IIT Bombay's English-Indonesian submission at WAT: **Integrating neural language models with SMT**

Sandhya Singh, Anoop Kunchukuttan, Pushpak Bhattacharyya
{sandhya, anoopk, pb}@cse.iitb.ac.in

Center for Indian Language Technology
IIT Bombay

Motivation

- At CFILT, English-Indonesian language pair is being experimented as a part of a Project.

- Relatively new language pair among Asian language Translations.

About English-Indonesian Language pair

- Script is Latin for both English and Indonesian.
- Sentence structure followed is SVO (Subject Verb Object).
- Not much structural divergence between English and Indonesian.
- Indonesian is highly agglutinative and morphologically rich as compared to English language.
- Indonesian is considered as resource poor language.

Experiment Description (1/4)

Four different systems were trained for both directions of language pair:

1. **Phrase Based SMT system (*Moses* baseline)**
   - *MGIZA++ for word alignment*
   - *grow-diag-final-end* heuristic
   - Lexicalized Reordering
   - Batch MIRA tuning
   - 5-gram LM with Kneser-Ney smoothing using SRILM

- **Data Statistics**

Language	Training Set	Tuning Set	Test Set	For LM
English	44939 sentences	400 sentences	400 sentences	50000 sentences
Indonesian	44939 sentences	400 sentences	400 sentences	50000 sentences

Experiment Description (2/4)

2. **System using Neural Language Model as a feature for translation(NPLM)**
   - Neural Language model with default NPLM settings (Vaswani et al. (2013))
   - Word embedding size as 700, 750, 800 for 5 epochs
   - One hidden layer
   - Integrated as a feature in PBSMT system

- **Data statistics**

Language	Training Set	Tuning Set	Test Set	For LM
English	44939 sentences	400 sentences	400 sentences	50000 sentences + 2M sentences (Europarl)
Indonesian	44939 sentences	400 sentences	400 sentences	50000 sentences + 2M sentences (CommonCrawl)

Experiment Description (3/4)

3. **System using Bilingual Neural Language Model as a feature for translation(NNJM)**
   - Neural network joint LM with Parallel data (Devlin et al. (2014))
   - 5-gram LM with 9 source context word
   - One hidden layer
   - Integrated as a feature in PBSMT system

- **Data Statistics**

Language	Training Set	Tuning Set	Test Set	For LM
English	44939 sentences	400 sentences	400 sentences	50000 sentences
Indonesian	44939 sentences	400 sentences	400 sentences	50000 sentences

Experiment Description (4/4)

4. System using Operation Sequence Model for translation(OSM)

- Integrates 5-gram-based reordering and translation in a single generative process (Durrani et al. (2013))
- Deals with words along with context of source & target.

- **Data Statistics**

Language	Training Set	Tuning Set	Test Set	For LM
English	44939 sentences	400 sentences	400 sentences	50000 sentences
Indonesian	44939 sentences	400 sentences	400 sentences	50000 sentences

Evaluation Process

1. Automatic Evaluation metrics
    - BLEU points
    - RIBES Scores
    - AMFM Scores

2. Pairwise Crowdsourcing Evaluation
    - Against the shared task baseline

3. JPO Adequacy Evaluation
    - For content transmission

English-Indonesian MT system

Automatic Evaluation of English – Indonesian MT system

Approach Used	BLEU score	RIBES score	AMFM score
Phrase based SMT	21.74	0.804986	0.55095
Operation Sequence Model	21.70	0.806182	0.552480
Neural LM with OE = 700	22.12	0.804933	0.5528
Neural LM with OE =750	21.64	0.806033	0.555
Neural LM with OE = 800	22.08	0.806697	0.55188
Joint neural LM*	**22.35**	**0.808943**	**0.55597**

- **Increase in BLEU score with NNJM by 0.61 points over PBSMT system**

* WAT Submission, OE: Output Embedding

Pairwise Crowdsourcing Analysis of EI system(1/2)

Crowdsourcing Evaluation method—

- 5 Evaluators scored the sentence translations against the shared task baseline translation as :
  - Better than baseline : 1
  - Tie with baseline : 0
  - Worse than baseline : -1
- All 5 scores were added and converted to :
  - 1 if >= 2
  - -1 if <= -2
  - 0 if between 2 & -2

Pairwise Crowdsourcing Analysis of EI system(2/2)

- Scores received from pairwise evaluations

Experiment	Approach Followed	Better than Baseline	Comparable to Baseline	Worse than Baseline	Scores
English-Indonesian	NNJM	23%	44.75%	32.25%	-9.0250

- Observations
  - For worse sentences, sentence length is found to be >= 25 words.
  - Words not getting translated is the most visible error.

JPO Adequacy Scores of EI system

- Adequacy evaluation method –

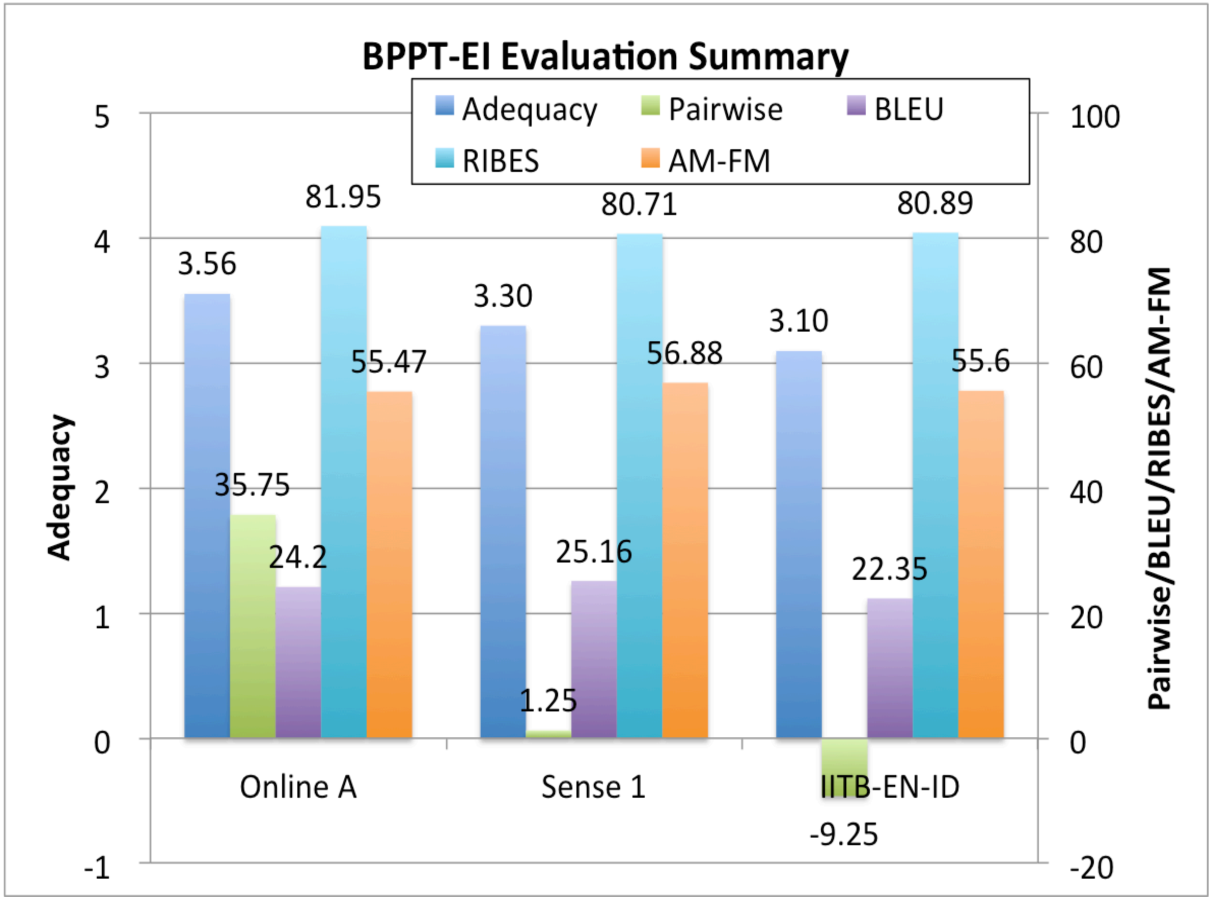
➢ 2 Annotators evaluated 200 translations for adequacy scores from 1 – 5

➢ Frequency of each score is used to compare.

- Scores :

Experiment	Approach Followed	Adequacy distribution					Adequacy Score
		5	4	3	2	1	
English-Indonesian	NNJM	17.75%	25.25%	23.25%	16.5%	17.25%	3.10

Summary of all evaluations for EI system (NNJM)



- Our systems adequacy scores suggests that the sentences are able to convey the meaning well.

Indonesian-English MT system

Results for Indonesian – English MT system

Approach Used	BLEU score	RIBES score	AMFM score
Phrase based SMT	22.03	0.78032	0.564580
Operation Sequence Model*	22.24	0.781430	0.566950
Neural LM with OE= 700	**22.58**	0.781983	**0.569330**
Neural LM with OE = 750	21.99	0.780901	0.56340
Neural LM with OE = 800	22.15	**0.782302**	0.566470
Joint Neural LM	22.05	0.781268	0.565860

- **Increase in BLEU score with NPLM by 0.55 points over PBSMT system**

* WAT Submission, OE: Output Embedding

Pairwise Crowdsourcing Analysis of IE system

- **Scores of crowdsourcing evaluation**

(refer to slide-11 for evaluation method)

Experiment	Approach Followed	Better than Baseline	Comparable to Baseline	Worse than Baseline	Scores
Indonesian-English	OSM approach	20%	34%	46%	-26.00

- **Observations**
  - For worse sentences, Sentence length is found to be >= 25 words

JPO Adequacy Scores of IE system

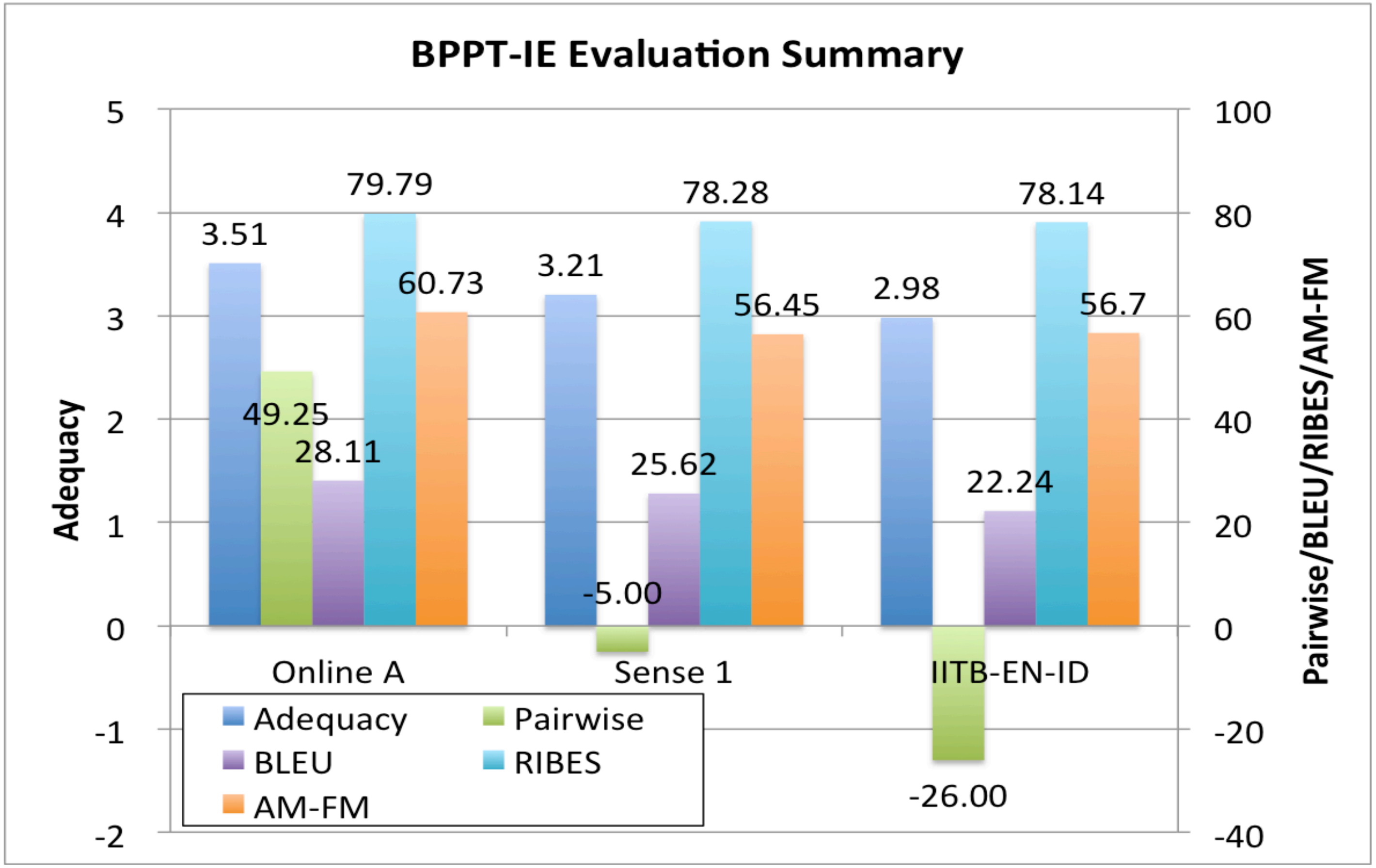
- **Scores** (refer to slide-13 for evaluation method):

Experiment	Approach Followed	Adequacy distribution					Adequacy Score
		5	4	3	2	1	
Indonesian-English	OSM approach	12%	18.75%	31.75%	30.5%	7%	2.98

- Observation:

-From adequacy distribution, it can be observed that > 50% of translations are adequate enough to convey the meaning.

Summary of all evaluations for Indonesian-English system(OSM)



- Our systems scores with OSM approach are not very promising against the baseline system.

Output Analysis of Indonesian-English System

Reference Sentence	Translated Sentence	Error Analysis
Moreover, syariah banking has yet to become a national agenda, Riawan said.	In addition, the banking industry had not so national agenda, said Riawan who also director of the main BMI.	Phrase insertion
Of course, we will adhere to the rules, Bimo said.	We will certainly patuhi regulations, Bimo said.	All words not translated
The Indonesian government last year canceled 11 foreign-funded projects across the country for various reasons, the Finance Ministry said.	The government has cancel foreign loans from various creditors to 11 projects in 2006 because various reasons.	Phrase dropped
As the second largest Islamic bank with a 29% market share of the Islamic banking industry's total assets at end-2007 albeit only 0.5% of overall banking industry's total assets, net financing margin NFM on Muamalat's financing operations increased to 7.9% in 2007 from 6.4% in 2004 due to better funding structure.	As the second largest bank of the market by 29 percent of the total assets syariah banking loans at the end of December 2007 although the market only 0.5 percent of the total assets banking industry as a whole, financing profit margin Muamalat rose to 7.9 percent in 2007 from 6.4 percent in 2004 thanks to funding structure.	Phrase dropped

* Text in blue represents error

Observations by Language Experts

Output analysis of Indonesian-English system

- The Sentences were adequate and fluent to some extent.
- The major error was of dropping and insertion of phrases.
- Some Indonesian words could not be translated to English due to lack of vocabulary learnt.
  - ➢Though OOV word percentage was found to be only 5% of the total words in the test set.
- Error in choice of function words used for English language.
  - ➢Require some linguistic insight on the Indonesian side of the language to understand the usage of function words in the source language.

Conclusion

- Due to structural similarity, translation outputs are adequate to understand.
- Integrating Neural Probabilistic LM (NPLM) with additional data as a feature in PBSMT system improves the translation quality.
- Integrating Neural Network Joint Model (Bilingual LM) trained on parallel data as a feature in PBSMT system improves translation quality.

Future Work

- Investigate the hyperparameters for the neural language model.
- Experiment with pure neural MT system for English-Indonesian language pair.

References

- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. "*Neural machine translation by jointly learning to align and translate*." In ICLR.
- Devlin, Jacob, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard M. Schwartz, and John Makhoul. 2014. "*Fast and Robust Neural Network Joint Models for Statistical Machine Translation."* In conference of the Association of Computational Linguistics.
- Durrani, Nadir, Helmut Schmid, and Alexander Fraser. 2011. *"A joint sequence translation model with integrated reordering."* Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics.
- Durrani, Nadir, Alexander M. Fraser, and Helmut Schmid. 2013. *"Model With Minimal Translation Units, But Decode With Phrases."* HLT-NAACL.
- Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan. 2007. *"Moses: Open source toolkit for statistical machine translation."* In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. Association for Computational Linguistics.
- Nakazawa, Toshiaki and Mino, Hideya and Ding, Chenchen and Goto, Isao and Neubig, Graham and Kurohashi, Sadao and Sumita, Eiichiro. 2016. *"Overview of the 3rd Workshop on Asian Translation."* Proceedings of the 3rd Workshop on Asian Translation (WAT2016), October.
- Niehues, Jan, Teresa Herrmann, Stephan Vogel, and Alex Waibel. 2011. *"Wider context by using bilingual language models in machine translation."* InProceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics.
- Vaswani, Ashish, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. *"Decoding with Large-Scale Neural Language Models Improves Translation."* In *EMNLP*.

Thank You!