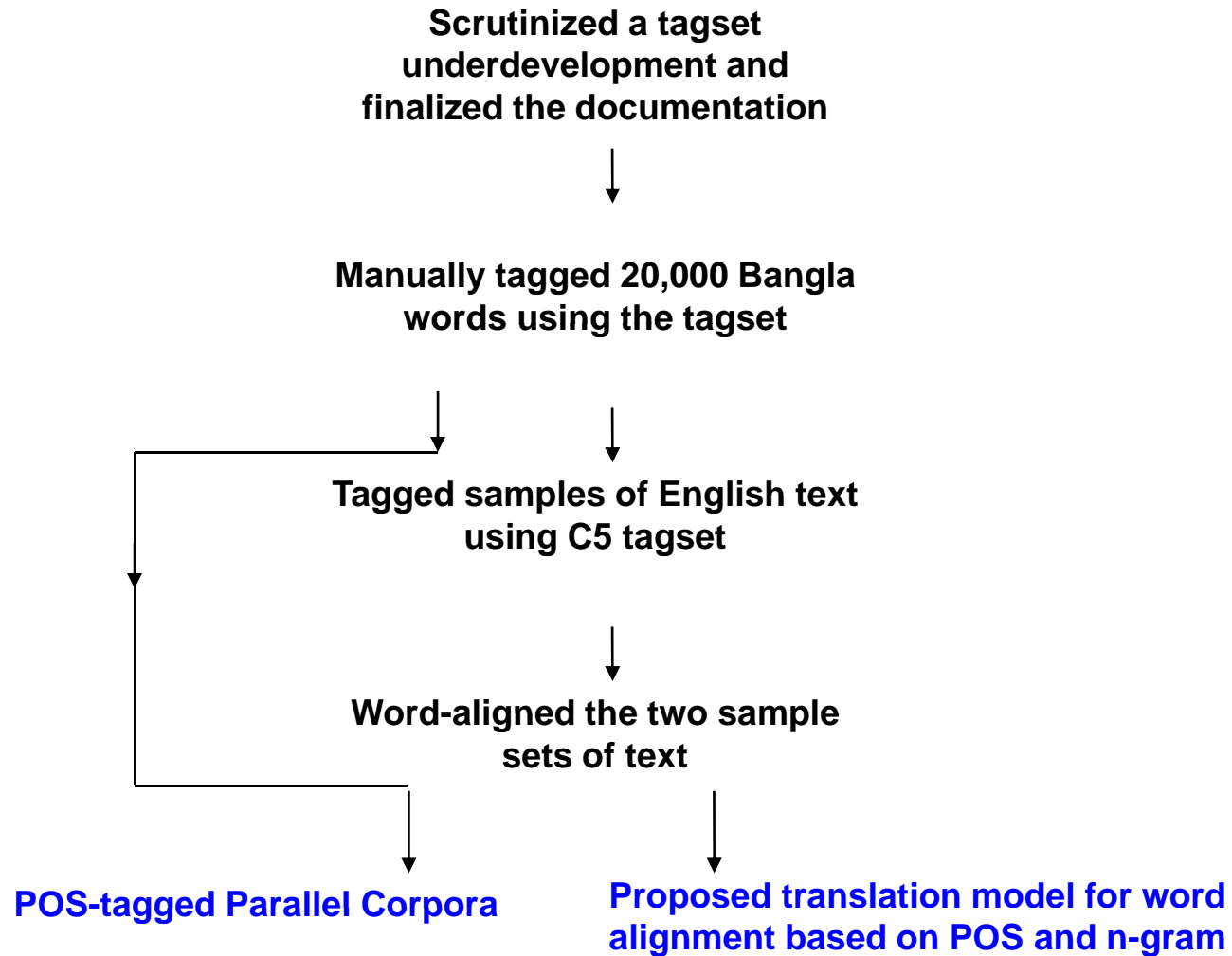
Presented by:

Rabia Sultana Ummi
Fahmina Huda

Some Definitions..

- **Machine Translation-** Used to translate text or speech from one natural language to another
- **Corpus-** Structured set of text (now electronic) used to perform statistical analysis, checking occurrences or validating linguistic rules
- **Parallel Corpora-** Multilingual corpora (same set of texts in different languages) specially formatted for side by side comparison
- **Statistical Machine Translation-** Generates translation using statistical methods based on bilingual corpora.

Annotated English/বাংলা Parallel Corpora



An Extract from the Developed Corpus

- Rupali is now undergoing treatment at Dhaka Medical College Hospital.
- Her sister and another worker suffered injuries in the accident and were sent to hospital.
- রূপালি বর্তমানে ঢাকা মেডিকেল কলেজ হাসপাতালে চিকিৎসাধীন রয়েছে।
- দুর্ঘটনায় তার বোন ও অপর এক শ্রমিক আহত হন এবং তাদের হাসপাতালে পাঠানো হয়েছে।

The Bilingual Corpus has been built using the Bangla and English versions of bdnews24.com

POS-Tagging the Text

55 Tags in the Bangla Tagset

রূপালি/NNP বর্তমানে/NNT+SFON ঢাকা/NNPC মেডিকেল/NNPC
কলেজ/NNPC হাসপাতালে/NNP+SFON চিকিৎসাধীন/JJ রয়েছে/VB ।/.

Level 1	Level 2	Tag	Examples
Noun	Proper	NNP	অক্টোবর, ঢাকা, রহমান
Noun	Temporal	NNT	গতকাল, আজ
Noun	Compound Proper Noun	NNPC	আব্দুর/NNPC রহমান/NNPC বিশ্বাসি/NNP
Adjective	Simple	JJ	সুন্দর, শ্রেষ্ঠ, দ্রুততম
Verb	Main Finite Verb	VB	করি, করলাম, করব
Suffixes	Adpositional	SFON	এ, য়, তে
Punctuation Marks	Sentence Final Punctuation	.	

PosAlign

- Jyun-Sheng Chang and Huey-Chyun Chen, Department of Computer Science, National Tsing Hua University
- To identify word correspondence in our parallel text, we suggest using PosAlign

The steps of alignment

- 1. Tag the parallel text with part-of-speeches
- 2. Initial alignment with the help of cognate lookup
- 3. Train the translation model iteratively using the unaligned part of the POS sequences

Advantage of using POS alignment - Parallel text of limited size is required for training

Using Corresponding English and Bangla POS-tagged Texts

- **The Bangla text was tagged manually**

আহতদের/JJ+SF\$ নাজিরহাট/NNPC স্বাস্থ্য/NNPC কেন্দ্রে/NNP+SFON
ভর্তি/NNV করা/VBM হয়েছে/VB ।/.

- **The English text was tagged using C5**

The/AT0 injured/AJ0 were/VBD admitted/VVN to/PRP Nazirhat/NP0
health/NN1 complex/NN1 ./.

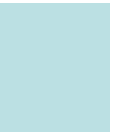
Initial Alignment using Cognate Lookup

The/AT0 injured/AJ0 were/VBD admitted/VVN to/PRP

Nazirhat/NP0 health/NN1 complex/NN1 (./.)

আহতদের/JJ+SF\$ নাজিরহাট/NNPC স্বাস্থ্য/NNPC

কেন্দ্রে/NNP+SFON ভর্তি/NNV করা/VBM হয়েছে/VB (./.)



Removing Suffix Tags

The/~~AT0~~ injured/~~AJ0~~ were/~~VBD~~ admitted/~~VVN~~ to/~~PRP~~

Nazirhat/~~NP0~~ health/~~NN1~~ complex/~~NN1~~ (./.)

{AT0, AJ0, VBD, VVN, PRP, NN1}

আহতদের/~~JJ+SF\$~~ নাজিরহাট/~~NNPC~~ স্বাস্থ্য/~~NNPC~~

কেন্দ্রে/~~NNP+SFON~~ ভর্তি/~~NNV~~ করা/~~VBM~~ হয়েছে/~~VB~~ (./.)

{JJ, NNPC, NNP, NNV, VBM, VB}

Alignment Using N-gram Language Model

The/AT0 injured/AJ0 were/VBD admitted/VVN to/PRP Nazirhat/NP0
health/NN1 complex/NN1 ./.

আহতদের/JJ নাজিরহাট/NNPC স্বাস্থ্য/NNPC কেন্দ্রে/NNP ভর্তি/NNV করা/VBM
হয়েছে/VB ।/.

English--{AT0, AJ0, VBD, VVN, PRP, NN1}

বাংলা --{JJ, NNPC, NNP, NNV, VBM, VB}

বাংলা/English	Conditional Probability	বাংলা/English	Conditional Probability
NNPC/NN1	0.40	NNPC/AJ0	0.40
NNP/NN1	1.00		
NNV,VBM,VB/VBD,VVN	1.00		
VBM,VB/VBD,VVN	1.00		
JJ/AT0,AJ0	1.00		
JJ/AJ0	0.67	JJ/NN1	0.33
Φ/AT0	0.33	Φ/PRP	0.50

Future Work

- 20, 000 word Bangla annotated corpus will be useful for future work in the field
- The bilingual English/বাংলা corpus can be used for Statistical Machine Translation
- Exact Translation in English and Bangla versions of bdnnews24.com is not available- Exact translation must be formed before use
- Suggestions and guidelines are available for those who wish to work in Sentence Alignment, Cognate Dictionary, etc.



Special Thanks to

- Professor Dr. Mumit Khan
- Altaf Mahmud
- Ibrar Ahmed