WAT 2016

# The 3rd Workshop on Asian Translation

**Proceedings of the Workshop** 

December 11-16, 2016 Osaka, Japan Copyright of each paper stays with the respective authors (or their employers).

ISBN978-4-87974-714-3

# Preface

Many Asian countries are rapidly growing these days and the importance of communicating and exchanging the information with these countries has intensified. To satisfy the demand for communication among these countries, machine translation technology is essential.

Machine translation technology has rapidly evolved recently and it is seeing practical use especially between European languages. However, the translation quality of Asian languages is not that high compared to that of European languages, and machine translation technology for these languages has not reached a stage of proliferation yet. This is not only due to the lack of the language resources for Asian languages but also due to the lack of techniques to correctly transfer the meaning of sentences from/to Asian languages. Consequently, a place for gathering and sharing the resources and knowledge about Asian language translation is necessary to enhance machine translation research for Asian languages.

The Workshop on Machine Translation (WMT), the world's largest machine translation workshop, mainly targets on European languages and does not include Asian languages. The International Workshop on Spoken Language Translation (IWSLT) has spoken language translation tasks for some Asian languages using TED talk data, but these is no task for written language.

The Workshop on Asian Translation (WAT) is an open machine translation evaluation campaign focusing on Asian languages. WAT gathers and shares the resources and knowledge of Asian language translation to understand the problems to be solved for the practical use of machine translation technologies among all Asian countries. WAT is unique in that it is an "open innovation platform": the test data is fixed and open, so participants can repeat evaluations on the same data and confirm changes in translation accuracy over time. WAT has no deadline for the automatic translation quality evaluation (continuous evaluation), so participants can submit translation results at any time.

Following the success of the previous WAT workshops (WAT2014, WAT2015), WAT2016 brings together machine translation researchers and users to try, evaluate, share and discuss brand-new ideas about machine translation. For the 3rd WAT, we proudly include new Asian languages: Hindi and Indonesian in addition to Japanese, Chinese and Korean for the machine translation evaluation shared tasks. We had 15 teams who submitted their translation results, and more than 500 submissions in total.

In addition to the shared tasks, WAT2016 also feature scientific papers on topics related to the machine translation, especially for Asian languages. The program committee accepted 7 papers that cover wide variety of topics such as neural machine translation, simultaneous interpretation, southeast Asian languages and so on.

We are indebted to Hideto Kazawa (Google) who gave an invited talk. We are grateful to "SunFlare Co., Ltd.", "TOIN Corporation", "Baobab, Inc". "Asia-Pacific Association for Machine Translation (AAMT)" and "PostEdit.Tokyo Co., Ltd." for partially sponsoring the workshop. We would like to thank all the authors who submitted papers. We express our deepest gratitude to the committee members for their timely reviews. We also thank the COLING 2016 organizers for their help with administrative matters.

WAT2016 Organizers

## Organisers

Toshiaki Nakazawa, Japan Science and Technology Agency (JST), Japan Hideya Mino, National Institute of Information and Communications Technology (NICT), Japan Chenchen Ding, National Institute of Information and Communications Technology (NICT), Japan Isao Goto, Japan Broadcasting Corporation (NHK), Japan Graham Neubig, Nara Institute of Science and Technology (NAIST), Japan Sadao Kurohashi, Kyoto University, Japan Ir. Hammam Riza, Agency for the Assessment and Application of Technology (BPPT), Indonesia Pushpak Bhattacharyya, Indian Institute of Technology Bombay (IIT), India

## **Programme Committee**

Rafael E. Banchs, Institute for Infocomm Research, Singapore Hailong Cao, Harbin Institute of Technology, China Michael Carl, Copenhagen Business School, Denmark Marine Carpuat, University of Maryland, USA Chenhui Chu, JST, Japan Fabien Cromières, JST, Japan Hideto Kazawa, Google, Japan Anoop Kunchookuttan, IIT Bombay, India Jong-Hyeok Lee, Pohang University of Science and Technology, Korea Gurpreet Singh Lehal, Punjabi University, Patiala, India Haizhou Li, Institute for Infocomm Research, Singapore Qun Liu, Dublin City University, Ireland Liling Tan, Universität des Saarlandes, Germany Masao Utiyama, NICT, Japan Andy Way, Dublin City University, Ireland Dekai Wu, HKUST, Hong Kong Devi Xiong, Soochow University, China Dongdong Zhang, Microsoft Research Asia, China Jiajun Zhang, Chinese Academy of Sciences, China

## **Technical Collaborators**

Luis Fernando D'Haro, Institute for Infocomm Research, Singapore Rafael E. Banchs, Institute for Infocomm Research, Singapore Haizhou Li, Institute for Infocomm Research, Singapore

## **Invited Speaker**

#### Hideto Kazawa, Senior Engineering Manager, Google, Japan

Google's Neural Machine Translation System: Training and Serving a Very Large Neural MT Models

## Abstract

Recently Neural Machine Translation (NMT) systems are reported to outperform other approaches in machine translation. However, NMT systems are known to be computationally expensive both in training and in translation inference – sometimes prohibitively so in the case of very large data sets and large models. Several authors have also charged that NMT systems lack robustness, particularly when input sentences contain rare words. These issues have hindered NMT's use in practical deployments and services, where both accuracy and speed are essential. In this talk, I present GNMT, Google's Neural Machine Translation system, which attempts to address many of these issues. Our model consists of a deep LSTM network with 8 encoder and 8 decoder layers using residual connections as well as attention connections from the decoder network to the encoder. To improve parallelism and therefore decrease training time, our attention mechanism connects the bottom layer of the decoder to the top layer of the encoder. To accelerate the final translation speed, we employ low-precision arithmetic during inference computations. To improve handling of rare words, we divide words into a limited set of common subword units ("wordpieces") for both input and output. On the WMT'14 English-to-French and Englishto-German benchmarks, GNMT achieves competitive results to state-of-the-art. Using a human sideby-side evaluation on a set of isolated simple sentences, it reduces translation errors by an average of 60phrase-based production system.

#### Short bio

Hideto Kazawa received M.Sc from University of Tokyo and Dr. Eng. from Nara Adavanced Institute of Science and Technology. He is now a Senior Engineering Manager of Google Translate team.

# **Table of Contents**

Overview of the 3rd Workshop on Asian Translation Toshiaki Nakazawa, Chenchen Ding, Hideya MINO, Isao Goto, Graham Neubig and Sadao Kuro-
hashi 1
Translation of Patent Sentences with a Large Vocabulary of Technical Terms Using Neural MachineTranslationZi Long, Takehito Utsuro, Tomoharu Mitsuhashi and Mikio Yamamoto
Japanese-English Machine Translation of Recipe Texts Takayuki Sato, Jun Harashima and Mamoru Komachi
<i>IIT Bombay's English-Indonesian submission at WAT: Integrating Neural Language Models with SMT</i> Sandhya Singh, Anoop Kunchukuttan and Pushpak Bhattacharyya
Domain Adaptation and Attention-Based Unknown Word Replacement in Chinese-to-Japanese Neural Machine Translation Kazuma Hashimoto, Akiko Eriguchi and Yoshimasa Tsuruoka
<i>Global Pre-ordering for Improving Sublanguage Translation</i> Masaru Fuji, Masao Utiyama, Eiichiro Sumita and Yuji Matsumoto
Neural Reordering Model Considering Phrase Translation and Word Alignment for Phrase-based Trans- lation
Shin Kanouchi, Katsuhito Sudoh and Mamoru Komachi
System Description of bjtu_nlp Neural Machine Translation System Shaotong Li, JinAn Xu, Yufeng Chen and Yujie Zhang
Translation systems and experimental results of the EHR group for WAT2016 tasks Terumasa Ehara
Lexicons and Minimum Risk Training for Neural Machine Translation: NAIST-CMU at WAT2016 Graham Neubig
NICT-2 Translation System for WAT2016: Applying Domain Adaptation to Phrase-based Statistical Ma- chine Translation
Kenji Imamura and Eiichiro Sumita 126
<i>Translation Using JAPIO Patent Corpora: JAPIO at WAT2016</i> Satoshi Kinoshita, Tadaaki Oshio, Tomoharu Mitsuhashi and Terumasa Ehara
An Efficient and Effective Online Sentence Segmenter for Simultaneous Interpretation Xiaolin Wang, Andrew Finch, Masao Utiyama and Eiichiro Sumita
Similar Southeast Asian Languages: Corpus-Based Case Study on Thai-Laotian and Malay-Indonesian Chenchen Ding, Masao Utiyama and Eiichiro Sumita
Integrating empty category detection into preordering Machine Translation Shunsuke Takeno, Masaaki Nagata and Kazuhide Yamamoto
<i>Kyoto University Participation to WAT 2016</i> Fabien Cromieres, Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi166

Character-based Decoding in Tree-to-Sequence Attention-based Neural Machine Translation Akiko Eriguchi, Kazuma Hashimoto and Yoshimasa Tsuruoka
Faster and Lighter Phrase-based Machine Translation Baseline   Liling Tan 184
Improving Patent Translation using Bilingual Term Extraction and Re-tokenization for Chinese–Japanese   Wei Yang and Yves Lepage 194
Controlling the Voice of a Sentence in Japanese-to-English Neural Machine Translation Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato and Mamoru Komachi
Chinese-to-Japanese Patent Machine Translation based on Syntactic Pre-ordering for WAT 2016 Katsuhito Sudoh and Masaaki Nagata
<i>IITP English-Hindi Machine Translation System at WAT 2016</i> Sukanta Sen, Debajyoty Banik, Asif Ekbal and Pushpak Bhattacharyya
Residual Stacking of RNNs for Neural Machine Translation   Raphael Shu and Akiva Miura 223

# **Conference Program**

## December 12, 2016

## 9:00–9:25 Welcome and overview of WAT2016

Overview of the 3rd Workshop on Asian Translation Toshiaki Nakazawa, Chenchen Ding, Hideya MINO, Isao Goto, Graham Neubig and Sadao Kurohashi

## 9:25–10:05 Research paper I

Translation of Patent Sentences with a Large Vocabulary of Technical Terms Using Neural Machine Translation Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi and Mikio Yamamoto

Japanese-English Machine Translation of Recipe Texts Takayuki Sato, Jun Harashima and Mamoru Komachi

10:05-10:20 Break

## 10:20–10:50 System description I

IIT Bombay's English-Indonesian submission at WAT: Integrating Neural Language Models with SMT Sandhya Singh, Anoop Kunchukuttan and Pushpak Bhattacharyya

Domain Adaptation and Attention-Based Unknown Word Replacement in Chineseto-Japanese Neural Machine Translation Kazuma Hashimoto, Akiko Eriguchi and Yoshimasa Tsuruoka

## December 12, 2016 (continued)

#### **10:50–12:00** Poster presentation I (Research paper)

*Global Pre-ordering for Improving Sublanguage Translation* Masaru Fuji, Masao Utiyama, Eiichiro Sumita and Yuji Matsumoto

Neural Reordering Model Considering Phrase Translation and Word Alignment for Phrase-based Translation Shin Kanouchi, Katsuhito Sudoh and Mamoru Komachi

### 10:50–12:00 Poster presentation I (System description)

IIT Bombay's English-Indonesian submission at WAT: Integrating Neural Language Models with SMT Sandhya Singh, Anoop Kunchukuttan and Pushpak Bhattacharyya

Domain Adaptation and Attention-Based Unknown Word Replacement in Chineseto-Japanese Neural Machine Translation Kazuma Hashimoto, Akiko Eriguchi and Yoshimasa Tsuruoka

System Description of bjtu\_nlp Neural Machine Translation System Shaotong Li, JinAn Xu, Yufeng Chen and Yujie Zhang

*Translation systems and experimental results of the EHR group for WAT2016 tasks* Terumasa Ehara

Lexicons and Minimum Risk Training for Neural Machine Translation: NAIST-CMU at WAT2016 Graham Neubig

NICT-2 Translation System for WAT2016: Applying Domain Adaptation to Phrasebased Statistical Machine Translation Kenji Imamura and Eiichiro Sumita

*Translation Using JAPIO Patent Corpora: JAPIO at WAT2016* Satoshi Kinoshita, Tadaaki Oshio, Tomoharu Mitsuhashi and Terumasa Ehara

## December 12, 2016 (continued)

## 12:00–14:00 Lunch

## 14:00–14:45 Invited talk

Google's Neural Machine Translation System: Training and Serving a Very Large Neural MT Models Hideto Kazawa

## 14:45–15:45 Research paper II

An Efficient and Effective Online Sentence Segmenter for Simultaneous Interpretation Xiaolin Wang, Andrew Finch, Masao Utiyama and Eiichiro Sumita

Similar Southeast Asian Languages: Corpus-Based Case Study on Thai-Laotian and Malay-Indonesian Chenchen Ding, Masao Utiyama and Eiichiro Sumita

*Integrating empty category detection into preordering Machine Translation* Shunsuke Takeno, Masaaki Nagata and Kazuhide Yamamoto

## 15:45–16:00 System description II

*Kyoto University Participation to WAT 2016* Fabien Cromieres, Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi

## 16:00–16:05 Commemorative photo

## December 12, 2016 (continued)

## 16:05–17:00 Poster presentation II (System description)

*Kyoto University Participation to WAT 2016* Fabien Cromieres, Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi

Character-based Decoding in Tree-to-Sequence Attention-based Neural Machine Translation

Akiko Eriguchi, Kazuma Hashimoto and Yoshimasa Tsuruoka

*Faster and Lighter Phrase-based Machine Translation Baseline* Liling Tan

Improving Patent Translation using Bilingual Term Extraction and Re-tokenization for Chinese–Japanese Wei Yang and Yves Lepage

Controlling the Voice of a Sentence in Japanese-to-English Neural Machine Translation

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato and Mamoru Komachi

Chinese-to-Japanese Patent Machine Translation based on Syntactic Pre-ordering for WAT 2016 Katsuhito Sudoh and Masaaki Nagata

*IITP English-Hindi Machine Translation System at WAT 2016* Sukanta Sen, Debajyoty Banik, Asif Ekbal and Pushpak Bhattacharyya

*Residual Stacking of RNNs for Neural Machine Translation* Raphael Shu and Akiva Miura

17:00– Closing