;; NIST Open Machine Translation Evaluation (OpenMT15)
;; System Description
;;
;; Participants are to, as accurately as possible, fill in the sections
;; below and return to NIST by the submission deadline stated in the
;; evaluation plan.
;;
;; System descriptions should be submitted in PDF format.
;;
;; System descriptions must be submitted separately for each source
;; language and follow this naming convention:
;; <site>_<langpair>_sysdesc.pdf
;; where <site> identifies your affiliation and <langpair> identifies the language pair processed, e.g. NIST_ara2eng_sysdesc.txt.

NIST Open Machine Translation Evaluation (OpenMT15)
Participant System Description

1.0 Site affiliation

Institute of Computing Technology Chinese Academy of Sciences / Dublin University

2.0 Contact information

Jia Xu xujjia@gmail.com

3.0 Submissions

openmt15_eval_ict_chi2eng_cn_text_primary

4.0 Primary system specs

The primary system outputs above were generated by ICT's ensemble machine translation system, which is based on a number of novel and known system combination techniques and machine learning ensemble methods, including the recently developed design bagging algorithm.

4.1 Core MT engine algorithmic approach

- Translation systems:
We made use of various, distinct translation systems: Moses [10], Moses-Hiero [10], Moses-factor [10], Groundhog [11], Jane [12], Moses-OSM-OXLM [13],[17], CDEC [14], design-bagging [4] adapted to Moses, design-bagging adapted to CDEC, and design-bagging adapted to Jane.

- Tokenization:
We applied ICT-CLAS[1] and integrated Chinese word segmentation [2] for Chinese tokenization. We also applied Chinese monolingual spell checker, and OOV handling using Out-of-domain training corpus.

- Translation model:
Word alignments are generated based on GIZA++ [8] and mGIZA [9]. Training set include in-domain training data and selected out-of-domain training data based on the development corpus (Bolt Phase2/Phase3) similar to [18]. We automatically [18] selected 30K sentence pairs from the in-domain training corpus and applied the bilingual and sentence segmentation on these sentences [5].

- Language model:
The language model for each domain (SMS, CTS, Chat) is an interpolation of many language models [3]. It includes language models trained with different data: in-domain training data, development-data-selected out-of-domain training data, and English text for Arabic training data. We also incorporated language models learned with known and also newly developed approaches, including SRILM [10], RNNLM [16], and our design-bagging language model adapted on SRILM. The order of SRILM is 5-gram. We also used our newly developed phrase-based language model [6], [7] added in as additional features aiming to capture phrase-level dependencies.

- Tuning:
The tuning set is selected from the development set (Bolt Phase2/Phase3) with the same size as the eval set. We used MERT [10] for both single system tuning and for the system combination/ensemble methods.

- Domain adaptation:
We obtain three sets of training and tuning data set [3], one for domain SMS, one for Chat, and one for CTS. Each training set contains in-domain data and selected data from out-of-domain training set based on the development set (BOLT-phase2/phase3). The sentences close to the eval set are selected from the development set acting as a tuning set for each domain.

- Design bagging:
For Moses, Jane, and CDEC systems, the design bagging methodology [4] was applied with 35 bootstraps with each of the bootstrap 40% of the original training data.

- Post-processing:
We used the default true-casing tool of Moses, and used memory-based translation on nearly 200 sentences that are close to the Chinese sentences in the training set, where closeness is measured by BLEU.

4.2 Critical additional features and tools used

ICT-CLAS Chinese word segmentation [1]
Monolingual spell-checker in Chinese
Chinese-named entity transliteration

4.3 Significant data pre/post-Processing

OOV handling: we automatically find the translation of OOV words from the out-of-domain training corpus using word alignment information.

4.4 Other data used (outside the LDC training data)

No other data

5.0 Key differences in contrastive systems

Not applicable

6.0 SysCombo submissions

We applied MEMT [15] for system combination.

For the SMS domain we combined Moses-standard, Moses-factor, Moses-Hiero, Moses-OSM-OXLM, Groundhog, Jane, Design-bagging-for-Moses-standard, Design-bagging-for-CDEC, and Design-bagging-for-Jane.

For the Chat domain we combined Moses-standard, Moses-factor, Moses-Hiero, Moses-OSM-OXLM, Groundhog, Jane, Design-bagging-for-Moses-standard, and Design-bagging-for-CDEC.

For the CTS domain we combined Moses-standard, Moses-factor, Moses-Hiero, Moses-OSM-OXLM, Groundhog, and Jane.

7.0 References (if applicable)

[1] Zhang, Hua-Ping, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. "HHMM-based Chinese lexical analyzer ICTCLAS." In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pp. 184-187. Association for Computational Linguistics, 2003.

[2] Xu, Jia, Evgeny Matusov, Richard Zens, and Hermann Ney. "Integrated Chinese word segmentation in statistical machine translation." In *IWSLT*, pp. 131-137. 2005.

[3] Xu, Jia, Yonggang Deng, Yuqing Gao, and Hermann Ney. "Domain dependent statistical machine translation." In *MT Summit*. 2007.

[4] Papakonstantinou, Periklis A., Jia Xu, and Zhu Cao. "Bagging by Design (on the Suboptimality of Bagging)." *Twenty-Eighth AAAI Conference on Artificial Intelligence*. 2014.

[5] Xu, Jia, Richard Zens, and Hermann Ney. "Sentence segmentation using IBM word alignment model 1." *Proceedings of EAMT*. 2005.

[6] Chen, Geliang. "Phrase Based Language Model for Statistical Machine Translation - Empirical Study". Bachelor Thesis at Peking University. June, 2013.

[7] Xu, Jia, Geliang Chen. "Phrase Based Language Model for Statistical Machine Translation" http://arxiv.org/abs/1501.04324. 2015.

[8] Och, Franz Josef, and Hermann Ney. "Giza++: Training of statistical translation models." (2000).

[9] Gao, Qin, and Stephan Vogel. "Parallel implementations of word alignment tool." In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pp. 49-57. Association for Computational Linguistics, 2008.

[10] Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris

Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan et al. "Moses: Open source toolkit for statistical machine translation." In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pp. 177-180. Association for Computational Linguistics, 2007.
[11] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).
[12] Vilar, David, Daniel Stein, Matthias Huck, and Hermann Ney. "Jane: Open source hierarchical translation, extended with reordering and lexicon models." In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pp. 262-270. Association for Computational Linguistics, 2010.
[13] Paul, Baltescu, Blunsom Phil, and Hoang Hieu. "Oxlm: A neural language modelling framework for machine translation." *The Prague Bulletin of Mathematical Linguistics* 102, no. 1 (2014): 81-92.
[14] Dyer, Chris, Jonathan Weese, Hendra Setiawan, Adam Lopez, Ferhan Ture, Vladimir Eidelman, Juri Ganitkevitch, Phil Blunsom, and Philip Resnik. "cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models." In *Proceedings of the ACL 2010 System Demonstrations*, pp. 7-12. Association for Computational Linguistics, 2010.
[15] Heafield, Kenneth, and Alon Lavie. "Combining Machine Translation Output with Open Source: The Carnegie Mellon Multi-Engine Machine Translation Scheme." *The Prague Bulletin of Mathematical Linguistics* 93 (2010): 27-36.
[16] Mikolov, Tomas, Stefan Kombrink, Anoop Deoras, Lukar Burget, and Jan Cernocky. "RNNLM-Recurrent neural network language modeling toolkit." In *Proc. of the 2011 ASRU Workshop*, pp. 196-201. 2011.
[17] Durrani, Nadir, Helmut Schmid, and Alexander Fraser. "A joint sequence translation model with integrated reordering." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 1045-1054. Association for Computational Linguistics, 2011.
[18] Moore, Robert C., and William Lewis. "Intelligent selection of language model training data." In *Proceedings of the ACL 2010 Conference Short Papers*, pp. 220-224. Association for Computational Linguistics, 2010.