Manipulating Input Data for Machine Translation

Liling Tan and Francis Bond
Universität des Saarlandes and Nanyang Technological University
alvations@gmail.com and bond@ieee.org

Introduction

- Data quality/quantity affects MT easily
- What happens if we try to *remove, add or rearrange* input data?

Approaches

- Context sensitive data selection
- Paraphrase data extension
- Lexicon addition
- Character level MT (JP-ZH)

Context Sensitive Data Selection

- Medical domain made up the majority of the ASPEC corpus (~30%)
- Build separate model for medical text
- Use generic model to decode the rest

Paraphrase Data Extension

- Generate EN paraphrases with ERG and ACE (Flickinger, 2000)
- Append EN paraphrases with original JP

EN Input: The particle sizes of the products decreased as the amount of seed increased.

Paraphrase: As the amount of seed increased, the particle sizes of the products decreased.

JP Input: 種の量が増加する程，生成物の粒子寸法は減少した。

Lexicon Addition

- Added JP-EN translation dictionary (JICST, 2004)

System Setup

- Moses phrase-based SMT (Koehn et al. 2007)
- GIZA++ IBM4 (Och and Ney, 2003)
- Bi-directional lex reordering (Koehn et al. 2005)
- KenLM 5grams (Heafield, 2011)
- Kneser-Ney smoothing (Kneser and Ney, 1995)
- MERT , truecasing

Character based MT

- Phrase-based SMT depends on tokens
- Diff segmenters produce diff tokens
- *What if we use characters for MT?* (Nakov and Tiedemann, 2012)

Input: 'これらカテゴリーに含まれる要素数を検討した'

MeCab: ['これら', 'カテゴリー', 'に', '含ま', 'れる', '要素', '数', 'を', '検討', 'し', 'た']

Juman: ['これ', 'ら', 'カテゴリー', 'に', '含ま', 'れる', '要素', '数', 'を', '検討', 'した']

KyTea: ['これ', 'ら', 'カテゴリー', 'に', '含', 'ま', 'れ', 'る', '要素', '数', 'を', '検討', 'し', 'た']

Char: ['こ', 'れ', 'ら', 'カ', 'テ', 'ゴ', 'リ', 'ー', 'に', '含', 'ま', 'れ', 'る', '要', '素', '数', 'を', '検', '討', 'し', 'た']

Results (JP-EN)

JP-EN

- Paraphrases and lexicon have minor improvement from baseline
- Human evaluation shows implementation or config. errors

JP-ZH

- Character based MT achieves pretty good results for JP-ZH

	BLEU	RIBES	Human
NAIST	35.8	0.811	56.25
WEBLIO-EJ1	33.4	0.795	43.25
Organizer	32.1	0.760	42.50
Kyoto-U	31.7	0.771	38.00
SAS_MT	31.4	0.771	27.50
Paraphrase	28.7	0.703	-
Baseline	28.6	0.703	3.75
Lexicon	28.1	0.693	-
Context	27.1	0.697	-

Table 1: EN-JP Results

	BLEU	RIBES	Human
NAIST	23.82	0.7236	40.50
Kyoto-U	21.07	0.701	25.00
Toshiba	20.61	0.707	23.25
Organizer	20.36	0.683	25.50
EIWA	19.86	0.706	22.50
Lexicon	18.91	0.646	-
Paraphrase	18.82	0.646	1.25
Baseline	18.57	0.640	-
Context	18.00	0.641	-
NII	17.47	0.630	-5.75
TMU	15.95	0.648	-17.00

Table 2: JP-EN Results

	BLEU	RIBES	Human
NAIST	40.15	0.845	50.75
SAS_MT	37.07	0.833	22.50
ORGANIZER	36.64	0.825	16.00
Kyoto-U	34.83	0.802	7.50
Character	34.64	0.784	-1.00
Baseline	33.46	0.771	-
EIWA	33.87	0.808	15.00

Table 3: ZH-JP Results

	BLEU	RIBES	Human
NAIST	30.53	0.8296	17.75
ORGANIZER	28.65	0.8091	14.00
NICT	27.98	0.8060	6.50
Baseline	27.92	0.7938	-
Kyoto-U	27.67	0.7964	-8.75
TOSHIBA	27.42	0.8044	0.75
BJTUNLP	24.12	0.7948	-3.75
Character	23.09	0.7794	10.00

Table 4: JP-ZH Results

Conclusion

- Data manipulation seems to affect BLEU score minimally
- More experiments necessary to improve system before conclusive results

References

- Francis Bond, Eric Nichols, Darren Scott Appling, and Michael Paul. 2008. Improving statistical machine translation by paraphrasing the training data. In IWSLT.
- Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. Natural Language Engineering.
- Chris Callison-burch and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In HLT/NAACL.
- Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In WMT.
- JICST, editor. 2004. JICST Japanese-English translation dictionaries. Japan Information Center of Science and Technology.
- Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In ICASSP-95.
- Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In NAACL
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In ACL.
- Lucian Vlad Lita, Abraham Ittycheriah, Salim Roukos and Nanda Kambhatla. 2003. tRuEcasIng. In ACL.
- Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In EMNLP.
- Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In ACL.

Acknowledgements

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n ° 317471.