

The 20th Annual Conference of the European Association for Machine Translation

29-31 May 2017, Prague, Czech Republic



Conference Booklet

User Studies and Project/Product Descriptions

Contents

| | |
|---|-----------|
| Contents | 1 |
| Foreword | 3 |
| Preface from the Programme Chair | 5 |
| Organizers | 6 |
| Program | 7 |
| Sponsors | 13 |
| Project and Product Papers | 18 |
| <i>Pierrette Bouillon, Paula Estrella, Roxana Lafuente and Sabrina Girletti</i> MTTT – Machine Translation Training Tool: A tool to teach MT, Evaluation and Post-editing | 18 |
| <i>Vincent Vandeghinste, Tom Vanallemeersch, Liesbeth Augustinus, Frank Van Eynde, Joris Pelemans, Lyan Verwimp, Patrick Wambacq, Geert Heyman, Marie-Francine Moens, Iulianna van der Lek-Ciudin, Frieda Steurs, Ayla Rigouts Terryn, Els Lefever, Lieve Macken, Sven Coppers, Jan Van Den Bergh, Kris Luyten and Karin Coninx</i> SCATE - Smart Computer-Aided Translation Environment - Year 3 | 19 |
| <i>Nadira Hofmann</i> TM & MT – a happy couple [...or how to calculate the potential benefit] | 20 |
| <i>Gary Evans, Alexander Ferrein and Winfried Kock</i> Towards Deploying CAT Tools in University Classes for Improving Foreign Language Acquisition | 21 |
| <i>Josep Crego, Guillaume Klein, Jean Senellart, Yoon Kim, Yuntian Deng and Alexander M. Rush</i> OpenNMT: An Open-source Toolkit for Neural Machine Translation | 22 |
| <i>Celia Rico</i> IN-MIGRA2-CM Why the Third Social Sector does Matter to MT | 23 |
| <i>Yu Gong and Demin Yan</i> A Tool Set to Integrate OpenNMT into Production Workflow | 24 |
| <i>Michael Gasser</i> Minimal Dependency Translation | 25 |
| <i>Rico Sennrich, Antonio Valerio Miceli Barone, Joss Moorkens, Sheila Castilho, Andy Way, Federico Gaspari, Valia Kordoni, Markus Egg, Maja Popovic, Yota Georgakopoulou, Maria Gialama and Menno van Zaanen</i> TraMOOC - Translation for Massive Open Online Courses: Recent Developments in Machine Translation | 27 |
| <i>Luchezar Jackov</i> SkyCode MT – a translation system using deep syntactic and semantic analysis | 28 |
| <i>Ulrich Germann</i> Progress in ModernMT, a New Open-Source Machine Translation Platform for the Translation Industry | 29 |
| <i>Julia Epiphantseva</i> PROMT Machine Translation for Amadeus Fare Quote Notes Translator | 30 |

| | |
|--|-----------|
| <i>Antonio Toral, Víctor Manuel Sánchez-Cartagena and Mikel Forcada</i> Final Results of Abu-MaTran (Automatic building of Machine Translation) | 31 |
| <i>Christian Federmann</i> Appraise on Azure: A cloud-based, multi-purpose evaluation framework | 32 |
| <i>Barry Haddow, Alex Fraser, Marion Weller, Alexandra Birch, Ondrej Bojar, Fabienne Braune, Colin Davenport, Matthias Huck, Michal Kaspar, Kvetoslava Kovarikova, Josef Plch, Anita Ramm, Juliane Ried, James Sheary, Ales Tamchyna, Dusan Varis and Phil Williams</i> HimL : Health in my Language | 33 |
| <i>Tewodros Gebreselassie and Michael Gasser</i> A translation-based approach to the learning of the morphology of an under-resourced language | 34 |
| User Track Papers | 36 |
| <i>Nadira Hofmann and Maryse Lèpan</i> MT in real-world practice: Challenges and solutions at Swiss Federal Railways | 36 |
| <i>Anne Beyer, Vivien Macketanz, Aljoscha Burchardt and Philip Williams</i> Can Out-of-the-box NMT Beat a Domain-trained Moses on Technical Data? | 41 |
| <i>Pierrette Bouillon, Johanna Gerlach, Hervé Spechbach, Nikos Tsourakis and Sonia Halimi</i> BabelDr vs Google Translate: a user study at Geneva University Hospitals (HUG) | 47 |
| <i>Rei Miyata and Atsushi Fujita</i> Dissecting Human Pre-Editing Toward Better Use of Off-the-Shelf Machine Translation Systems | 53 |
| <i>Joachim Van den Bogaert, Bram Vandewalle and Roko Mijic</i> Bootstrapping Quality Estimation in a live production environment | 59 |
| <i>Adrià Martín-Mor, Gökhan Dođru and Sergio Ortiz</i> MTradumàtica: Free Statistical Machine Translation Customisation for Translators | 65 |
| <i>Lucia Comparin and Sara Mendes</i> Using error annotation to evaluate machine translation and human post-editing in a business environment | 68 |
| <i>Dimitar Shterionov, Pat Nagle, Laura Casanellas, Riccardo Superbo and Tony O’Dowd</i> Empirical evaluation of NMT and PBSMT quality for large-scale translation production | 74 |
| <i>Pavel Levin, Nishikant Dhanuka and Maxim Khalilov</i> Pavel Levin, Nishikant Dhanuka and Maxim Khalilov Machine Translation at Booking.com: Journey and Lessons Learned Machine Translation at Booking.com: Journey and Lessons Learned | 80 |
| <i>Nicola Bertoldi, Roldano Cattoni, Mauro Cettolo, Mohammad Amin Farajian, Marcello Federico, Davide Caroselli, Luca Mastrostefano, Marco Trombetti, Ulrich Germann and David Madl</i> MMT: New Open Source MT for the Translation Industry | 86 |

Foreword from the president of the European Association for Machine Translation

Mikel L. Forcada

Departament de Llenguatges i Sistemes Informàtics,
Universitat d'Alacant, E-03690 Sant Vicent del Raspeig, Spain
mlf@ua.es

As president of the European Association for Machine Translation (EAMT), it is a great pleasure for me to write the foreword to the book of proceedings for the user and projects/products tracks of the 20th annual conference of the EAMT in Prague, the Czech Republic.

The EAMT started organizing annual workshops in 1996; later, these workshops became annual conferences, and were hosted all around Europe. Years ago, the venue was steadily moving from west to east: from Barcelona (2009) to Saint-Raphaël (2010) to Leuven (2011) to Trento (2012) to Dubrovnik (2014)—after skipping one year to host the successful world-wide MT Summit 2013 in Nice—but recently turned around to go west again at Antalya (2015), to go to Riga (2016) and now Prague (2017). Again, you have guessed: EAMT 2018, our 21th annual conference, will surely be west from Prague. It will be announced at EAMT 2017 shortly after I am writing these lines. Those who miss our conference, will find out by visiting our Association's website, EAMT.org.

By the way, if you have not done so yet, please consider joining the EAMT. Our membership rates are low, particularly for students, and have not increased since the EAMT's inception. You will benefit from discounts when attending not only our conferences, but also the conferences held by our partner associations the Asia-Pacific Association for Machine Translation (AAMT) and the Association for Machine Translation in the Americas (AMTA). You will also have an exclusive chance to benefit from funding for your activities related to machine translation. And perhaps you can get even more involved and participate in serving the European machine translation community by becoming a member of the Executive Committee of the EAMT.

But let me go back to EAMT 2017. As in previous conferences, it is great to see the strong programme put together by our programme chairs: Alexander Fraser, research track chair, and Kim Harris, user track chair. As in previous editions, there will also be a projects and products session which showcases the advance of machine translation in Europe. And, last but not least, I also feel very fortunate to have João Graça from Unbabel as our invited speaker.

EAMT 2017 would have never been possible without the generous offer to host and the hard work subsequently done by the local organizing committee at the well-known machine translation group of Charles University, headed by Jan Hajič and Ondřej Bojar. I warmly thank them all. One important part of their work

has been to put together this book of proceedings that you are reading now. Note that the research papers of EAMT 2017 have been published in a special issue of the *Prague Bulletin of Mathematical Linguistics (PBML 108)*.¹

It is also with great pleasure that I thank our sponsors: Memsource (gold sponsor), Star Group (silver sponsor), text&form (bronze sponsor), and Prompsit and Apertium (supporting sponsors).

Finally, I would like to thank EAMT 2017 attendees for coming to Prague. I hope the conference leads to new friendships and fruitful collaboration.

Mikel L. Forcada
EAMT President
May 2017

¹<http://ufal.mff.cuni.cz/pbml>

Preface from the Programme Chair

Kim Harris

text&form, Germany
kim.harris@textform.com

It is my pleasure to welcome you to the 20th annual conference of the European Association for Machine Translation (EAMT) in Prague, the Czech Republic. I have really enjoyed serving as user programme chair for the user track in this edition of the conference. The EAMT conference has become the most important event in Europe in the area of machine translation for researchers, users, professional translators, etc.

As in previous editions, the conference is organised around three different tracks: research, user and projects/products. The research track papers will appear in volume 108 of the *Prague Bulletin of Mathematical Linguistics*. The user track reports users' experiences with machine translation, in industry, government, NGOs, etc. The project and product track offers projects and products the opportunity to be presented to the wide audience of the conference.

This year we have received 14 submissions to the research track and 17 descriptions of projects and products. Each submission to the user track was peer reviewed by at least two independent members of the Programme Committee. In the user track 11 papers out of 14 (79%) were accepted for publication. Aside from regular papers from the three tracks, the programme of EAMT 2017 includes an invited talk by João Graça, João Graça, CTO and co-founder of Unbabel, on the topic of "How to combine AI with the crowd to scale professional-quality translation".

I would like to thank the user programme committee members, whose names are listed below, for their high quality reviews and recommendations. These have been very useful to make decisions. We would also like to thank all the authors for trying their best to incorporate the reviewers' suggestions when preparing the camera ready papers. For those papers that were not accepted, we hope that the reviewers' comments will be useful to improve them. Special thanks to Mikel L. Forcada, who took care of the projects and products track.

Kim Harris
text&form, Germany

Organizers

Organizers

Institute of Formal and Applied Linguistics (ÚFAL)
Computer Science School, Faculty of Mathematics and Physics
Charles University, Prague, Czech Republic
The European Association for Machine Translation (EAMT)

General chair

Mikel L. Forcada

Local organizers

Ondřej Bojar
Jan Hajič

Project and product descriptions reviewers

Mikel L. Forcada
Lucia Specia

User track reviewers

| | |
|----------------------|--------------------|
| Aljoscha Burchardt | Tatjana Gornostaja |
| Manuel Herranz | Marion Wittkowsky |
| Gema Ramírez-Sánchez | Lena Marg |
| Ventsislav Zhechev | Pedro Díez-Orzas |
| John Tinsley | Jost Zetsche |
| Arle Lommel | Marcus Danei |
| Anne Beyer | Jörgen Danielson |
| Maxim Khalilov | Christian Lieske |
| Thomas Senf | Yuqi Zhang |
| Niko Papula | Bruno Pouliquen |
| Olga Beregovaya | Tony O Dowd |
| John Moran | Jörg Porsiel |
| Declan Groves | |

Program

Sunday, 28th May

18:00-19:00 Registration

18:30-21:00 Opening reception

Monday, 29th May

08:00-10:00 Registration

10:00-10:30 Opening of the conference
Session chair: Mikel Forcada

10:30-11:00 Keynote speech by João Graça, CTO and co-founder of Unbabel (Lisboa, Portugal)
"How to combine AI with the crowd to scale professional-quality translation"
Session chair: Mikel Forcada

11:00-11:30 Coffee break

11:30-13:00 Research presentations
Session chair: Lucia Specia

1. Parnia Bahar, Tamer Alkhoul, Jan-Thorsten Peter, Christopher Jan-Steffen Brix, Hermann Ney.
Empirical Investigation of Optimization Algorithms in Neural Machine Translation
2. Jan-Thorsten Peter, Arne Nix, Hermann Ney.
Generating Alignments Using Target Foresight in Attention-Based Neural Machine Translation
3. Praveen Dakwale, Christof Monz.
Convolutional over Recurrent Encoder for Neural Machine Translation

13:00-14:30 Lunch

14:30-15:30 Research presentations
Session chair: Ulrich Germann

4. Franck Burlot, François Yvon.
Learning Morphological Normalization for Translation from and into Morphologically Rich Languages
5. Anita Ramm, Riccardo Superbo, Dimitar Shterionov, Tony O'Dowd,

Alexander Fraser.
Integration of a Multilingual Preordering Component into a Commercial
SMT Platform

15:30-16:00 Coffee break

User presentations
Session chair: Kim Harris

16:00-17:30

1. Anne Beyer, Vivien Macketanz, Aljoscha Burchardt and Philip Williams - Can Out-of-the-box NMT Beat a Domain-trained Moses on Technical Data?
 2. Joachim Van den Bogaert, Bram Vandewalle and Roko Mijic - Bootstrapping Quality Estimation in a live production environment
 3. Dimitar Shterionov, Pat Nagle, Laura Casanellas, Riccardo Superbo and Tony O'Dowd - Empirical evaluation of NMT and PBSMT quality for large-scale translation production.
-

18:30-19:30 Staropramen: Prague Brewery

Tuesday, 30th May

8:00-9:00 Registration

User presentation
Session chair: Andy Way

9:00-9:30

4. Pavel Levin, Nishikant Dhanuka and Maxim Khalilov - Machine Translation at Booking.com: Journey and Lessons Learned
-

9:30-9:40

MT Summit XVI (September 18-22, Nagoya, Japan) Hiromi Nakaiwa, AAMT president.

Poster booster: projects and products
Session chair: Mikel Forcada

9:40-10:15

1. Pierrette Bouillon, Paula Estrella, Roxana Lafuente and Sabrina Girletti. MTTT – Machine Translation Training Tool: A tool to teach MT, Evaluation and Post-editing
2. Vincent Vandeghinste, Tom Vanallemeersch, Liesbeth Augustinus, Frank Van Eynde, Joris Pelemans, Lyan Verwimp, Patrick Wambacq, Geert Heyman, Marie-Francine Moens, Iulianna van der Lek-Ciudin, Frieda Steurs, Ayla Rigouts Terryn, Els Lefever, Lieve Macken, Sven Coppers, Jan Van Den Bergh, Kris Luyten and Karin Coninx. SCATE - Smart Computer-Aided Translation Environment - Year 3
3. Nadira Hofmann. TM & MT – a happy couple [...or how to calculate the

potential benefit]

4. Gary Evans, Alexander Ferrein and Winfried Kock. Towards Deploying CAT Tools in University Classes for Improving Foreign Language Acquisition
5. Josep Crego, Guillaume Klein, Jean Senellart, Yoon Kim, Yuntian Deng and Alexander M. Rush. OpenNMT: An Open-source Toolkit for Neural Machine Translation
6. Celia Rico. IN-MIGRA2-CM. Why the Third Social Sector does Matter to MT
7. Yu Gong and Demin Yan. A Tool Set to Integrate OpenNMT into Production Workflow
8. Michael Gasser. Minimal Dependency Translation
9. Rico Sennrich, Antonio Valerio Miceli Barone, Joss Moorkens, Sheila Castilho, Andy Way, Federico Gaspari, Valia Kordoni, Markus Egg, Maja Popovic, Yota Georgakopoulou, Maria Gialama and Menno van Zaanen. TraMOOC - Translation for Massive Open Online Courses: Recent Developments in Machine Translation
10. Luchezar Jackov. SkyCode MT – a translation system using deep syntactic and semantic analysis
11. Ulrich Germann. Progress in ModernMT, a New Open-Source Machine Translation Platform for the Translation Industry
12. Julia Epiphantseva. PROMT Machine Translation for Amadeus Fare Quote Notes Translator
13. Antonio Toral, Víctor Manuel Sánchez-Cartagena and Mikel Forcada. Final Results of Abu-MaTran (Automatic building of Machine Translation)
14. Christian Federmann. Appraise on Azure: A cloud-based, multi-purpose evaluation framework
15. Barry Haddow, Alex Fraser, Marion Weller, Alexandra Birch, Ondrej Bojar, Fabienne Braune, Colin Davenport, Matthias Huck, Michal Kaspar, Kvetoslava Kovarikova, Josef Plich, Anita Ramm, Juliane Ried, James Sheary, Ales Tamchyna, Dusan Varis and Phil Williams. HimL : Health in my Language
16. Tewodros Gebreselassie and Michael Gasser. A translation-based approach to the learning of the morphology of an under-resourced language

Poster boaster: users

Session chair: Kim Harris

10:15-10:30

1. Nadira Hofmann and Maryse Lèpan. MT in real-world practice: Challenges and solutions at Swiss Federal Railways
2. Pierrette Bouillon, Johanna Gerlach, Hervé Spechbach, Nikos Tsourakis and Sonia Halimi. BabelDr vs Google Translate: a user study at Geneva University Hospitals (HUG)

3. Rei Miyata and Atsushi Fujita. Dissecting Human Pre-Editing Toward Better Use of Off-the-Shelf Machine Translation Systems
4. Adrià Martín-Mor, Gökhan Doğru and Sergio Ortiz. MTradumàtica: Free Statistical Machine Translation Customisation for Translators
5. Lucia Comparin and Sara Mendes. Using error annotation to evaluate machine translation and human post-editing in a business environment
6. Nicola Bertoldi, Roldano Cattoni, Mauro Cettolo, Mohammad Amin Farajian, Marcello Federico, Davide Caroselli, Luca Mastrostefano, Marco Trombetti, Ulrich Germann and David Madl. MMT: New Open Source MT for the Translation Industry

10:30-11:00 Coffee break

11:00-12:45 Poster session: users, projects and products
Session chairs: Kim Harris and Mikel Forcada

12:45-14:00 Lunch

Research presentations
Session chair: Matteo Negri

14:00-15:30 6. Pintu Lohar, Haithem Afli, Andy Way.
Maintaining Sentiment Polarity in Translation of User-Generated Content
7. Eva Martínez Garcia, Carles Creus, Cristina España-Bonet, Lluís Màrquez.
Using Word Embeddings to Enforce Document-Level Lexical Consistency in Machine Translation
8. Daniel Torregrosa, Juan Antonio Pérez-Ortiz, Mikel L. Forcada.
Comparative Human and Automatic Evaluation of Glass-Box and Black-Box Approaches to Interactive Translation Prediction

15:30-16:00 Coffee break

16:00-16:30 Best thesis award
Session chair: Lucia Specia

16:30-17:30 EAMT general assembly

19:00-23:00 Banquet

Wednesday, 31st May (half day)

08:00-09:00 Registration

Research presentations
Session chair: Maja Popovic

- 09:00-10:00** 9. Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, Andy Way.
Is Neural Machine Translation the New State of the Art?
10. Filip Klubička, Antonio Toral, Víctor M. Sánchez-Cartagena.
Fine-Grained Human Evaluation of Neural Versus Phrase-Based Machine Translation
-

Poster booster: research
Session chair: Alexander Fraser

1. Arda Tezcan, Véronique Hoste, Lieve Macken.
A Neural Network Architecture for Detecting Grammatical Errors in Statistical Machine Translation
2. Rei Miyata, Anthony Hartley, Kyo Kageura, Cécile Paris.
Evaluating the Usability of a Controlled Language Authoring Assistant
3. Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, Philip Williams.
A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines
4. Jinhua Du, Andy Way.
Pre-Reordering for Neural Machine Translation: Helpful or Harmful?
5. Mikel L. Forcada, Felipe Sánchez-Martínez, Miquel Esplà-Gomis, Lucia Specia.
10:00-10:45 Towards Optimizing MT for Post-Editing Effort: Can BLEU Still Be Useful?
6. Chiraag Lala, Pranava Madhyastha, Josiah Wang, Lucia Specia.
Unraveling the Contribution of Image Captioning and Neural Machine Translation for Multimodal Machine Translation
7. Maja Popović.
Comparing Language Related Issues for NMT and PBMT between German and English
8. Francis M. Tyers, Hèctor Alòs i Font, Gianfranco Fronteddu, Adrià Martín-Mor.
Rule-Based Machine Translation for the Italian–Sardinian Language Pair
9. Marco Turchi, Matteo Negri, M. Amin Farajian, Marcello Federico.
Continuous Learning from Human Post-Edits for Neural Machine Translation
10. Alberto Poncelas, Gideon Maillette de Buy Wenniger, Andy Way.
Applying N-gram Alignment Entropy to Improve Feature Decay Algorithms
11. Nasser Zalmout, Nizar Habash.
Optimizing Tokenization Choice for Machine Translation across Multiple

Target Languages

12. Peyman Passban, Qun Liu, Andy Way.

Providing Morphological Information for SMT Using Neural Networks

13. Álvaro Peris, Mara Chinea-Ríos, Francisco Casacuberta.

Neural Networks Classifier for Data Selection in Statistical Machine Translation

14. Miguel Domingo, Mara Chinea-Rios, Francisco Casacuberta.

Historical Documents Modernization

15. Eleftherios Avramidis.

Comparative Quality Estimation for Machine Translation Observations on Machine Learning and Features

16. Vinit Ravishankar.

Finite-State Back-Transliteration for Marathi

17. Duygu Ataman, Matteo Negri, Marco Turchi, Marcello Federico.

Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English

18. Carla Parra Escartín, Hanna Béchara, Constantin Orăsan.

Questing for Quality Estimation A User Study

19. Ankit Srivastava, Georg Rehm, Felix Sasaki.

Improving Machine Translation through Linked Data

10:45-11:15 Coffee break

11:15-13:00 Poster session: research
Session chair: Alexander Fraser

13:00-13:15 Closing of the conference
Session chair: Mikel Forcada

14:30-18:30 Workshop: Social Media and User Generated Content Machine Translation

Sponsors

We are very grateful to our sponsors for their support:

Gold sponsor



Memsource

Silver sponsor



Star Group

Bronze sponsor



text & form

Also sponsored by



Charles University



Apertium



Prompsit



The FREE
Memsourse Academic Edition

Including project management

Used by academic institutions worldwide



Univerzita Palackého
v Olomouci



Aston University
Birmingham



Introducing the Memsourse Academic Edition

About 100 academic institutions all over the world already use the Memsourse Academic Edition. Why? To train their students to become proficient CAT tool users and make them competitive on the translation market.

Why do universities choose Memsourse?



Full functionality

Including project management features



It's free

Full functionality at zero cost



Easy setup

No installation, no paperwork




Unlimited users

A big department? No problem!

Get your free Academic Edition today!

Contact us at academic@memsource.com



text & form

Thinking about thinking machines?

TEXT&FORM HELPS YOU MAKE THE MOST OF
YOUR TRANSLATION AUTOMATION.

Integrated language solutions in any language you need.
www.textform.com

MTTT – Machine Translation Training Tool: A tool to teach MT, Evaluation and Post-editing

Pierrette Bouillon, Sabrina Girletti

University of Geneva, FTI/TIM
Boulevard du Pont-d'Arve 40
CH - 1211 Genève 4, Switzerland

Pierrette.Bouillon@unige.ch

Sabrina.Girletti@unige.ch

Paula Estrella, Roxana Lafuente

University of Córdoba, FaMAF/NLP
Medina Allende s/n
5000, Córdoba, Argentina

pestrella@famaf.unc.edu.ar

roxana.lafuente@gmail.com

Abstract

MTTT is an open-source tool conceived to help students and non-savvy users get started with the core technologies involved in a classical workflow of MT+PE without having to deal with the purely technical aspects of installing, training and evaluating MT models. In that sense, this tool is a graphical user interface abstracting the underlying command; it also provides post-editing functionalities, which would be the final stage in the workflow. MTTT is available at <http://pln.famaf.unc.edu.ar/?q=node/6>.

1 Description

The translation industry has widely accepted the so-called MT+PE or PEMT workflow, which involves machine translation and post-editing to deliver translations. Accordingly, many institutions have incorporated these topics in courses at different levels (MA, BA) and in different disciplines that could be involved in the process of developing MT or applying PE (Gaspari et al., 2015; Kenny & Doherty, 2014; O'Brien, 2002). In order to avoid any bias due to the use of a particular commercial software for the practical exercises, we have explored the use of open-source solutions. However, despite the many open-source tools available for MT, evaluation and PE, it is difficult to carry out practical exercises on these topics because not all of them provide graphical user interfaces (GUI), highly convenient for non-technical students, and more importantly none of them implements the whole

MTPE workflow. This has motivated the development of an open-source prototype, MTTT, conceived to help students and non-savvy users get started with the core technologies without having to deal with the purely technical aspects of installing, training and evaluating MT models, usually done through the command line. MTTT is a GUI that abstracts the commands needed to create statistical models using Moses (Koehn et al., 2007). It also provides functionalities to: (a) evaluate the models generated with standard automatic metrics; (b) post-edit machine translated text; and (c) generate basic statistics about post-editing productivity. Additionally, we are planning to extend its functionalities by allowing the user to access the resulting models to explore its contents and gain more insights about the internals of the PEMT process.

References

- Gaspari, F. Almaghout, H. and Doherty, S. 2015. A survey of machine translation competences: insights for translation technology educators and practitioners. *Perspectives*, 23(3), 333–358.
- Kenny, D. and Doherty, S. 2014. Statistical machine translation in the translation curriculum: overcoming obstacles and empowering translators. *The Interpreter and Translator Trainer*, 8(2), 276–294.
- Koehn, P. et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th ACL*. 177–180.
- O'Brien, S. 2002. Teaching post-editing: a proposal for course content. In *6th EAMT Workshop on Teaching Machine Translation*, 99–106.

SCATE – Smart Computer-Aided Translation Environment – Year 3 (/4)

Vincent Vandeghinste
Tom Vanallemeersch
Liesbeth Augustinus
Frank Van Eynde
Joris Pelemans
Lyan Verwimp
Patrick Wambacq
Geert Heyman
Marie-Francine Moens
Iulianna van der Lek-Ciudin
Frieda Steurs
University of Leuven
first.lastname@kuleuven.be

Ayla Rigouts Terryn
Els Lefever
Arda Tezcan
Lieve Macken
Ghent University
first.lastname@ugent.be
Sven Coppers
Jan Van den Bergh
Kris Luyten
Karin Coninx
UHasselt – tUL – EDM
first.lastname@uhasselt.be

Abstract

We aim to improve translators' efficiency through improvements in the technology. Funded by Flemish Government IWT-SBO, project No. 130041. <http://www.ccl.kuleuven.be/scate>

1 Tree-based MT and TM

We have aligned parse trees based on semantic predicates and roles, and building a tree-to-tree decoder for syntax-based SMT. We create parallel node-aligned treebanks and make them available online. We investigate different fuzzy matching metrics and how to integrate them with MT.

2 Detecting grammatical errors in SMT

As grammatical errors are the most frequent error types in MT output, we develop a methodology that detects grammatical errors in SMT output by using monolingual morpho-syntactic word representations in combination with surface and syntactic context windows.

3 Term Extraction from Comparable Corpora

Framing the induction of translations as a classification problem, we learn from a seed dictionary what word pairs are translations. We combine word and character-level features and induce fea-

tures on character-level from training data. For evaluation we developed an annotation scheme with detailed guidelines, resulting in high inter-annotator agreement. In addition to monolingual annotations, we are also working on a bilingual gold standard, where terms are linked with their translations.

4 Post-Editing via ASR

We are investigating domain adaptation by boosting language model probabilities of domain-specific terminology. The terminology is inferred from the already corrected material, either directly by keeping a word cache or indirectly, by using word and/or topic similarity. In addition, the language model is enriched with character-level information which enables modeling out-of-vocabulary words, which are very common in new domains.

5 Intelligent Translator Interfaces

A thorough redesign of translator interfaces has been established, integrating the different types of MT and TM, term corpora and consistency checks in such a way translators can minimize focus shifts and optimize usage of these tools. We included support for multiple translators working on different pieces of the same text and personalized workflows as part of the online translator interface.

6 Integration

We have built a demo system which combines the different research aspects into one demo, and are working with translators to collect feedback on the interface.

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

TM & MT – a happy couple ...or how to calculate the potential benefit

Nadira Hofmann
STAR Group
Wiesholz 35, 8262 Ramsen
Switzerland
nadira.hofmann@star-group.net

Abstract

More and more customers with an established translation process are planning to use a machine translation (MT) system to derive further benefit from their extensive translation memory (TM) and validated terminology. Before potentially introducing an MT system, questions are raised regarding the added value and quality such a solution can deliver in a professional translation environment – combining a translation memory system (TMS) with an MT system. STAR has developed a **three-phase proof of concept** that can answer these questions. This service provides customers with conclusive statistics and a solid decision-making process that are based on “real-life” projects.

1 Phase 1: Engine training and initial analysis with real jobs

At the beginning of Phase 1, STAR sets up a machine translation (MT) system that trains MT engines using customer-specific translation memory (TM) and terminology only, thereby guaranteeing that translation results are consistent in terms of style and terminology.

STAR then does an initial analysis with real jobs from previous months that have been translated using a translation memory system (TMS), e.g. Transit, but without MT support. These jobs are translated again (except for 100% matches) using the trained MT engines. Then, each MT translation is compared with the existing human translation using Transit’s fuzzy algorithm. This way the MT results can be

mapped into the fuzzy ranges. The resulting statistical overview gives the customer a precise impression of the following: 1) How many MT suggestions would the translators have been able to additionally benefit from – instead of having a lower quality fuzzy match or none at all? 2) Which language directions and domains are suitable for being processed with MT?

2 Phase 2: Pilot phase in the live process

Phase 2 shows how those involved in the process handle MT and TM in practice. To answer this question under real-life conditions, the trained MT engines are integrated into the customer’s existing translation process. This is done by one-off adjustment of the TM project templates or the parameters of the corporate language management (CLM) system. The project management workflow remains the same: The MT suggestions are requested during project import and are sent to the translators included in the project packages.

Translators do not need to take any additional steps: Transit e.g. displays the MT suggestions in the translation editor along with the fuzzy matches, but without indicating a quality score.

3 Phase 3: Productive analysis of the results from the pilot phase

For the productive analysis, the translation jobs that were processed in the pilot phase are analysed in the same way as the jobs in the initial analysis. But now, this analysis determines how the translators have actively benefitted from the MT suggestions.

It shows at a glance if the expectations raised by the initial analysis have been met, as well as reliably indicating what needs to be adjusted and optimised before the MT solution goes live.

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

FOUNDCAT: Towards Deploying CAT Tools in University Classes for Improving Foreign Language Acquisition

Gary Evans and Alexander Ferrein and Winfried Kock

FH Aachen University of Applied Sciences

Aachen, Germany

{gary.evans, ferrein, kock}@fh-aachen.de

Abstract

The FOUNDCAT project (Free, Open UNiversity Development using Computer-Aided Translations) aims at integrating state-of-the-art CAT tools into learning management software platforms such as Moodle for teaching German undergraduate students to broaden their English language skills.

The FOUNDCAT project has recently received funding from the German "Stifterverband" for the development of software and teaching concepts as part of "Fellowship für Innovationen in der digitalen Hochschullehre" (Fellowship for innovations in digital university teaching).¹ The project began work in March 2017 and aims to be completed by the end of the winter semester 2017/18.

With the advent of massive open online courses (MOOCs) and flipped classroom concepts, teachers are becoming aware that eLearning has much greater potential than just providing a collection of PDF documents, or videos on a download server. Computer-aided translation (CAT) tools can be successfully applied in a number of teaching activities.

Our objective is not to teach language students to become proficient in using computer-aided translation tools. We are primarily teaching German undergraduate students to broaden their English language skills. To help students memorize technical terms more easily and also enhance their language proficiency in general, we have been deploying CAT tools in our English

classes, resulting in positive responses from participants. While translating into a second language is unusual in the translation world, it has proven to be educational when learning a second language. Duolingo Immersions (which is no longer available online) utilised this method as part of language acquisition. The focus is not so much on the product (the translation), but rather the process of translating. Peer and client (teacher) reviews offer the opportunity to analyse translated segments and provide feedback in the form of comments and tracked changes to help improve L2 language proficiency.

Numerous open source CAT tools are available (e.g. OmegaT, Pootle, Weblate etc.). The main elements of CAT tools include term bases, machine translations and translation memories. Students either create translatable content themselves or select open source content (e.g. Wikipedia, or FH Aachen content). Segments are then chosen by students, translated and then peer reviewed in an iterative process resulting in translations for further analysis. The ability to comment on segments allows students to flag errors and target specific areas for improvement, hence individualising students' needs in a scalable learning environment. The inclusion of gaming elements (peer grading, levels, badges etc.) aims to add to student motivation.

We aim to extend our LMS-based courses so that FOUNDCAT can be embedded into the online LMS course. We are currently evaluating Weblate for suitability. We will assess the usability for our purpose of teaching English and evaluate how such tools can best be integrated into LMS platforms for language acquisition in general.

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

¹<https://www.stifterverband.org/lehrfellows/2016/ferrein>

OpenNMT: Open-source Toolkit for Neural Machine Translation

Guillaume Klein[†], Yoon Kim^{*}, Yuntian Deng^{*}, Josep Crego[†]
Jean Senellart[†], Alexander M. Rush^{*}
Harvard University^{*}, SYSTRAN[†]

Abstract

We introduce an open-source toolkit for neural machine translation (NMT) to support research into model architectures, feature representations, and source modalities, while maintaining competitive performance, modularity and reasonable training requirements.

1 Introduction

Neural machine translation has become a set of standardised approaches that has led to remarkable improvements, particularly in terms of human evaluation. It has now been successfully applied in production environment by major translation technology providers.

*OpenNMT*¹ is an open (MIT licensed) and joint initiative by SYSTRAN and the Harvard NLP group to develop a NMT toolkit for researchers and engineers to benchmark against, learn from, extend and build upon. It focuses on providing a production-grade system with an extensive set of model and training options to cover a large set of needs of academia and industry.

2 Description

OpenNMT implements the complete sequence-to-sequence approach that achieved state-of-the-art results in many tasks including machine translation. Based on the Torch framework, this model comes with many extensions that are known useful including multi-layer RNN, attention, bidirectional encoder, word features, input feeding, residual connections, beam search, and several others. The toolkit also provides various options to customize the training process depending on the task

and data with multi-GPU support, re-training, data sampling and learning rate decay strategies.

Toolkits like *Nematus*² or Google's *seq2seq*³ share similar goals and implementation but with frequent limitations on efficiency, tooling, features or documentation which *OpenNMT* tries to solve.

3 Ecosystem

More than the core project, *OpenNMT* aims to propose an ecosystem around NMT and sequence modelling. It comes with an optimised C++ inference engine based on the Eigen library to make deployment and integration of models easy and efficient. The library has also been used on multiple tasks, including image-to-text, speech-to-text and summarisation. We also provide recipes to automatise the training process, demo servers to quickly showcase results and a benchmark platform⁴ to compare approaches.

4 Community

OpenNMT is also a community⁵ providing various supports on using the project, addressing specific training processes and discussing the current and future state of neural machine translation research and development. The online forum counts more than 100 users and the project has been starred by over 1,000 users on GitHub.

5 Conclusion

We introduce *OpenNMT*, a research toolkit for neural MT that prioritises efficiency and modularity. We hope to maintain strong machine translation results at the research frontier, providing a stable framework for production use while enlarging an active and motivated community.

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

¹<http://opennmt.net>

²<https://github.com/rsennrich/nematus>

³<https://github.com/google/seq2seq>

⁴<http://nmt-benchmark.net/>

⁵<http://forum.opennmt.net/>

IN-MIGRA2-CM

Why the Third Social Sector does Matter to MT

Celia Rico

Facultad de Comunicación
Universidad Europea
C/ Tajo, s/n, 28670 Villaviciosa de Odón, Madrid
celia.rico@universidadeuropea.es

Abstract

Now that MT is increasingly used in multilingual contexts, contributing, from a market perspective, for speeding up processes, reducing costs and improving quality, it is interesting to note how the multilingual demands of the third social sector seem to have fallen into oblivion as far as this technology is concerned. In this regard, IM-MIGRA2-CM, as an interdisciplinary project, seeks, among other objectives, to cater for the multilingual needs of stakeholders working in not-for-profit contexts. One such need is the implementation of a customized MT engine following the trail of much more profitable sectors (automotive, travel or engineering, to name but a few).

1 Description

The third sector is a pillar that helps to build bridges between the state and the civil society by detecting social needs, providing a response, and developing frameworks for social participation, with high dependence on public funding and a workforce based on volunteer work. This compels actors involved to find new ways to respond to the demands of the millions of people at risk of poverty or other forms of social exclusion.

In this context, translation usually plays a key role, nonetheless usually carried out by voluntary contributions from professional translators who, altruistically, use their own resources to perform the job. In this scenario, work conducted in the area of translation technology in IN-MIGRA2-

CM concentrates in the implementation of MT as a means to help facilitate the work of the volunteer translator.

IN-MIGRA2-CM is as an interdisciplinary project that aims at carrying out a needs analysis of the third social sector, and more specifically, of migrant population in Spain, from different, yet complementary, perspectives: discourse analysis, language learning, sociolinguistics, and translation technology. The project is lead by Universidad de Alcalá de Henares and the consortium includes four research teams at Universidad Europea and Universidad Nebrija. The contribution from the team at Universidad Europea in the area of machine translation evolves around three main questions:

- How good are generic MT engines for translation in the third sector domain?
- Is domain adaptation of MT engines feasible in third sector translation?
- Is out-of-domain data useful in this context?

The research framework considers different translation engines (both rule-based and statistical machine translation), and different sets of training data (parallel corpus for general purposes, proprietary translation memories, and sample translations) with the purpose of carrying out a series of user experiments and evaluation. IN-MIGRA2-CM is still at a preliminary stage of work (year 1): setting up the evaluation context and methodology. Full results will not be available until the project ends (year 3).

Acknowledgements

IN-MIGRA2-CM is a project jointly funded by the Autonomous Community of Madrid (Spain) and the European Social Fund under grant H2015/HUM3404 (start date: 1 Jan 2016; end date: 31 Dec 2018).

A Toolset to Integrate OpenNMT into Production Workflow

Yu Gong

Product Globalization

VMware

Beijing, China

gongy@vmware.com

Demin Yan

Product Globalization

VMware

Palo Alto, USA

dyan@vmware.com

Abstract

In recent months, machine translation (MT) using deep learning has attracted attention for its improved quality over statistical MT. Harvard University and Systran introduced an open-source tool, OpenNMT, to the public for training neural machine translation models. OpenNMT is easy to use yet, there are still some limitations when applying it into an enterprise production environment.

In most enterprise production environments, output from the localization workflow is in Translation Memory eXchange (TMX) format. To feed this kind of human-translated parallel data into OpenNMT, users have to write their own tools or make use of some third-party tools to manipulate the data.

To quickly set up a workable machine translation engine with less cost and effort, we developed a toolset, called OMTS (OpenNMT Toolset) ^[1], to accelerate the process. OMTS contains two major features:

- TMX parsing and corpus cleaning;
- OpenNMT model training and controlling;
- RESTful APIs to call an OpenNMT model.

In the beginning, OMTS uses TMX file(s) as input, and then calls the corpus cleaning tool in m4loc (Moses for Localization)^[2] to generate clean and

tokenized corpus required by the pre-processing step in OpenNMT. A training job is automatically kicked off right after the corpus is ready to generate the final model.

OMTS evaluates the results by giving it a BLEU score. A dashboard gives the users a sense of how good the model is. Users also have an option to let OMTS automatically choose the best model (with the highest BLEU score). Finally, to integrate the model into localization workflow, a connector is required to link the model to the production environment. This connector is usually done by the localization management system (e.g. SDL WorldServer) provider and currently not in the scope of OMTS.

In conclusion, OMTS streamlines the process of creating workable NMT models by making use of the enterprise's own raw data and integrating it into the current localization workflow. With minimal effort, users are then able to set up their own OpenNMT systems.

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

[1] We're intended to get OMTS open source and it's currently in internal review process.

[2] <https://github.com/achimr/m4loc>

Minimal Dependency Translation: a Framework for CAT for Under-Resourced Languages

Michael Gasser
Indiana University, School of Informatics and Computing
Bloomington, Indiana, USA
gasser@indiana.edu

For under-resourced languages (URLs), the communities of speakers suffer from a lack of written material in their mother tongues. A partial solution to the problem is the translation of documents from other languages into the URLs. Computer-assisted translation (CAT) can speed up this process, but CAT systems require sizable translation memories, which are not available when one of the languages is under-resourced.

This paper describes an ongoing project to develop a lexical-grammatical framework for CAT with URLs as the target languages (TLs), relying on the grammatical resources and bilingual dictionaries that are available for many URLs. Called Minimal Dependency Translation (MDT), the framework is built on a lexicon of phrasal units called **groups**. Translation of a sentence results in an unordered set of translations of instantiated source-language (SL) groups.

Processing in MDT is illustrated below for the translation into Guarani of the Spanish sentence *no vamos a hablar con los maestros* ‘we aren’t going to speak with the teachers’ (1). The sentence is first subjected to POS tagging and morphological analysis, and a series of morphosyntactic transformation rules brings the input closer to TL structure (2). For example, the negator *no* and periphrastic future marker *vamos a* ‘we are going to’ are incorporated into the verb *hablar* ‘speak’, corresponding to Guarani morphology. Next the system searches for groups matching the input; three are shown (3). Two of these groups have heads that are lexemes rather than wordforms. For example, the group `<con $n>` matches sequences consisting of the

preposition *con* ‘with’ followed by any noun. Next, constraint satisfaction is used to find a set of groups that covers the input sentence. In this process, group instantiations may be **merged**; in the example, the `$n` element in `<con $n>` unifies with the head of `<maestro_n>` ‘teacher’ to form a single dependency structure. Next TL groups are accessed for each selected SL group (4). Cross-linguistic feature agreement constraints in the group entries are applied (for example, TL verbs agree with SL verbs on the negation feature), and merged groups are merged for the TL (5). Thus, the `$n` element in `<$n ndive>` ‘with \$n’, unifies with the head of `<mbo’ehára_n>` ‘teacher’. Finally, morphological generation is applied to the resulting TL lexemes and features (6). A single possible translation is shown for each SL phrase: *nañañe’ëmo’ãi* ‘we will not speak’, *mbo’eharakuéra ndive* ‘with teachers’.

- (1) No vamos a hablar con los maestros.
- (2) `hablar_v[t=fut,+neg,pn=1p]`
`con maestro_n[+pl]`
- (3) `<hablar_v>`, `<con $n>`, `<maestro_n>`
- (4) `<ñe’ë_v>`, `<$n ndive>`, `<mbo’ehara_n>`
- (5) `ñe’ë_v[t=fut,+neg,pn=1p]`,
`mbo’ehara_n[+pl] ndive`
- (6) *nañañe’ëmo’ãi*; *mbo’eharakuéra ndive*

The goals of the project are (1) the development of a set of open-source tools for creating MDT implementations and (2) two functioning MDT implementations, one for Spanish–Guarani

(<http://guarani.soic.indiana.edu/mainumby/>), the other for English–Amharic. The project began in 2016; following user testing in early 2018, the projected end date is late 2018. We are collaborating with the translation community in Paraguay through the Ateneo de Lengua y

Cultura Guaraní and with the IT PhD Program at Addis Ababa University.

Ongoing research is concerned with methods for handling ambiguity (SL morphology and syntax, group assignment during constraint satisfaction, group translation) and for extending and correcting the lexicon-grammar based on user feedback and the limited bilingual corpora that are available.

TraMOOC - Translation for Massive Open Online Courses: Recent Developments in Machine Translation

Rico Sennrich and Antonio Valerio Miceli Barone

University of Edinburgh

rico.sennrich@ed.ac.uk, amiceli@inf.ed.ac.uk

Joss Moorkens and Sheila Castilho and Andy Way and Federico Gaspari

ADAPT Centre

{joss.moorkens, sheila.castilho}@adaptcentre.ie, {away, fgaspari}@computing.dcu.ie

Valia Kordoni and Markus Egg and Maja Popovic

Humboldt-Universität zu Berlin

{evangelia.kordoni, markus.egg}@anglistik.hu-berlin.de, popovicm@hu-berlin.de

Yota Georgakopoulou and Maria Gialama

Deluxe Media Europe

{yota.georgakopoulou, maria.gialama}@bydeluxe.com

Menno van Zaanen

Tilburg University

mvzaanen@uvt.nl

Abstract

Massive open online courses have been growing rapidly in size and impact. TraMOOC¹ aims at developing high-quality translation of all types of text genre included in MOOCs from English into eleven European and BRIC languages that are hard to translate into and have weak MT support.

1 Recent developments

In TraMOOC, we have developed machine translation prototypes for 11 target languages, from English into German, Italian, Portuguese, Dutch, Bulgarian, Greek, Polish, Czech, Croatian, Russian, and Chinese. The translation systems are based on phrase-based SMT and neural machine translation. The latter has achieved state-of-the-art performance in recent evaluation campaigns (Bojar, 2016). We use the Nematus toolkit (Sennrich, 2017) for training; the translation server is based on the amuNMT toolkit (Junczys-Dowmunt et al., 2016). The translation systems have been adapted to MOOC texts via fine-tuning of the model parameters on in-domain training data to maximize translation quality on this domain.

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

¹TraMOOC is a H2020 Innovation Action project funded by the European Commission (H2020-ICT-2014-1-ICT-17-2014/644333) and runs from February 2015 to February 2018. For more details on the project, please, visit <http://www.tramooc.eu>

We have also completed a comparative human evaluation of phrase-based SMT and NMT for four language pairs to compare educational domain output from both systems using a variety of metrics. These include automatic evaluation, human rankings of adequacy and fluency, error-type markup, and technical and temporal post-editing effort. The results show a preference for NMT in side-by-side ranking for all language pairs, texts, and segment lengths. In addition, perceived fluency is improved and annotated errors are fewer in the NMT output. However, results are mixed for some error categories. Despite far fewer segments requiring post-editing, document-level post-editing performance was not found to have significantly improved when using NMT in this study, suggesting that NMT may not show an enormous improvement over SMT when used in a production scenario. We have subsequently prepared data and a slightly amended quality evaluation methodology to apply to all TraMOOC NMT systems later in 2017.

References

- Bojar, Ondřej et al. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.
- Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. In *Arxiv*.
- Sennrich, Rico et al. 2017. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain.

SkyCode MT – a translation system using deep syntactic and semantic analysis

Luchezar Jackov

SkyCode Ltd., Sofia, Bulgaria

PhD student at the Institute for Bulgarian Language,
Bulgarian Academy of Sciences

lucho@skycode.com

Abstract

SkyCode MT is a rule-based machine translation system that evaluates all possible parsing hypotheses and ranks them using dependency relations. It uses Princeton WordNet (PWN) (Fellbaum, 1998) synsets as universal dictionary and has separate per-language analysis and synthesis modules which enables translation between any two of the seven languages of the system. It has been developed as a complete solution used in commercial applications. The small footprint allows its use on mobile devices (smartphones and tablets). The system has participated as a translation vendor in the 7th FP project iTranslate 4 (<http://itranslate4.eu>).

1 System description

The system translates between English, German, French, Spanish, Italian, Turkish and Bulgarian by means of a deep internal syntactic and semantic representation of the input text. This allows the translation of the 21 language pairs (42 translation directions) in just 150 MB. The sense inventory is based on the original PWN synonym sets (concepts) extended with lexicalizations having the following synset coverage: 74124 in Bulgarian, 62015 in Turkish, 79553 in German, 84345 in Spanish, 88955 in French and 78718 in Italian.

The lexicalizations are used for morphological analysis of the source, creating initial hypotheses for simple concepts (the various readings of single words and collocations). The system uses manually defined rules to generate all possible

parses (parsing hypotheses) for the source by applying them in a bottom-up fashion on adjacent hypotheses, building an entire sentence parse tree. The rules are based on Chomsky-normal-form context-free grammar extended with dependency relations on the constituents. As a result each hypothesis identifies concepts (PWN synsets) and dependency relations between them. The relations between the concepts are used for evaluation of how 'sensible' the hypothesis is by consulting a relations knowledge base. It is defined on the PWN synsets and is language-independent for most of the relations.

The translation is synthesized using the PWN synset lexicalizations for the target language and manually defined synthesis rules, transferring the semantic relations to the translation.

Both the synthesis and the analysis rules are shared between languages that have common linguistic phenomena such as the same word order, e.g. $S \rightarrow NP VP$, $VP \rightarrow V NP$, $VP \rightarrow V PP$.

The use of PWN synsets as universal dictionary and knowledge base as well as splitting the analysis from the synthesis allow for the translation between the languages of the system without having to define per-language-pair rules. This also makes adding a new language relatively easy by only defining PWN lexicalizations, and analysis and synthesis rules specific to the new language.

The system is implemented in C++, which makes it portable across various operating systems and platforms including mobile devices. A detailed description is given in (Jackov, 2014).

References

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Jackov, L. 2014. *Machine translation based on WordNet and dependency relations*. In *Computer Linguistics In Bulgaria 2014*, p. 64–72.

Progress in ModernMT, a New Open-Source Machine Translation Platform for the Translation Industry

<http://www.modernmt.eu>

The ModernMT Consortium:

Translated srl, Rome, Italy

Fondazione Bruno Kessler (FBK), Povo, Italy

University of Edinburgh, Edinburgh, Scotland, Europe

Translation Automation User Society (TAUS), Amsterdam, Netherlands

Corresponding author: Ulrich Germann (ugermann@inf.ed.ac.uk)

Abstract

We report progress made in Year 2 of ModernMT, a three-year EU *Horizon 2020 Innovation Action* (2015–2017) that develops new open-source machine translation technology for use in translation production environments. ModernMT is designed to facilitate both fully automatic translation and interactive post-editing scenarios.

1 Project Goals

ModernMT aims to improve the state of the art in open-source machine translation technology by developing scalable, cloud-ready software that offers the following benefits.

- A **simple installation** procedure for turn-key RESTful¹ machine translation services.
- **Very fast set-up times** for systems built from scratch using existing parallel corpora (e.g., translation memories). Incoming data can be ingested at approximately the same speed at which it is uploaded.
- **Immediate integration of new data** (e.g., from newly post-edited MT output). Rebuilding or retuning the system will not be necessary.
- **Instant domain adaptation** by considering translation context beyond the individual sentence, without the need for domain-specific custom engines. The ModernMT system uses the translation input (from a single translation unit to an entire document), as well as additional context keywords (if provided by the user) to retrieve similar texts from its bitext database and to bias translations towards the style and lexical choice of these similar texts.

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

¹https://en.wikipedia.org/wiki/Representational_state_transfer

- **High scalability** with respect to throughput, concurrent users, and the amount of data the system can handle.

In addition, ModernMT is actively collecting, curating, cataloguing, and — where possible — releasing parallel data from web crawls and parallel data contributions from translation stakeholders, so that ModernMT users have access to data to build their own custom systems. Furthermore, additional data is being collected to set up a new MT service provider that offers high-quality MT services at an affordable price to MT users who prefer not to have to maintain their own systems.

2 Project Phases

The current roadmap of ModernMT can be described as follows.

Year 1 was dedicated to integrating existing statistical machine translation technology, mostly based on the *Moses* toolkit,² and prototyping of instant system adaptation and dynamic model updates.


Year 2 saw the development of a cloud-ready infrastructure and successful integration of adaptation and instant updates into the system. This included development of new database-backed back-ends for the language and translation models.

Year 3 will put focus on development of a ready-to-launch product and investigations into Deep Learning for use within the framework of ModernMT.

3 ModernMT is Open-source

The software is available at <https://github.com/ModernMT/MMT>.

Acknowledgements

 ModernMT has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 645487 (call ICT-17-2014).

²<http://www.statmt.org/moses>

PROMT Machine Translation for Amadeus Fare Quote Notes Translator

Julia Epiphantseva

PROMT LLC

17E Uralskaya str. building 3, 199155,

St. Petersburg, Russia

`Julia.Epiphantseva@prompt.com`

Abstract

This document provides an overview of the implementation of PROMT Cloud solution into the [Amadeus Translator](#) application in the Amadeus booking system for translating Fare Quote Notes (FQN) to optimize the process of airline tickets sale and improve the quality of service for travel agencies' clients. FQN contain the rules, regulations and conditions that apply to a specific fare. FQN are created automatically in English and have specific format and language features.

1 Challenge

The Amadeus booking system helps travel agencies to find and book tickets for domestic and international flights. Airline companies store the information about their vacant seats and terms and conditions of the flights in the Amadeus database. The travel agencies use this database for booking tickets and explaining terms and conditions of the flights to travelers. Since this information available in English only, the travel agency staff deals with important information in foreign language, which could lead to misunderstanding and wrong decisions. To address this challenge, Amadeus decided to provide travel agencies with custom machine translation. The MT solution should meet the following requirements: understandable translation conveying the meaning of FQN and taking into account terms and abbreviations; high performance and reliability; integration in the Amadeus system's interface.

2 Solution Overview

PROMT suggested its solution PROMT Cloud with a programming interface (API) ready to process a large number of translation requests. The provided solution consisted of two components: dedicated customized translation module and dedicated web service for easy integration of MT into the Amadeus booking system.

To achieve better translation quality the following algorithms and customization data were added to the translation system:

- Special algorithm of FQN preprocessing taking into account their format and structure (deletion of mid sentence line breaks was implemented);
- Additional dictionaries
~1,200 terms typical for FQN
~20,000 names of airline companies and airport codes;
- Translation memory with professional translations of 150 most frequent sentences (including titles).

3 Results

The implemented solution translates more than 6,000 translation requests per week. Each translation request consists of about 700 words, which generates more than 4,000,000 translated words per week. The new application provides customized machine translation of FQN from English into Russian. Due to the use of advanced technologies and customized settings, the professional terminology of travel industry is taken into account in the application. With its help travel agents are able to quickly and easily obtain necessary information and provide passengers with complete and up-to-date information on terms of travel and restrictions of the chosen fare.

Final Results of Abu-MaTran (Automatic building of Machine Translation)

| | | |
|---|---|---|
| Antonio Toral Faculty of Arts University of Groningen NL-9712 EK Groningen a.toral.ruiz@rug.nl | Víctor Sánchez-Cartagena Prompsit Language Engineering E-03202 Elx vmsanchez@prompsit.com | Mikel L. Forcada Dept. Lleng. i Sist. Inform. Universitat d'Alacant E-03071 St. Vicent del Raspeig mlf@ua.es |
|---|---|---|

Abstract

We present the final results of Abu-MaTran (<http://www.abumatran.eu>), a 4-year project (January 2013–December 2016) on rapid development of machine translation for under-resourced languages. It was funded under the Marie Curie's Industry-Academia Partnerships and Pathways 2012 programme. The Abu-MaTran consortium had 5 partners (4 academic and 1 industrial) in four different countries.

1 Introduction

Abu-MaTran sought to enhance industry-academia cooperation as a key aspect to tackle one of Europe's biggest challenges: multilingualism. We aimed to increase the hitherto low industrial adoption of machine translation (MT) by identifying crucial cutting-edge research techniques, making them suitable for commercial exploitation. We also aimed to transfer back to academia the know-how of industry to make research results more robust. We worked on a case study of strategic interest for Europe: MT for the language of a new member state (Croatian) and for related languages. All the resources produced have been released as free/open-source software, resulting in effective knowledge transfer beyond the funded period.

2 Results

At EAMT 2017 we will present a selection of the final results of the project, including the following:

- **Web crawling:** A novel pipeline to crawl massive amounts of parallel and monolingual

data from the Internet's top level domains that is ready for commercial exploitation.

- **Acquisition of language resources** (bilingual dictionaries and transfer rules): We have developed methodologies (i) to enable non-expert users to improve the coverage of morphological dictionaries and (ii) to learn automatically translation rules from very small parallel corpora.
- **Language models:** Implementation of a novel cloud-based language model that allows us to use effectively vast amounts of monolingual data in phrase-based statistical MT.
- **Linguistically-augmented approaches**, including morph-segmentation approaches, to phrase-based and neural MT.
- **Improved data selection** of training data for MT using linguistic information and quality estimation techniques.
- **Collaborative development of MT:** development of state-of-the-art rule-based MT between closely-related languages through a collaborative process.
- **Dissemination:** Workshops on (i) tools for teaching MT and on (ii) methodologies for rapid development of MT for under-resourced languages; and the establishment of a linguistics Olympiad in Spain.

All the tools and data sets developed within the project were released according to free/open-source licenses and can be found at the project's website.¹

¹<http://www.abumatran.eu/>

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

Appraise on Azure: A cloud-based, multi-purpose evaluation framework

Christian Federmann
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
chrife@microsoft.com

Abstract

We present Appraise on Azure, an extension to the Appraise evaluation framework which enables users to host evaluation campaigns in the Microsoft Azure cloud. This allows to scale annotation efforts on demand and makes it easier to set up and run multiple annotation tasks in parallel. Both the Appraise framework and the code adding Azure support are released under an open license. The demo will give details on the architecture of the system, discuss Azure integration and demonstrate annotation options in the framework.

1 Motivation

Over the last years, the Appraise evaluation toolkit has seen growing interest from industry and research community. This has resulted in a number of forks of the source code which is publicly available.¹ Work on the updated version presented in this demo was motivated by problems with the original version:

- **Outdated forks:** As the underlying software packages have been updated over time, the integration of external code is difficult. Potential security issues might remain unfixed.
- **No support for parallel campaigns:** The original version of Appraise was implemented as an open framework which allowed to build annotation platforms for specific scenarios. However, the framework lacked support to run multiple such collection efforts in parallel.

- **Too WMT centric:** Appraise is the platform for the yearly evaluation campaign conducted as part of the WMT Conference on Statistical Machine Translation. Hence, focus shifted from supporting a wide range of annotation tasks to only two: relative ranking and, later, direct assessment. Also, the software became harder to configure for non-WMT campaigns, adding unnecessary complexity to the setup.

2 Appraise on Azure Improvements

We have addressed the aforementioned issues with the old codebase and added support for the latest Django version 1.11. Also, we implemented support to host Appraise on Microsoft Azure. The Microsoft Azure cloud² is a collection of integrated cloud services, including hosting, storage and compute solutions as well as high-level APIs such as Cognitive Services. The simplified setup process makes it easier to focus on the creation of new annotation views. The same holds if a user only wants to run an evaluation campaign. The end user benefits in multiple ways: 1) all software dependencies have been updated. 2) it is now possible to configure multiple annotation campaigns in a single Appraise instance. 3) users can set up non-WMT tasks while still getting updated status views and other features introduced for WMT.

3 Conclusion

Our demo presents Appraise on Azure, a cloud-based, multi-purpose evaluation framework. We discuss the architecture, annotation views and give an outlook to future extensions. All features work both standalone and when hosting Appraise on Microsoft Azure.

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

¹ See <https://github.com/cfedermann/Appraise/>

² See <https://azure.microsoft.com/>

HimL: Health in my Language

Barry Haddow¹, Alexandra Birch¹, Ondřej Bojar², Fabienne Braune⁵,
Colin Davenport³, Alex Fraser⁵, Matthias Huck⁵, Michal Kašpar⁶,
Květoslava Kovaříková⁶, Josef Pích⁶, Anita Ramm⁵, Juliane Ried⁴,
James Sheary³, Aleš Tamchyna², Dušan Variš², Marion Weller⁵, Phil Williams¹

¹School of Informatics, University of Edinburgh, Scotland

² Faculty of Mathematics and Physics, Charles University in Prague, Czech Republic

³ NHS 24, Caledonia House, Glasgow, Scotland ⁴ Cochrane, Freiburg, Germany

⁵ LMU Munich, Germany ⁶ Lingea s.r.o., Brno, Czech Republic

Coordinator email: bhaddow@inf.ed.ac.uk

Abstract

HimL (www.himl.eu) is a three-year EU H2020 Innovation Action, which started in February 2015. Its aim is to increase the availability of public health information via automatic translation. Targeting languages of Central and Eastern Europe (Czech, German, Polish and Romanian) we aim to produce translations which are adapted to the health domain, semantically accurate and morphologically correct.

1 Description

In HimL we aim to deploy and evaluate machine translation systems for the public health domain, addressing domain adaptation, semantic accuracy and target morphology. The project is now in its third year, and we have made two releases of our translation systems and used them to translate the user partner websites. These have been subjected to automatic evaluation, human evaluation, and are undergoing user evaluation.

The HimL system releases so far were built as phrase-based MT systems using large, diverse training sets and applying language model and translation model interpolation to adapt to the medical domain. In Year 2, we applied the corrective approach to morphology to the English→Czech system, and the two-step approach to the English→German system. We also filtered the phrase tables to remove phrase-pairs

that would clearly result in semantically incorrect translations.

Our work on human evaluation has led us to develop a semantic evaluation measure based on the UCCA (Universal Conceptual Cognitive Annotation) framework. We are currently developing an automatic version of this metric to give rapid feedback on the semantic accuracy of translations.

We have recently shown that neural MT can produce better results for most of our language pairs, using continued training with synthetic data for adaptation, and will be rolling out NMT systems in Year 3. We are investigating how our work on semantic accuracy and treatment of morphology can be applied to NMT, for instance by incorporating semantic roles into the NMT system, or by using additional signal from back-translation to confirm the semantic accuracy. Our machine-learning version of the corrective morphology tool *depfix* (known as *MLfix*) will be used in the Year 3 system releases.

Finally, we are sponsoring this year's WMT biomedical translation task¹, providing test sets for the HimL language pairs, and collaborating in the release of a medical MT training set (UFAL Medical Corpus).

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 644402.

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

¹<http://www.statmt.org/wmt17/biomedical-translation-task.html>

A translation-based approach to the learning of the morphology of an under-resourced language

Tewodros Abebe

Addis Ababa University, IT-PhD
Addis Ababa, Ethiopia
wolaytta.boditti@gmail.com

Michael Gasser

Indiana University, School of Informatics
Bloomington, Indiana, USA
gasser@indiana.edu

Morphological analysis and generation are essential to many natural language processing tasks. There are now a number of tools for developing finite-state transducers (FSTs), which can be run either as analysers or as generators, for languages that are well studied and increasingly sophisticated algorithms for the automatic learning of morphology for languages with sufficient data. However, for most languages, there are neither sufficient linguistic resources nor sufficient data. One way of creating computational resources for languages that do not have many is to start with the resources that exist for other, closely related languages and then to learn differences based on the limited data available (Pretorius and Bosch, 2009). In this project we apply this general idea to the problem of morphology learning and implement it for the specific case of the languages Wolaytta and Gofa.

Wolaytta and Gofa are members of the poorly researched Omotic family, spoken in southwestern Ethiopia. Wolaytta is the most spoken and best studied of the roughly 30 Omotic languages, while the closely related language Gofa has very few resources of any sort.

Given a target language (TL) whose morphology we would like to learn, our approach starts with a related, “source” language (SL) whose morphology is known. We assume the availability (or development as part of the project) of an FST for the SL. We also assume a small set of bilinguals who are literate in both SL and TL and can provide the system with word–translation pairs.

The system begins with the assumption that SL and TL have identical morphology; a copy of the FST for the SL is the initial state for the TL. Given a translation pair, the basic idea is to

attempt to translate the SL word, using the current state of the system’s TL knowledge, and compare the result with the correct TL translation. Small differences between the predicted and correct TL words lead to learning: the TL FST is modified in some way. Possible updates to the FST include modifications to the form of roots or affixes, to morphotactics (the sequence of potential affixes), and to alternation rules, which are responsible for the morphophonological changes that may take place at the boundaries between morphemes.

The project began in spring 2016 and will terminate in spring 2018. The first author has received funding for the research from the IT PhD program of Addis Ababa University. The expected outcomes of the project, all free, open-source, and available on GitHub under a GNU GPL3.0 license, are: (1) morphological analyser/generators (FSTs) for Wolaytta and Gofa, (2) a toolkit for the learning of FSTs for under-resourced languages based on the known morphology of a related language and a set of translation pairs.

Efforts so far have focused on developing an FST for Wolaytta, using the Helsinki Finite-State Transducer toolkit (Lindén et al., 2009), on collecting a data set of translation pairs from Wolaytta–Gofa bilinguals, and on solving the basic task of isolating roots and affixes when one or the other of these differs in a translation pair.

We are currently focusing on the more complex tasks of learning differences in the order or number of affixes and learning modified or new alternation rules. In both cases, it may be necessary to constrain the search space with phonological biases, for example, towards alternation rules that implement assimilation.

References

Lindén, Krister, Silfverberg, Miikka, and Pirinen, Tommi. 2009. HFST Tool for Morphology: an Efficient Open-Source Package for

Construction of Morphological Analyzers.
SFCM 2009.

Pretorius, L., and Bosch, S. 2009. Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology. In *Proceedings of the First Workshop on Language Technologies for African Languages* (pp. 96-103).

MT in real-world practice: Challenges and solutions at Swiss Federal Railways

Nadira Hofmann

STAR Group

Wiesholz 35, 8262 Ramsen
Switzerland

nadira.hofmann@star-group.net

Maryse Lepan

SBB AG

Hilfikerstrasse 1, 3000 Bern 65
Switzerland

maryse.lepan@sbb.ch

Abstract

This user study uses the example of the Swiss Federal Railways (SBB) to show how an MT system is evaluated and introduced in practice. The first part describes the motivation and requirements for the company when it comes to introducing machine translation. Part two explains how the benefits can be determined before the system is launched by reliably analysing real product jobs and how a conclusive pilot phase can be implemented in a company's real-world setting. The third part deals with the findings from the pilot phase and uses specific examples to show how they have been taken into consideration and implemented for the product going live.

1 Introduction

1.1 Multilingualism as a tradition

Switzerland has several official languages and has written the promotion of “*understanding and exchange between the linguistic communities*” into the constitution. As the national rail company, the Swiss Federal Railways (SBB) therefore also has a long tradition of linguistic diversity: Translation has been part of operations at SBB since it was founded over 100 years ago. Multilingual project teams and multilingual communication with employees, customers and suppliers are part of day-to-day business.

1.2 SBB Language Services

The 15 people in the SBB Language Services team are supported by 10 external translation agencies and freelance translators. It is responsible for all translations and for the centralisation and management of the corporate language and terminology, which is developed in collaboration with technical experts and language specialists.

The following language technology is used:

- STAR CLM (Corporate Language Management) for managing the language processes,
- Transit as the translation memory system,
- TermStar and WebTerm as the terminology management systems.

Thanks to this technology, since 2001, SBB Language Services has developed a large, well-structured translation memory and comprehensive dictionaries with validated terminology. The SBB dictionary that resulted from this is available to the entire workforce at SBB.

1.3 Creating added value from existing data

The company was prompted to think about introducing an MT system by an SBB talent programme for first-line managers where the task was to create added value for SBB.

It was an obvious choice to use machine translation to generate added value from the linguistically validated data from the translation memory and terminology. Two approaches were pursued for this:

- Integrating the MT into SBB Language Services' existing translation workflow in order to support professional translators by offering additional MT-generated translation suggestions,
- Integrating the MT into the SBB intranet portal in order to support all employees by offering ad-hoc translations for their communications (“*SBB Translate*”).

The following requirements were present as framework conditions:

- Develop a valid decision-making tool to decide on the sense, practicability and economic efficiency of an MT system,
- Use the company's own terminology and formulations in the railway jargon that is approved by SBB,
- Seamless integration into existing processes and into the IT environment at SBB,
- Scalable and expandable for future requirements.

2 Evaluating the added value

To decide whether the solution actually generates added value, the decision-makers required reliable information regarding the benefits and quality in everyday translation.

For “facts and figures”, engine training and an initial analysis with real productive jobs from the SBB Language Services were initially carried out. The prerequisites for integrating the MT solution into its IT environment were also checked.

The pilot phase that followed involved the evaluation of how those involved in the process handle MT in practice.

2.1 Training the evaluation engine

The sample evaluation is carried out with *one* language combination, for which the involved parties can handle and evaluate the source and target language.

Therefore, in this case, German-French was trained with the SBB training material (2,593,609 segments for the translation memory and 42,788 dictionary entries). The engine was trained and hosted by the system provider.

2.2 Initial analysis with real production jobs

846 production jobs from recent months that were human-translated and human-reviewed without MT support were then analysed.¹ The jobs were repeated using MT and automatically compared with the existing results from the human translators and reviewers.²

¹ The document types were categorised as follows: 72% Word, 20% PowerPoint, 7% Excel, 1% Visio. The total volume was 587,927 words or nearly 4.7 million characters.

² In order to get meaningful results, these production jobs were *not* part of the SBB training material and therefore not used to train the engines.

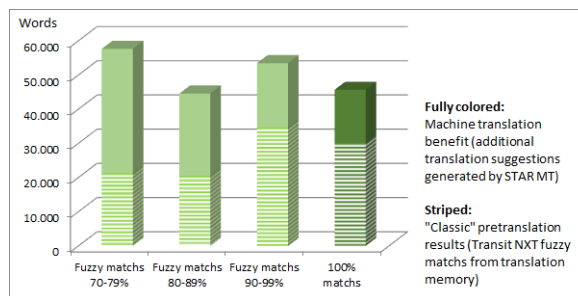


Figure 1. Additional MT-generated translation suggestions from the MT system

Result:

- The “perfect matches” (suggestions that could be applied without any changes) increased by 50%.
- The number of “good” translation suggestions (fuzzy quality 70-99%) more than doubled.

2.3 Pilot phase in translator's everyday work

The figures from the initial analysis proved the theoretical benefits of using MT. The pilot phase had to show whether this would result in real added value in a translator's everyday work. In addition to pure figures, acceptance, usability and “perceived” benefits also play a leading role.

To answer this question under real-life conditions, the MT should and must be integrated into real production jobs that are part of ongoing operations. As for all production jobs, the project management was therefore the responsibility of the customer. However, the engines were hosted by the system provider because, from experience, customers/interested parties in the pilot phase do not want to concern themselves with how the MT works or its infrastructure.

There are several scenarios for integrating the hosted MT solution. In this specific case, the project managers created their projects without machine translation and sent the project packages to the system provider. Here, segments that were not pretranslated from the translation memory were enriched using translation suggestions from the hosted MT system³ and the packages were sent back to the project manager who was then able forward them to the translator.

³ MT suggestions are not only created for “no matches” but also for segments with fuzzy matches from the TM. The procedure is still unusual, but it is logical: High fuzzy quality implies that the segment to be translated is very similar to the TM. Since the TM also acts as the basis for the engine training, the statistical machine translation for these segments provides particularly good results.

With this, nothing changes for internal and external translators: They do not require any additional tools or work steps; the MT suggestions are offered as fuzzy matches in the translation editor window of the translation memory system and can then be used with the familiar functions.

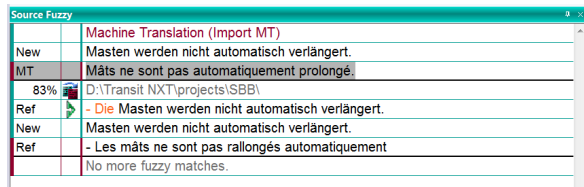


Figure 2. MT suggestion and classic fuzzy match in Transit's translation editor

The translators also do not need to access to the MT system or the hosted engines: They receive the additional MT suggestions automatically with the project packages. This means that external service providers and employees who work from home can be easily integrated into the pilot phases without any technical hurdles.

2.4 Web application for specific translation tests

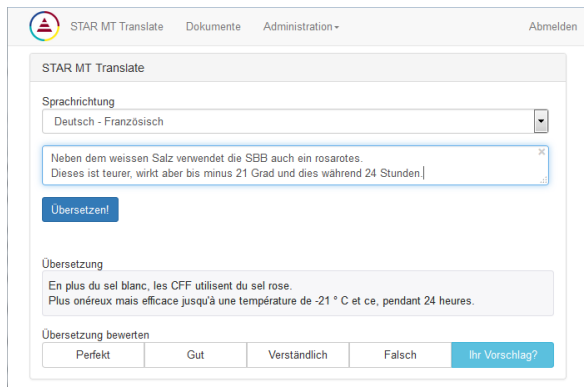


Figure 3. Prototype for the online solution – not yet customised and adjusted to the customer's corporate identity

In parallel to this, a web application was provided in order to request machine translations from the hosted engine via the browser. This means that the experts from SBB Language Services were able to carry out tests to determine how individual sentences or paragraphs are translated by the MT system and how the system performs to different source texts.

At the same time, the application was a prototype for SBB Translate as an online solution for translating individual sentences, but also entire

documents⁴. Here, selected SBB employees were able to test whether the solution meets the demands of everyday work.

3 Findings and challenges from the pilot phase

The pilot phase produced the following findings, which could then be taken into consideration when the system goes live.

3.1 All language directions directly and without pivot languages

In accordance with the requirement from SBB, all possible combinations and directions of the four languages (German, French, Italian and English) had to be supported.

Analysis of the existing data (volumes of the text corpora and terminology) showed that, for all translation directions requested by the customer, separate engines can be trained – i.e. a total of 12 engines for four languages.

This means that all translations can be carried out directly. It does *not* result in any of the adverse effects on quality or performance that are expected from machine translations using the “detour” pivot language.⁵

3.2 No differentiation according to subject area

The well-structured translation memory would have allowed a differentiation in order to train separate engines according to subject area.

However, the pilot phase showed that this differentiation is neither necessary nor useful. For this reason, the SBB material was used to train just one engine for each language direction.

⁴ All of the file formats that are relevant in practice can be supported. For the specific customer, support for Office documents, PDFs and text files is provided.

⁵ It makes sense to use the language that is used most frequently (in this case, German) as the pivot language. Language combinations that do not involve pivot languages would be translated using two successive machine translations.

Example French-Italian translation: The French text would be machine-translated into German and the German text would then be translated into Italian.

This increases potential MT errors and the server load increases because two machine translations would have to be performed for one translation request.

3.3 Generic back-up engines for general-language translation requests

When testing the web applications, the experts from the language department predominantly request translations of railway-specific texts with specialist terminology. In contrast to this, the requests from testers outside of the language department were significantly more general-language.

The engines were not initially trained for this; the translation results did not always meet the expectations of the “translation laymen”.

To translate these types of text with better results, additional, generic back-up engines are used for the web application. These are enriched with freely available corpora (e.g. from Europarl) and are automatically taken into consideration if the SBB-specific trained engines cannot generate a suitable translation.

3.4 The human factor

We know that the human factor decides on the acceptability and usability of the MT. The various target groups (professional translators vs “normal” employees) have differing knowledge, expectations and reservations, which all have to be taken into consideration when the MT goes live.

In terms of staff numbers, the target group of translation professionals is a known quantity and these people can therefore be informed directly and individually. A key aspect is the use of MT suggestions in the translation process. This is made easier by the fact that the translation process itself does not change – it is “only” supported by additional translation suggestions.

An individual approach is not possible for the numerous users of the web application (30,000 SBB employees). They are informed about the opportunities and limitations of machine translation via an attractive FAQ area: MT as a tool for understanding foreign-language texts – but not as a replacement for professional translation by SBB Language Services for documents that are to be published.

3.5 Text corpora with a broad range of formats

Many companies call upon the content from the projects from the language department as well as the extensive data stocks from translations that could be used as training corpora. This content is also not usually available in translation-typical exchange formats (e.g. TMX or XLIFF) because

the sustainable data usage and managed language processes are not often at the forefront of such translations.

In this specific case, the contents of the SBB website were localised by external agencies, for example, but could not be retrieved from the translation memory for SBB's language department.⁶ The content was then only available in the form of more than 20,000 HTML files.⁷

To prepare such contents for engine training, the translation memory system's filter technology is used. It generates format-neutral language files so that content from any source and format can be used.

3.6 Morphologically generated additional information

Terminology plays an important role in engine training and has a significant impact on MT quality: The more validated terminology is used for engine training, the better the translation results provided by MT.

SBB Language Services has carried out extensive terminology work that has, to date, been used in collaboration with the TM. The dictionaries also usually contain the base form of nouns, verbs and adjectives, while the texts that are to be translated usually contain inflected forms. A large proportion of the terminological potential would remain untapped if MT engines were only trained with canonical forms.

Morphology is used to close the gaps between dictionary entries and real texts. The technology for this comes from the translation memory system, which provides morphological support for over 80 languages and language variants. In this case, “morphology” means linguistic expertise mapped out in tried-and-tested rules, and not just simple stemming.

As an example, the values from the French-German engine show what morphology can offer:

- Text corpus: Translation memory with 3,007,240 segments/36,882,122 words

⁶ The website contents were particularly valuable for training the English engines: In the SBB Language Services' translation memory, English was heavily under-represented when compared with the other languages.

⁷ The HTML files were asynchronous material as the individual languages differed in structure and contents. Therefore the material was used as basis for monolingual data to train the language models.

In scenarios with synchronous bilingual or multilingual documents, bilingual language file pairs can be generated and used for engine training.

- Terminology: 35,815 language entries
- Morphologically generated additional terminology and segments in which they occur: 2,063,522 segments

The larger terminology base and the additional context information means that the BLEU score for this engine increases from 35 to 48. Irrespective of the BLEU score, the translations that were enhanced by extra morphologically generated terminology were clearly preferred by translators during a manual evaluation of sentence BLEU lists.

3.7 Web application with automatic TM pretranslation

The trained SBB engines usually provide good results for texts that were not previously professionally translated. However, the human translations that are validated by SBB Language Services and are available in the translation memory are of a higher quality. By definition, they have a BLEU value of 100.

The web application therefore uses the same process that is common and established for professional translation systems: For segments that are already contained in the translation memory, the translation from the translation memory is used (100% matches); for the rest, the MT engines are used. Thanks to the high performance of the TM indices, the user does not notice any increase in the response times.

In addition to the increased translation quality, the two-stage process has another advantage: Newly translated segments from SBB Language Services' projects immediately flow into the web application from the translation; new formulations, terminology and text types are immediately available. This means that the intervals for re-training the engine can be increased.

3.8 Focus for optimisation strategies

MT processes and MT engines are complex and have many influencing parameters that help to further improve the quality of the translation results. The theoretical opportunities are almost infinite but, in practice, it is useful to focus on the relevant areas.

This decision is supported by analysis functions. Examples:

- The evaluation of the engine-specific access figures shows which language combinations are used particularly intensively.
- The interactive feedback function of the web application reproduces the “per-

ceived” translation quality and allows for targeted improvements, where required.

- With OOV statistics (“out of vocabulary”), SBB Language Services can determine which terms in the terminology work should be prioritised.
- Quantitative evaluations of peak loads and load distribution provide IT with information about the sizing of the system.

4 Long-term perspectives

The introduction of an MT system is no rush job: It may take months before the solution takes hold and is accepted by the employees in their everyday business.

It is even more important that this investment in time and resources offers a long-term perspective and is open to future requirements that cannot currently be foreseen by the parties concerned. Future-proof interfaces that allow the solution to be integrated into changing IT structures are particularly relevant.

In this specific case, the following scenarios are envisaged and, from a technical point of view, could already be implemented:

- Apps for iOS/Android so that the MT can be easily used on mobile devices.
- Integration into Office in order to request MT directly from Word, Excel, Power-Point or Outlook
- An API that can be used to translate texts and documents using third-party applications.

Can Out-of-the-box NMT Beat a Domain-trained Moses on Technical Data?

Anne Beyer*, Vivien Macketanz^Δ, Aljoscha Burchardt^Δ and Philip Williams[◊]

* beo Gesellschaft für Sprachen und Technologie mbH
Ruppmannstrae 33b, 70565 Stuttgart, Germany
anne.beyer@beo-doc.de

^Δ German Research Center for Artificial Intelligence (DFKI)
Language Technology Lab, Alt-Moabit 91c, 10559 Berlin, Germany
firstName.lastName@dfki.de

[◊] University of Edinburgh
School of Informatics, 10 Crichton Street, Edinburgh, EH8 9AB, UK
pwillia4@inf.ed.ac.uk

Abstract

In the last year, we have seen a lot of evidence about the superiority of neural machine translation approaches (NMT) over phrase-based statistical approaches (PBMT). This trend has shown for the general domain at public competitions such as the WMT challenges as well as in the obvious quality increase in online translation services that have changed their technology. In this paper, we take the perspective of an LSP. The questions we want to answer with this study is if now is already the time to invest in the new technology. To answer this question, we have collected evidence as to whether an existing state-of-the-art NMT system for the general domain can already compete with a domain-trained and optimised Moses (PBMT) system or if it is maybe already better. As it is well known that automatic quality measures are not reliable for comparing the performance of different system types, we have performed a detailed manual evaluation based on a test suite of domain segments.

1 Introduction

In the last year, we have seen a lot of evidence about the superiority of neural machine translation approaches (NMT) over phrase-based statistical approaches (PBMT). This trend has shown for the general domain at public competitions such as the WMT challenges (Bojar et al., 2016) as well as

in the obvious quality increase in online translation services that have changed their technology.¹

When it comes to particular domains in the context of commercial translation services, the interest in NMT is huge, but we are not aware of systematic public studies about the performance of NMT in comparison to PBMT. While bigger companies are already in the process of changing their technology, smaller language service providers (LSP) have limited resources in their day-to-day-business both in terms of humans and compute power for undertaking the necessary experiments. For researchers, it is still difficult to obtain suitable training data in order to assess the potential of the new technology on in-domain data.

The background for this study was simply the question of an LSP if now is already the time to invest in the new technology. To answer this question, we wanted to collect evidence as to whether an existing state-of-the-art NMT system for the general domain can already compete with a domain-trained and optimised Moses (PBMT) system or if the former can maybe even outperform the latter already.

As we did not want to rely solely on automatic measures, we have performed a manual evaluation based on a phenomenon-driven test-suite, a method we have applied for evaluations in the technical domain before, e.g., in (Avramidis et al., 2016).

2 Experiment

2.1 Data

The customer data used in this study came from translations of catalogues for technical tools. Our

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

¹<https://research.googleblog.com/2016/09/a-neural-network-for-machine.html>

dataset consisted of translation tasks from German into British English assigned to beo over a course of two months. Overall, the set contained around 5,000 segments.

2.2 Phrase-based Statistical MT System

The PBMT system used is based on Moses (Koehn et al., 2007) and was adapted to integrate MT into the translation workflow at beo.

As training data we used the customer’s translation memory (TM) and terminology, which yielded a total of 337,600 segments. Formatting tags were removed from the data and it was tokenized and lower cased. As we translate from German, compounds were also split on the source side in order to reduce data sparseness in terms of unknown words. A 3-gram language model was built using IRSTLM (Federico et al., 2008).

The training procedure follows the baseline Moses setup², but the model was not tuned further, as no tuning setup was found yet which improved the system’s performance over the baseline, according to an internal evaluation with our translators. This is similar to what we found for other customer set-ups. It could be due to the fact that the training-data and the translations are very similar, as we only used in-domain data for training. We have not yet tried to add more out-of-domain data because this did not improve the usefulness of systems trained for other customers, but might look into that at a later point as well. As we are only concerned with the application of MT for post-editing, the quality requirements are different from other tasks such as quality evaluation and we rely more on post-editor feedback that automated quality scores.

For the translation, we used the M4Loc integration tools³, a wrapper for Moses which extracts formatting tags before the translation and inserts them into the target afterwards according to the word alignment (Hudk and Ruopp, 2011). Furthermore, we ran a few test rounds on the customer data together with our translators and created a set of hand-crafted rules based on regular expressions which are applied after the MT to fix certain errors (e.g., with casing or spaces).

²<http://www.statmt.org/moses/?n=Moses>.
Baseline

³<https://github.com/achimr/m4loc>

2.3 Neural MT System

The neural system that was used in this study was built by the University of Edinburgh. This MT engine is the top-ranked system that was submitted to the WMT ’16 news translation task (Sennrich et al., 2016). The system was built using the Nematus toolkit.⁴

As training data, only the official WMT task data was used – this system did not have access to the customer-specific data during training. The data was tokenized and truecased, and tokens on both the English and German sides were split into subword units using byte-pair encoding (BPE), a frequency-based method that aims to improve the handling of rare words.

The full training configuration and scripts for this system have been publicly released.⁵

2.4 Manual Evaluation Procedure

For the manual evaluation process, two professional (computational) linguists went through the data and identified reoccurring linguistic phenomena that are characteristic for this domain-specific data.⁶ In a second step, all the phenomena detected were narrowed down to the most prominent ones, namely formal address, genitive, modal construction, negation, passive voice, predicate adjective, prepositional phrase, terminology and tagging. Thereafter, 100 segments per phenomenon were extracted, resulting in a total of 900 segments. For each segment, the total occurrences of the respective phenomenon were counted. Then, the total occurrences of the phenomena in the MT outputs were counted. Consequential, translation accuracy was calculated by dividing the number of occurrences in the MT output by the total number of occurrences in the segments.

When evaluating the correctness of the translations, the focus lies solely on the respective phenomenon under consideration, other errors are ignored. For a translated phenomenon to be counted as correct, it does not necessarily exactly have to match the reference, but it can also be realized in a different linguistic construction expressing the same semantic meaning, e.g., a passive construction that is translated in active construction will

⁴<https://github.com/rsennrich/nematus>

⁵<https://github.com/rsennrich/wmt16-scripts>

⁶These “linguistic phenomena” are understood in a pragmatic sense and include a wide range of issues that can influence the translation quality.

have less components but if the meaning is translated correctly, the counting should be adjusted to the instances in the source accordingly.

3 Evaluation Results

Due to the repetitive nature of the customer data, some of the segments in our dataset were already part of the TM or very similar to segments in the TM and therefore part of the training data for the Moses system. In order not to distort the results too much, those segments where Moses exactly matched the reference translations were omitted from the automatic evaluation. For the manual evaluation, we did not exclude those segments.

3.1 Automatic Evaluation Results

Even though BLEU is not intended to be used in order to compare different MT systems, this is a practice that is performed quite often. In order to show how much different translation quality evaluation methods can vary, we also carried out an evaluation on BLEU and METEOR, cf. Table 1. For calculating the automatic score, all tags were removed from the segments and the reference, furthermore all numbers were replaced by “10” because there were cases in which the reference involved different tags/numbers than the segments.

| | NMT | Moses |
|--------|-------|-------|
| BLEU | 23.68 | 47.98 |
| METEOR | 28.46 | 38.26 |

Table 1: BLEU and METEOR scores.

As described above, the automatic evaluation has a clear bias towards Moses. This is amplified by the fact that the references were derived from post-edits of the Moses output. These segments are thus naturally more similar to the Moses output than to the completely independent NMT output. Despite removing the segments for which the translation by Moses exactly matched the reference, both BLEU and METEOR show distinctly better scores for Moses compared to the NMT system. Taking into account the manual evaluation, though, gives a different picture.

3.2 Manual Evaluation Results and Examples

Table 2 shows the results of the manual evaluation on segment-level. For the 900 segments extracted, 1,453 phenomena could be found altogether, as

there was often more than one occurrence of the phenomenon per segment. Phenomena like terminology occur more frequently than phenomena like negation that rarely appear more than once within one segment. Percentage values in boldface indicate that the systems is significantly better on the respective phenomenon with a 0.95 confidence level.

| | # | NMT | Moses |
|----------------------|------|------------|-------------|
| formal address | 138 | 90% | 86% |
| genitive | 114 | 92% | 68% |
| modal construction | 290 | 94% | 75% |
| negation | 101 | 93% | 86% |
| passive voice | 109 | 83% | 40% |
| predicate adjective | 122 | 81% | 75% |
| prepositional phrase | 104 | 81% | 75% |
| terminology | 330 | 35% | 68% |
| tagging | 145 | 83% | 100% |
| sum | 1453 | | |
| average | | 89% | 73% |

Table 2: Manual evaluation translation accuracy focusing on particular phenomena.

The NMT system outperforms Moses on three categories: genitive, modal construction and passive voice. Moses on the other hand outperforms NMT on terminology and tagging – which is not surprising as terminology was part of the TM and tagging was handled by an extra module. For the remaining phenomena, the systems show no statistical significant variance. Additionally, the NMT system also outperforms Moses on the overall average.⁷ Nevertheless, it is important to keep in mind that the values of the manual evaluation only give insights on certain phenomena and do not necessarily represent the systems’ overall performance but can rather be interpreted as revealing a tendency. Interestingly, the tendency the manual evaluation displays is counter to that of the automatic scores shown in Table 1. This can be traced back to the training material for Moses which included the the customer’s translation memory and terminology which has a high influence on the BLEU and METEOR scores. The manual evaluation results on the other hand imply that even if a translation deviates substantially from a given reference it can

⁷Average calculation: division of the sum of the absolute numbers of correct segments by the sum of all segments for each system.

still be correct, a fact that is not taken into account in the automatic scores.

The following examples depict interesting findings from the analysis and comparison of the two systems. The relevant component of the sentence is underlined. When a system created a correct output for the respective phenomenon, the system name is marked in boldface.

- (1) Source: Schweißbänder erhöhen wesentlich den Tragekomfort eines Helmes.
Ref.: Sweatbands significantly increase the wearing comfort of a helmet.
NMT: Welding tapes significantly increase the comfort of a helmet.
Moses: Welding belts significantly increase the wearing comfort of a Helmes.

Example (1) contains the genitive *eines Helmes* that should correctly be translated as *of a helmet*. As can be seen, the NMT correctly translates the genitive while Moses leaves *Helmes* untranslated which makes it hard to tell whether it correctly translates the genitive. This was a systematic problem for Moses, as Moses left unknown words untranslated. The NMT system on the other hand often generated sentences that were grammatical and contained “only” mistranslated unknown words rather than untranslated unknown words. As a result, syntactic features like the genitive in example (1) can be maintained.

- (2) Source: Dazu kann das Board werkzeuglos gedreht und wieder eingehängt werden.
Ref.: The board can be turned and re-attached without using tools.
NMT: The board can be rotated and re-mounted.
Moses: To do this, the board can be rotated and back.

Example (2) includes a modal verb construction. A modal verb is always followed by at least one other verb. In the construction above, the modal verb *kann* is followed by the two verbs *gedreht* and *eingehängt* as well as the verb *werden*. Those verbs form a processual passive construction. In order to count as correctly translated, the English MT outputs should also exhibit four verbs, as the construction is formed the same way in English. While the NMT system correctly translated all four verbs, Moses leaves out one verb. Note that the fact that both systems do not translate *werkzeug-*

los (*without using tools* in the reference) can be ignored in this evaluation as the focus lies exclusively on the phenomenon of modal verb constructions.

- (3) Source: Die Panoramascheibe mit integriertem Seitenschutz sorgt für eine <g id="1004">optimale Augenraumabdeckung</g>.
Ref.: The panoramic lens with integral side protection ensures <g id="1004">optimum coverage of the eye area</g>.
NMT: The panorama disc with integrated side protection ensures a <g id="1004">optimal eye room cover</g>.
Moses: The panoramic lens with integral side protection ensures <g id="1004">optimum Augenraumabdeckung</g>.

The third example given here is taken from the terminology category. Additionally, it contains tagging which can be ignored in this case. The source sentence contains three terms: *Panoramascheibe*, *Seitenschutz* and *Augenraumabdeckung* which should be translated as *panoramic lens*, *side protection* and *coverage of the eye area*, respectively. The NMT system only correctly translates *side protection* while mistranslating the other two terms, giving literal translations. Moses correctly translates two of the three terms, leaving *Augenraumabdeckung* untranslated. Nevertheless, at first glance the NMT output looks “better” because it does not leave words untranslated. When taking a closer look though, this assumption does not hold.

As Moses benefits in terms of knowing a subset of the terminology, we considered it reasonable to also analyze segments without terminology in order to draw some more general conclusions about the comparison between the two systems, independent of the domain. For this purpose, 90 segments without domain-specific terminology were extracted from the data set. These segments comprise 30 short (< 40 characters), 30 medium-length (40 - 79 characters) and 30 long (> 79 characters) items. Two annotators were asked to evaluate these segments individually, rating them on a scale from 1 - 3, with 1 = perfect translation, 2 = small errors, content still understandable, and 3 = unintelligible. The mean values of the two annotators can be found in Table 3. While the NMT’s

performance is judged better for the longer segments, Moses’ performance is judged better for short and medium-length segments. Nevertheless, conducting a *t*-test showed that the differences in the mean values are not statistically significant. Yet, it should be kept in mind at this point that we did not expect the differences to be statistically significant as the population of segments examined was very small. We interpret the scores solely as a tendency.

Below, we will discuss an example from this category:

(4) Source: Neben den Bedingungen zur Aufstellung und Inbetriebnahme wird eine Vielzahl von technischen und gesetzlichen Anforderungen an das Lager selbst gestellt, um z. B. wassergefährdende Flüssigkeiten, Säuren und Laugen oder auch entzündbare Flüssigkeiten gesetzeskonform aufzubewahren und zu lagern.

Ref.: In addition to the conditions for erection and commissioning there are a wide variety of technical and legal requirements on the storage location itself, relating for instance to water-polluting liquids, acids and alkalis or also flammable liquids, which must be kept safe and stored in accordance with regulations.

NMT: In addition to the conditions for installation and commissioning, a wide range of technical and legal requirements will be placed on the warehouse itself in order to maintain and store, for example, water-hazardous liquids, acids and foliage, or even flammable liquids.

Moses: In addition to the conditions for erection and commissioning is a wide variety of technical and legal requirements of the stored even, e. g. for water-polluting liquids, acids and alkalis or flammable liquids legally compliant aufzubewahren and storage.

Example (4) belongs to the long segments, having 293 characters. While there were long segments that consisted of several sentences, this segment comprises only one sentence. It contains an in-

| | ∅ NMT | ∅ Moses |
|------------------------|-------|---------|
| short segments | 1.7 | 1.5 |
| medium-length segments | 2.1 | 1.9 |
| long segments | 2.2 | 2.3 |

Table 3: Mean values for segments without terminology.

finitive clause that reaches from the conjunction *um* to the verb *zu lagern*. While in German, objects are located between the conjunction and the last verb, in English the conjunction *in order to* is immediately followed by the verb in the infinitive with the objects being located behind the verb. The NMT system successfully manages to resolve this construction, placing the verbs at the right position while Moses not only leaves the verbs at the end of the sentence but also leaves one verb untranslated. This example depicts our finding that NMT can handle long sentences better than Moses.

At the same time, this sentence also highlights difficulties that can arise, e.g., for post-editing, by the fact that the NMT system substitutes unknown words in the source with similar words in order to be able to translate them. While in some cases this might work out well, there are other cases where it does not, as in example (4) above: The word *Laugen (alkalis)* was treated as the word *Laub* which means *foliage*, resulting in a rather curious translation. For post-editing this means that in order to detect erroneous translations it is crucial to check the NMT output very thoroughly because mistranslations might be harder to find than in a system output that contains untranslated words.

4 Conclusion and Outlook

From the viewpoint of the linguistic phenomena we have studied in our experiment, the answer to the question in the title of this paper would probably be a sentence beginning with “Yes, but ...”. The reason for the restriction is that the two categories NMT can not yet handle as good as Moses are of high importance in the language business: tags and terminology.

Still, sooner rather than later there will be tag-handling components for NMT systems and the issues with terminology will probably vanish once the NMT is trained on customer domain data. So, from the analytic perspective we took here, NMT could indeed become a valid alternative to PBMT

for commercial use in the future.

The purpose of this study was to determine if now is already the time for LSPs to start investing in NMT. Our comparison showed that even an out-of-the-box system can perform quite reasonably, although it was not trained on the specific data. Our next step will be to look into the OpenNMT system⁸ and to compare models trained on the same dataset. Here, we will also take a closer look at other important factors, such as the time and effort needed for setting up such a system, the different training and decoding times and the impact of different kinds of errors on the post-editing effort.

For this purpose, We plan to also perform productivity tests with post editors to get a second, less phenomenon-driven comparison between the systems. In this course, we may also re-calculate automatic scores using post-edits as reference translations to rule out the Moses bias we have clearly observed in the figures we have presented here. For scenarios without post-editing, it would also be interesting to repeat task-based evaluations like the one we present in (Gaudio et al., 2016).

Another follow-up study that could be conducted might focus on a comparison of systems which are more similar with regard to their setup of the training data. In doing so, it would be interesting to investigate whether, for instance, an NMT system’s BLEU and METEOR scores might get closer to those of an SMT system, and if the bias towards the NMT system in the manual evaluation scores persists or even increases.

Acknowledgement

This article has received support from the EC’s Horizon 2020 research and innovation programme under grant agreements no. 645452 (QT21). We thank the anonymous reviewers for their valuable feedback.

References

Avramidis, Eleftherios, Vivien Macketanz, Aljoscha Burchardt, Jindrich Helcl, and Hans Uszkoreit. 2016. Deeper Machine Translation and Evaluation for German. In Hajic, Jan, Gertjan van Noord, and Antnio Branco, editors, *Proceedings of the 2nd Deep Machine Translation Workshop. Deep Machine Translation Workshop (DMTW), October 21, Lisbon, Portugal*, pages 29–38. Charles University Prague, Charles University, Prague, 10.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany, August. Association for Computational Linguistics.

Federico, Marcello, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *9th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1618–1621, Brisbane, Australia, September.

Gaudio, Rosa, Aljoscha Burchardt, and Antonio Branco. 2016. Evaluating Machine Translation in a Usage Scenario. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Hudk, Tom and Achim Ruopp. 2011. The integration of moses into localization industry. In *Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT)*, pages 47–53, Leuven, Belgium, May.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. *CoRR*, abs/1606.02891.

⁸<http://opennmt.net/>

BabelDr vs Google Translate: a user study at Geneva University Hospitals (HUG)

Pierrette Bouillon¹, Johanna Gerlach¹, Hervé Spechbach², Nikos Tsourakis¹, Sonia Halimi¹

¹ FTI/TIM, University of Geneva

{pierrette.bouillon, johanna.gerlach, nikolaos.tsourakis, sonia.halimi}@unige.ch

² Geneva University Hospital (HUG)

herve.spechbach@hcuge.ch

Abstract

This paper presents a user study carried out at Geneva University Hospitals (HUG) where we compared BabelDr, a flexible phraselator, with Google Translate (GT). French speaking doctors were asked to use both systems to diagnose Arabic speaking patients. We report on the user's interactions with both systems, the quality of translation, the participant's ability to reach a diagnosis with the two systems as well as user satisfaction.

1 Introduction

In the context of the current European refugee crisis, hospitals are more and more often obliged to deal with patients who have no language in common with the staff, and may also fail to share the same culture. For example, at the Geneva University Hospitals (HUG), Geneva's main hospital, 52% of the patients are foreigners and 10% speak no French at all. In 2015, the languages which caused most problems were Tigrinya, Arabic and Farsi; refugees from Eritrea, Syria and Afghanistan make up about 60% of all new demands for asylum in the area (SEM Newsletter, October 2015). The problems are not only linguistic. Cultural differences mean that these patients may have different conceptualizations of medicine, health care (Hacker et al., 2015), illness and treatment (Priebe et al., 2011). A situation of this kind, with barriers in language, culture and medical understanding, creates serious problems for quality, security and equitability of medical care, as has

been pointed out by several researchers ((Flores et al., 2003) and (Wasseman et al., 2014)). Others underline the negative impact these issues have on health care costs (Jacobs et al., 2007).

In absence of qualified interpreters, a number of solutions are available today, but they all have their drawbacks. Phone-based interpreter services are very expensive (3 CHF/minute in Switzerland), not always available for some languages, and known to be less satisfactory than face-to-face interaction through a physically present interpreter (Wu et al., 2014). Google Translate (GT), increasingly often used when no other alternatives exist, is known to be unreliable for medical communication (Patil and Davies, 2014). Other tools like MediBabble and Universal Doctor have been developed specifically for the medical diagnosis scenario and translate a set of fixed questions, but are technically unsophisticated and content cannot easily be changed. Similar remarks apply to medical resources developed for refugees in conflict zones, like the *Medical Handbook for Refugees*¹, which are non-interactive databases.

BabelDr² is a joint project of Geneva University's Faculty of Translation and Interpreting (FTI) and Geneva University Hospitals (HUG) which specifically addresses this problem of lack of qualified interpreters in hospitals in languages spoken by refugees. The BabelDr application can be characterised as a flexible speech-enabled phrasebook (Rayner et al., 2016). Semantic coverage consists of a prespecified set of utterance-types, but users can use a wide variety of surface forms when speaking to the system. Each utterance-type is associated with a canonical source language version, which is rendered into the target languages

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

¹<https://www.refugeephasebook.de/medical%20phrases/>

²<http://babeldr.unige.ch/>

by suitably qualified translation experts. The central design goals are to ensure that a) translations are reliable, b) new target languages can easily be added, enabling flexibility in the face of changing patient demographics and c) content can be changed depending on the context.

In this paper, we present a user study done at Geneva University Hospitals (HUG) where we compared the baseline version of BabelDr with the online desktop version of GT in real hospital settings. We report on the 1) interactions in both systems, 2) quality of translation and 3) impact on diagnosis and satisfaction. Our hypotheses are that GT is not precise enough for this domain and that BabelDr is robust enough to make the diagnosis possible. Section 2 presents BabelDr in more detail; Section 3 describes the experiment and Section 4 the results.

2 BabelDr

The baseline version of BabelDr used for this experiment has been designed to assist in triaging of non-French-speaking patients visiting HUG's A&E department. It allows medical professionals to perform a preliminary medical examination dialogue, using a decision-tree method, to determine the nature of the patient's problem and the appropriate action to take. The coverage of the current version of the system consists of yes-no questions and instructions, and the patient is expected to respond non-verbally, e.g. by nodding or pointing with their fingers.

BabelDr differs from general speech translation systems like GT in several important respects. In particular, both speech recognition and translation are performed by domain-specific rule-based methods, as opposed to GT's general-purpose data-driven methods. As explained in (Rayner et al., 2016), they are for convenience split into multiple pieces, one for the source language and one for each target language, with the parts relevant to each language placed in different files; source and target languages are linked through canonical representations of the source-language utterances. The files are combined at compile-time, and the result is converted first into Synchronised Context-Free Grammar form (Aho and Ullman, 1969), and then into a GrXML grammar which can be compiled and run on the Nuance Toolkit 10.2 platform. This means that speech recognition, parsing and translation are all performed by the Nuance Toolkit

engine.

At runtime, the system echoes back the canonical form of the sentence to the source-language user, only producing a translation if the source-language user approves. The canonical form thus acts both as a pivot for translation and as a back-translation to verify recognition. It was designed with the help of HUG to be the less ambiguous and the most explicit as possible, for example a sentence such as *avez-vous l'impression d'être févreux ?* "do you feel you're running a temperature?" is mapped to *avez-vous de la fièvre ?* "do you have a fever?". Similarly *où va la douleur ?* "which way is the pain going?" corresponds to *pouvez-vous montrer avec le doigt où irradie la douleur ?* "could you show me with your finger the direction in which the pain is radiating?".

Target-language utterances can be realised in spoken form either using the Nuance Text-to-Speech Engine (TTS), or using prerecorded multimedia files. This functionality is needed for low-resource languages like Tigrinya, which currently lack TTS engines, and also for translation into sign language (Ahmed et al., 2017). The platform is entirely web-based. The runtime system runs on a cloud server and is accessed through a thin client running on a normal web browser. Content is remotely uploaded and compiled through a web interface. The methods used were developed on previous projects and have been described elsewhere (Fuchs et al, 2012; Rayner et al, 2015).

Linguistic coverage is organised into domains, centered around body parts (abdomen, head, chest and kidneys/back); there is nontrivial overlap, since some questions are common to all domains. At the time of writing, each of the four domains has a semantic coverage of around 2000 utterance types, with an associated grammar that uses a vocabulary of about 2000–2500 words and expands to on the order of tens of millions of surface sentences. The system supports translation from French to Arabic and Spanish, and there are partial sets of translations for Tigrinya, English, LSF-CH (Swiss French sign language) and Auslan.

The BabelDr interface was designed to resemble the GT interface, but presents several important differences (both interfaces are shown in Figure 1). First, since BabelDr is a phraselator, it provides help and gives access to the list of possible canonical sentences covered by the system. After each recognition event, the list of examples is up-

Figure 1: BabelDr and GT interfaces



dated and the system automatically moves to the recognized sentence, allowing to see related questions. Second, in BabelDr input is by speech only. If the system does not recognize the utterance correctly, the user has to speak again. In GT, users can edit the recognition result by typing, or bypass speech recognition entirely and type input. Third, instead of displaying a recognition result, BabelDr displays the canonical form of the spoken utterance. Finally, the way to use the microphones differs, GT being push-and-talk and BabelDr push-and-hold.

3 Experiment

3.1 Goal

The aim of this user study is to measure the impact of the medium (BabelDr, GT) on the diagnosis made by doctors. Both systems were used by doctors at HUG or medical students to perform a medical diagnosis, based on two scenarios. For each scenario (appendicitis and cholecystitis), a patient was standardized by HUG. The two patients both received the a priori list of symptoms for the disease they present. They were instructed to give a negative or noncommittal answer for all other symptoms. The order of system and scenario versions were balanced among participants, each participant performing two diagnoses, one with BabelDr and one with GT, in an alternate order. The experiment ends when the doctor reaches a diagnosis.

3.2 Languages and domain

The language pair for the study was French into Arabic. For BabelDr, the "abdomen" domain was used. In both systems, TTS was used for speech output.

3.3 Participants

All participants were recruited at the hospital and were paid for the task:

Arabic speaking patients: two standardized Syrian patients, one male and one female.

French speaking doctors: four medical students and five doctors (clinical chiefs) working at HUG.

3.4 Location and duration

The study took place at the HUG evaluation lab and was organized in two main sessions, a pre-test with the four students and the main user study with the five doctors. One week before each test, participants received a short introduction to both systems and were given 30 minutes to practice and ask questions.

3.5 Data collected

The following data were collected during the experiments: video recordings of the room, screen capture videos, eye tracking data, diagnoses reached by doctors after each scenario, demographic and satisfaction questionnaires.

The videos and screen captures were transcribed, in particular what was said by the doctors, the recognition result and the translation into Arabic. In the following sections, we analyse these results focusing on the doctor-patient communication rather than system performance. We will therefore look mainly at interactions which reached the patients. Section 4.1 focuses on interactions sent to translation by the doctors, Section 4.2 on the quality of their translation, section 4.3 on diagnosis and section 4.4 on user satisfaction.

4 Results

4.1 Interactions with the system

Table 1 summarizes the interactions with the two systems. Overall, the number of interactions was similar for both. On average, the doctors did 30 interactions per dialog, while students have an average around 45.

Table 1: Interactions with the systems

| | Total | Translated |
|----------------|-------|------------|
| GT | | |
| Students | 181 | 179 (99%) |
| Doctors | 150 | 140 (93%) |
| BabelDr | | |
| Students | 187 | 128 (68%) |
| Doctors | 156 | 109 (70%) |

Since the two systems function differently both in terms of recognition and translation, as described in Section 2, the definition of a successful interaction with each of the systems is not straightforward. Since the source language users do not understand the target language, they can only judge the correctness of speech recognition. In this section, we consider accepted interactions, namely those where the user has found the recognition result to be satisfactory, and has validated this either by sending the utterance to translation (in BabelDr) or by oralizing the translation (in GT, where translation is enabled by default). This does not necessarily imply that the recognition result exactly matches the spoken utterance, but rather that it expresses the meaning intended by the user.

Table 1 shows that the number of interactions sent to translation and oralized for the patient is higher in GT than in BabelDr, with 99% (students) and 93% (doctors) of interactions accepted *vs* 68% and 70% in BabelDr.

Table 2: BabelDr: non oralized interactions

| | Students | Doctors |
|------------------------------|-----------|-----------|
| 1. Out of coverage | 39 | 33 |
| a. Out of domain | 16 | 8 |
| b. Out of grammar | 23 | 25 |
| 2. In coverage | 15 | 11 |
| a. Canonical rejected | 3 | 6 |
| b. Recognition error | 12 | 5 |
| 3. Interaction issues | 5 | 3 |
| Total not translated | 59 | 47 |

A closer analysis of the rejected interactions in

BabelDr shows different causes. These interactions are detailed in Table 2. About two thirds of rejected interactions are cases where the user produced an utterance that was not covered by the system (1). These can be split into two types. First, interactions that were not among the canonical sentences included in the domain coverage of the system (1a). These were mostly wh-questions (*quel est votre problème ?* "what is your problem?") and declarative sentences that were not part of the usual anamnesis questions (*je vais appeler l'infirmière* "I will call the nurse"). Second, interactions using surface forms not covered by the grammar (1b). This accounts for 23 of the student's and 25 of the doctor's interactions. They are due either to gaps in the coverage or to users not complying with the instructions. Although participants were instructed not to use ellipsis, coordination, complex sentences or informal language during the introduction, some used them anyway, often resulting in incorrect recognition results. This category also includes disfluencies.

A second group (2) includes failed interactions for in coverage utterances. Some were rejected because users did not find the canonical appropriate (2a), or decided to ask something else instead. The rest (2b) were caused by recognition errors, sometimes due to a long silence at the beginning of the interaction.

Finally, for a small number of cases, interaction issues led to bad recording or translation (3).

Another aspect to consider when comparing the number of rejected interactions is that BabelDr only allows speech input while in GT, participants could also type when recognition did not work. We observe that the students corrected or typed 50 (28%) and the doctors 5 (3%) of their GT interactions. Table 3 shows the detail of these interactions. Between 2 and 3% of GT recognition results were corrected manually (for example, *allez-vous à sel normalement* "can you go to the salt normally" → *allez-vous à selles normalement* "can you go to the bathroom normally; *est-ce que la couleur et brune* "and the colour brown" → *est-ce que la douleur est brune* "is the colour brown"). For the students, we also observe a larger number of cases where modifications were related to incorrect interactions with the system, e.g. where they forgot to stop the microphone after the interaction. Finally, also for the students, we have a number of cases where users preferred typing input rather

than speaking (12%).

Table 3: Interactions modified by typing in GT

| | Students | Doctors |
|---------------------------|----------|---------|
| Correction of rec. result | 6 (3%) | 3 (2%) |
| Bad interaction | 22 (12%) | 2 (1%) |
| Typed input only | 22 (12%) | - |
| Total | 50 | 5 |

4.2 Translation quality

The sentences sent to translation by doctors (canonical form for BabelDr, recognition result or typed input for GT) were evaluated in terms of adequacy and comprehensibility by three Arabic advanced translation students of the Faculty of Translation and Interpreting of Geneva University. Adequacy was annotated on a four point scale (nonsense/mistranslation/ambiguous/correct) and comprehensibility on a four point scale (incomprehensible/syntax errors/non idiomatic/fluent). Evaluation was carried out in context and taking into account the sex of the patient (male or female). Table 4 presents the majority judgements for adequacy and comprehensibility as well as the number of cases where no majority was reached. Inter-annotator agreement for both evaluations is moderate (Light's Kappa for adequacy: 0.483; for comprehensibility: 0.44) according to (Landis and Koch, 1997).

Table 4: Translation evaluation (doctor's interactions sent to translation)

| | BabelDr | | GT | |
|--------------------------|---------|-----|-----|-----|
| Adequacy | | | | |
| no majority | 1 | 1% | 10 | 7% |
| nonsense | 0 | 0% | 53 | 38% |
| mistranslation | 0 | 0% | 0 | 0% |
| ambiguous | 7 | 6% | 24 | 17% |
| correct | 101 | 93% | 53 | 38% |
| Total | 109 | | 140 | |
| Comprehensibility | | | | |
| no majority | 0 | 0% | 14 | 10% |
| incomprehensible | 0 | 0% | 52 | 37% |
| syntax errors | 3 | 3% | 18 | 13% |
| non idiomatic | 3 | 3% | 3 | 2% |
| fluent | 103 | 94% | 53 | 38% |
| Total | 109 | | 140 | |

We observe that GT is less adequate and com-

prehensible than BabelDr. The evaluators also fail to reach a majority judgement more often for GT than BabelDr, suggesting that these translations are more difficult to evaluate. Interestingly, the BabelDr translations are not always considered as correct. In BabelDr, translations account for the gender of the patient, but were intended to be neutral in respect to cultural, educational and formality aspects. Evaluators disagree with some translation choices. An interesting improvement of the system would include more different patient profiles. The translators were also strict, marking some BabelDr translations as incorrect although they were completely meaningful (for example, *pouvez-vous me montrer avec le doigt où est la douleur ?* "could you indicate with your finger the pain location ?" -> هل يمكنك وضع اليد على ؟ منطقة الألم ؟ "could you indicate with your hand the pain location ?". This shows the subjectivity of human evaluation and the need to give better evaluation guidelines and to focus more on oral comprehension in future evaluations.

4.3 Diagnosis

Each of the 9 subjects had to find 2 diagnoses (1 appendicitis and 1 cholecysticis), one with each system. With GT, 5/9 doctors found the correct diagnosis, against 8/9 with BabelDr, which suggests that BabelDr is more suitable for the diagnostic task. If we look at the doctors only, they all found the right diagnosis with BabelDr, against 4/5 with GT. These results suggest that, even if it is possible to reach a correct diagnosis with bad translations, correct translations facilitate the task. It is interesting to see that BabelDr seems to help students to perform better diagnoses (1/4 diagnoses correct with GT and 3/4 with BabelDr), perhaps because the system gives access to the list of canonical sentences and helps them ask relevant questions.

4.4 Satisfaction

At the end of the task, doctors completed a questionnaire which confirms the quantitative results. Even if they felt constrained with both systems, they agreed that with BabelDr:

- they could ask enough questions to be sure about the diagnosis (only 1/9 negative opinion with babelDr vs 4/9 in GT)
- they were confident in the translation to the target language (1/9 negative opinion with BabelDr vs 8/9 in GT)

- they liked the way recognition results are presented (0/9 negative opinion vs 3/9 with GT).
- they could integrate the system in their everyday medical practice (1/9 negative opinion with BabelDr vs 5/9 with GT)

The participants had the same subjective perception of recognition quality with both systems. 3/9 participants think that they couldn't be recognized easily and 4/9 think they could. Others were neutral. In the post-experiment interviews, the doctors often mentioned the difficulty of expressing their questions as yes/no questions, as this is unusual in the anamnesis dialogue.

5 Conclusion

The data collected in this user study show that despite a very good speech recognition component, GT's translations are far less adequate and less comprehensible than BabelDr's. Along with the lower confidence expressed by the doctors in this system, this suggests that GT is not precise enough for the task, corroborating our first hypothesis. Despite this, GT allows some users, mainly the doctors, to reach a correct diagnosis. However, correct diagnoses were far more frequent with BabelDr.

This study has also provided insights into the suitability of a limited coverage phraselator such as BabelDr for this task. Although we observe more rejected interactions than for GT due to the rule-based approach, this was not perceived as particularly limiting by the users, who felt they could ask enough questions. This suggests that BabelDr is a promising tool for the task. Future enhancements of the system include training of a statistical recognizer and implementation of robust matching methods to reduce the number of failed speech interactions.

This experiment allowed us to collect a corpus of 18 diagnostic dialogues performed with two different tools, which can be used to study many different aspects of doctor-patient communication or for a shared task.

6 Acknowledgements

This project is financed by the "Fondation Privée des Hôpitaux Universitaires de Genève". We thank Manny Rayner for his comments on this paper.

References

- Ahmed, F., Bouillon P., Destefano, C., Gerlach, J., Hooper, A., Rayner, M., Strasly, I., Tsourakis, N. and Weiss, C. 2017. Rapid Construction of a Web-Enabled Medical Speech to Sign Language Translator Using Recorded Video. to appear in *Future and Emergent Trends in Language Technology*. Seville (Spain) - 2016, Springer.
- Aho, A.V. and Ullman, J.D. 1969. Properties of syntax directed translations. *Journal of Computer and System Sciences* 3, 3:319–334.
- Flores, Gl. et al. 2003. Errors in medical interpretation and their potential clinical consequences in pediatric encounters. *Pediatrics*.
- Fuchs, M., Tsourakis, N. and Rayner, M. 2006. A Scalable Architecture For Web Deployment of Spoken Dialogue Systems. *Proceedings of LREC 2012*, Istanbul, Turkey.
- Hacker, K., Anies, M., Folb, B.L. and Zallman, L. 2015. Barriers to health care for undocumented immigrants: a literature review *Risk Management Healthcare Policy*, 156:1108–1113.
- Jacobs, E.A., Sadowski, L.S. and Rathouz, P.J. 2003. The impact of an enhanced interpreter service intervention on hospital costs and patient satisfaction. *Journal of General Internal Medicine*, 22(2):306–311.
- Landis, J.R. and Koch, G.G. 1997. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159-174.
- Patil S. and Davies P. 2014. Use of Google Translate in medical communication: evaluation of accuracy *BMJ*, 349.
- Priebe, S., Sandhu, S., Dias, S et al. 2011. Good practice in health care for migrants: views and experiences of care professionals in 16 European countries. *BMC Public Health*, 11.
- Rayner, M., Baur, C., Chua, C., Bouillon, P. and Tsourakis, N. 2015. Helping Non-Expert Users Develop Online Spoken CALL Courses. *Proceedings of the Sixth SLATE Workshop*, Leipzig, Germany.
- Rayner, M., Armando, A., Bouillon, P., Ebling, S., Gerlach, J., Halimi, S. and Tsourakis, N. 2016. Helping Domain Experts Build Phrasal Speech Translation Systems. José F. Quesada et al.. *Future and Emergent Trends in Language Technology*. Seville (Spain) - 2015, Springer, 2016:41-52.
- Wassermann, W. et al. 2014. Identifying and Preventing Medical Errors in Patients with Limited English Proficiency: Key Findings and Tools for the Field. *Journal for Healthcare Quality*.
- Wu A.C., Leventhal J.M., Ortiz J., Gonzales E.E. and Forsyth B. 2006. The interpreter as cultural educator of residents. *Archive of Pediatric and Adolescent Medicine* 160, 1145–1150, 160:1145-1150.

Dissecting Human Pre-Editing toward Better Use of Off-the-Shelf Machine Translation Systems

Rei Miyata[†] Atsushi Fujita[‡]

[†]Nagoya University

Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
miyata@nuee.nagoya-u.ac.jp

[‡]National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Souraku-gun, Kyoto, 619-0289, Japan
atsushi.fujita@nict.go.jp

Abstract

Machine translation (MT) systems are not able to always produce translations of human-level quality. As a practical means of such MT systems, we investigated the potential of pre-editing strategy, by collecting actual pre-edit instances using a human-in-the-loop protocol. In our study, targeting Japanese-to-English translation on four different datasets and using an off-the-shelf MT system, we collected a total of 12,287 pre-edit instances for 400 source sentences and showed promising results; more than 85% of source sentences turned out to be accurately translated by the MT system. We also found that the pre-edited Japanese source sentences were better translated into Chinese and Korean, confirming the usefulness of pre-editing strategy in a multilingual setting. Through decomposing the collected pre-edit instances, we built a typology of primitive edit operations comprising 53 types, which unveils the subjects for further research.

1 Introduction

Given the improved quality of machine translation (MT) and the increased demand for rapid delivery of translations, a number of off-the-shelf MT systems have become available. However, none of them can guarantee that their raw outputs are always of sufficient quality. When we consider embedding such MT systems in computer-aided

translation (CAT) settings, it is indispensable to explore practical means to obtain high-quality translations without configuring the MT systems.

One option to make better use of such MT systems is to edit source text (ST) so that it is amenable to the targeted MT system, i.e., *pre-editing*. As demonstrated in the literature, pre-editing ST leads to improved MT quality (Bernth and Gdaniec, 2001; Miyata et al., 2015) and reduced post-editing effort (Pym, 1988; O’Brien and Roturier, 2007; Aikawa et al., 2007). Controlling ST is particularly effective in multilingual settings (Ó Broin, 2009).

Several studies have examined human-in-the-loop protocols that include pre-editing ST in order to improve MT quality. Uchimoto et al. (2006) have used back translation as a means to spot non-machine-translatable spans in ST, which are subsequently served to humans to be edited. Resnik et al. (2010) have taken advantage of monolingual human knowledge of the target language to identify spans of ST that are likely to cause translation errors. Mirkin et al. (2013) have devised an interactive tool for monolingual authors. It suggests appropriate alternatives along with confidence scores for MT outputs.

In this paper, we investigate the capability of the pre-editing strategy and provide an overview of possible edit operations used for pre-editing. First, we empirically demonstrate the potential usefulness of the pre-editing strategy, i.e., how often STs turn out to be accurately translated by a targeted MT system. To this end, we designed a human-in-the-loop protocol, in which human editors incrementally edit given STs (Section 2), and experimented with Japanese-to-English translation tasks on four different datasets (Section 3). Using the original and the best-edited STs, we also

© 2017 Rei Miyata and Atsushi Fujita. Licensed under the Creative Commons BY-ND 4.0 license. Some rights reserved. <https://creativecommons.org/licenses/by-nd/4.0/>

Status
 “In progress”
 “Complete”
 “Give up”

| No | 日本語文 オリジナル | Source text (ST) | 自動翻訳文 | MT output |
|----|------------|---------------------------|-------|--|
| 9 | | お住まいの市区町村の役所に出生届をご提出ください。 | | Please submit notification of birth to the public office of the municipality where you live. |

Translate Complete ユニット選択画面に戻る

| No | 書き換え履歴 | Versions of ST | 自動翻訳履歴 | Corresponding MT output |
|-----|--------|---------------------------|--------|--|
| ★ 8 | | お住まいの市区町村の役所に出生届をご提出ください。 | | Please submit notification of birth to the public office of the municipality where you live. |
| ★ 7 | | 出生届を、居住地の市区町村の役所にご提出ください | | Birth report, submit to the public office of the municipality where you live. |
| ★ 6 | | 居住地の市区町村の役所に、出生届をご提出ください。 | | The public office of the municipality of the place of residence, please submit birth certificates. |
| ★ 5 | | 居住地の役所に出生届をご提出ください。 | | Please submit birth registrations to the city hall of the residence. |
| ★ 4 | | 居住地の市区町村の役所に出生届をご提出ください。 | | Please submit birth registrations to the city hall of the city, ward, town or village of the place of residence. |
| ★ 3 | | 出生届を居住地の市区町村の役所にご提出ください。 | | Submit them to the office of city, ward, town or village of the place of residence birth certificates. |
| ★ 2 | | 出生届を居住地の市区町村の役所に提出してください | | Please submit it to the city hall of the city, ward, town or village of the place of residence registration of a new birth |
| ★ 1 | | 出生届を居住地の市区町村の役所に提出 | | Birth registrations submitted to the public office of the municipality where you live. |

Best version
 Child node(s)
 Parent node
 Original version

Figure 1: Our platform for collecting pre-edit instances: when an edited ST in the upper pane is submitted, it is registered with its MT output in a sequential order as shown in the bottom pane.

investigated the usefulness of the pre-edited STs in translating Japanese STs into Chinese and Korean (Section 4). To give an overall picture of pre-editing, we built a typology of edit operations upon actual *pre-edit instances*, i.e., pairs of STs before/after minimal pre-editing, collected through the above protocol followed by manual decomposition (Section 5). The typology can act as a guidepost to determine useful operations, such as those having the largest impact on the MT quality and those that are easy to automate.

2 Protocol for Collecting Pre-editing Instances

As in Miyata et al. (2015), we ask human editors to incrementally edit STs relying on their introspection, so that improved MT quality is achieved. Miyata et al. (2015) collected only the final versions of edited STs and directly compared them with the originals. In contrast, we aim to observe the trials and errors of editors and to achieve translations of satisfactory quality as much as possible. To that end, we developed a Web-based platform, shown in Figure 1, with the following two features.

- We record ST after every minimal edit is performed in order to capture the detailed process of pre-editing.
- We allow editors to resume editing from any given past version of ST in order to facilitate their trial and error.

Editors are asked to follow the iterative procedure given below for each original ST. We refer to the set of collected versions of STs for the same original ST as a *unit*.

- Step 1.** Assess the MT output for the present ST according to the 5-point scale criterion in Table 1. Go to Step 4, if it has satisfactory quality;¹ otherwise, go to Step 2.
- Step 2.** Select one version of ST to be edited, from the past versions of STs, referring to the corresponding MT outputs, and go to Step 3; if none is likely to achieve satisfactory quality even if edited, go to Step 4.
- Step 3.** Minimally edit the selected version of ST, while keeping the meaning of the ST, referring to the MT output for it, so that the MT system would be able to generate a better translation. When the edited ST is submitted, its MT output is automatically generated and registered together. Go back to Step 1.
- Step 4.** Choose one version of ST that achieves the best MT quality among all the versions registered in the unit (called the *Best ST*), and terminate the procedure for this ST.

To observe fine-grained edit operations, we instructed editors to make edits primitive as much as possible in Step 3, showing some examples. Table 2 shows an example; a phrase reordering for sentence (a) makes sentence (b), and a passiviza-

¹“Perfect” or “Good” quality in our criterion in Table 1.

| | |
|-----------------------|--|
| 5. Perfect | Information of the original text has been completely translated. There are no grammatical errors in the translation. Word choice and phrasing is natural even from a native speaker’s point-of-view. |
| 4. Good | Word choice and phrasing is slightly unnatural, but the information of the original text has been completely translated and there are no grammatical errors in the translation. |
| 3. Fair | There are some minor errors in the translation of less important information of the original text, but the meaning of the original text can be easily understood. |
| 2. Acceptable | Important parts of the original text are omitted or incorrectly translated, but the core meaning of the original text can still be understood with some efforts. |
| 1. Incorrect/nonsense | The meaning of the original text is incomprehensible. |

Table 1: Criterion for evaluating MT quality.

| |
|---|
| (a) 来院しなくても十日前後で登録のクレジットカードから引き落としを行います。 |
| (b) 来院しなくても登録のクレジットカードから十日前後で引き落としを行います。 |
| (c) 来院しなくても登録のクレジットカードから十日前後で引き落としが行われます。 |

Table 2: Examples of primitive edits on a Japanese sentence whose meaning is “You’ve registered your credit card. We will charge on that card in around 10 days regardless of your visit.”

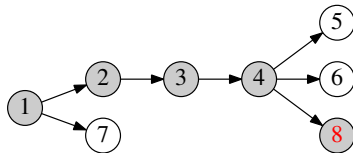


Figure 2: Tree representation of versions of STs shown in Figure 1.

tion of sentence (b) leads to sentence (c). In this case, the edit from sentence (a) to sentence (c) is not considered as primitive.

Also in Step 3, we prohibited editors from registering an ST identical to any past versions of ST. With this constraint, versions of ST in each unit form a tree structure, as illustrated in Figure 2. Each node comprises a version of ST accompanied by the MT output for it; the number in the node stands for the chronological order of the version in each unit, with one node, No. 8 in this example, labeled the Best ST. Every node, except for the original one (No. 1), is derived from a parent node. It is guaranteed that the path between the Best ST and the original one (henceforth, *best path*) in each unit, e.g., gray nodes in Figure 2, contains edit operations effective in improving MT quality.

3 Pilot Run

Using our protocol presented in Section 2, we collected versions of STs and pre-edit instances in Japanese-to-English translation of four sets of STs

in three domains: hospital conversation² (*hosp*), living information provided by municipalities³ (*muni*), and two types of news articles, Japanese-origin ones from BCCWJ⁴ (*bccwj*) and English-origin ones from Reuters⁵ (*reuters*). While *hosp* is spoken, the others are written; sentence length is markedly diverse (see also Table 3). These domains are so different from each other that we expect that the applicability of our proposed protocol can be evaluated from diverse points of view. For each dataset, we randomly sampled 100 Japanese sentences and used them as original STs.

As the off-the-shelf MT system, we used TexTra,⁶ a freely-available, state-of-the-art phrase-based statistical MT system, through its REST API. We assigned the pre-editing task to one native Japanese speaker who has a good command of English and ample experience in evaluating the quality of various types of MT systems according to the criterion in Table 1, while she has no prior knowledge of TexTra.

As a result, 12,687 versions of STs and thus 12,287 pre-edit instances were collected; see Table 3 for statistics. As shown in the rightmost column, more than 85% of the STs were ended with MT outputs of satisfactory quality. This demonstrates the high potential of the MT system when proper human intervention is incorporated. In general, the longer the original ST was, the more edit operations were required to attain satisfactory quality. Table 4 shows an example of the Best ST of a unit in *reuters*, which was obtained after 25 consecutive edits in the best path and the MT output of which met satisfactory quality.

²An in-house speech transcription corpus of conversational utterances in a hospital.

³Excerpts from websites of municipalities in Japan (Miyata et al., 2015).

⁴http://pj.ninjal.ac.jp/corpus_center/bccwj/

⁵http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/jea/reuters/

⁶<https://mt-auto-minhon-mlt.ucri.jgn-x.jp/>

| Dataset | Mode | Avg. num. of tokens in original ST (s.d.) | Num. of versions of ST | | | | Num. of units | |
|----------------|---------|--|------------------------|------|------|-----|---------------|----------------------|
| | | | Total | Avg. | Med. | Max | Original=Best | Satisfactory quality |
| <i>hosp</i> | spoken | 12.1 (4.5) | 1199 | 12.0 | 3 | 105 | 40 | 97 |
| <i>muni</i> | written | 21.3 (12.0) | 2119 | 21.2 | 14 | 89 | 3 | 97 |
| <i>bccwj</i> | written | 26.9 (16.0) | 3823 | 38.2 | 26 | 209 | 0 | 86 |
| <i>reuters</i> | written | 34.8 (12.6) | 5546 | 55.5 | 45 | 258 | 4 | 93 |

Table 3: Statistics of the collected data.

| | ST | MT output |
|-----------|---|--|
| Original | 同国は、前年の過剰輸出と、今年の減産によって、穀物不足に直面しており、大量の小麦輸入の計画を表明している。 | Excess exports in the previous year, and reduced production this year, is facing a shortage of grain, a large amount of wheat imports plan. |
| Best | 当年の減産と前年の過剰輸出による穀物の不足をふまえ、この国は小麦を大量に輸入する計画を表明している。 | Based on the shortage of grain due to production cuts in the current year and excessive exports last year, this country has announced plans to import a large amount of wheat. |
| Reference | The country, currently battling an acute grain shortage due to excessive exports last year, faces a poor harvest this year and intends to import large quantities of wheat. | |

Table 4: An example of Best ST with satisfactory MT quality.

| | ST | MT output |
|-----------|---|---|
| Original | WSCによると、4日には弱い複数の降雨の可能性があるので、5-6日には全般に乾燥した天候が戻る見通し。 | WSC, although the possibility of weak more rainfall within 4 days, the weather in general dry return to 5-6 days. |
| Best | WSCによると、4日には弱い降雨の可能性が存在する一方で、5日から6日にかけては、乾燥した天候が全般に戻ってくる見込み、とのこと。 | WSC said, while the possibility of a weak rain exists on June 4 , from June 5 to 6, the dry weather comes back, in general. |
| Reference | WSC said the outlook was for a chance of a few light showers on 4th, and generally dry conditions on 5th and 6th. | |

Table 5: An example of Best ST for which our protocol cannot achieve satisfactory MT quality.

It should also be noted that 27 out of 400 units did not attain satisfactory quality in our human-in-the-loop protocol. Among these “Give up” cases, we identified that mis-translation of proper nouns and incorrect lexical choices were the most difficult types of MT errors to rectify. For example, the Best ST in Table 5 contains expressions for dates, “4日,” “5日,” and “6日,” proper translations of which are “4th,” “5th,” and “6th,” respectively. The MT system specified “June” improperly. This error stems from the wrong phrase alignment in the statistical model. These types of errors should be addressed during training the models and/or post-editing, rather than pre-editing. Our protocol enables us to identify MT errors that are difficult to amend only by the pre-editing strategy. This will eventually help us streamline the overall translation workflow using off-the-shelf MT systems.

4 Machine Translatability for Different Languages

We examined the effectiveness of the pre-editing strategy in a multilingual translation setting, i.e., whether an ST, edited so that it is better translated into one target language, can also be better translated into other languages. First, all the original

and the Best STs in the four datasets (800 sentences in total) were translated into Chinese and Korean using the corresponding models of TexTra. Then, for each set of Chinese and Korean translations, one human evaluator was asked to assess the MT quality using the 5-point scale in Table 1.

As shown in Table 6, the MT quality for the Best STs was, on average, higher than that for the original STs for all the datasets, indicating that edit operations that improved English-translatability of Japanese STs are portable to Chinese- and Korean-translatability to a certain degree. For both languages, the MT quality for the Best STs in *hosp* and *bccwj* well surpassed that for the original STs, while there were no significant improvements in *muni* and *reuters*. Further scrutiny into the language dependency of machine-translatability is important to justify the pre-editing approach to other target languages and domains.

5 Typology of Edit Operations

We analyzed the diversity of edit operations exhibited during our pre-editing exercise. As mentioned in Section 2, it is likely that the best path contains edit operations effective in improving MT quality. We therefore focused on pre-edit instances in the

| Chinese | Avg. score | | Num. of units (Org vs. Best) | | |
|----------------|------------|--------|------------------------------|----|----|
| | Org | Best | > | = | < |
| <i>hosp</i> | 2.73 | 2.93** | 7 | 70 | 23 |
| <i>muni</i> | 2.84 | 2.89 | 32 | 31 | 37 |
| <i>bccwj</i> | 2.39 | 2.75** | 13 | 42 | 45 |
| <i>reuters</i> | 2.61 | 2.77 | 22 | 45 | 33 |

| Korean | Avg. score | | Num. of units (Org vs. Best) | | |
|----------------|------------|--------|------------------------------|----|----|
| | Org | Best | > | = | < |
| <i>hosp</i> | 3.32 | 3.56** | 12 | 57 | 31 |
| <i>muni</i> | 3.58 | 3.67 | 32 | 29 | 39 |
| <i>bccwj</i> | 3.37 | 3.60* | 18 | 47 | 35 |
| <i>reuters</i> | 3.31 | 3.36 | 24 | 47 | 29 |

Table 6: Results of human evaluation of MT quality: “*” and “**” indicate significant differences over “Org(inal ST)” tested by Wilcoxon signed-rank test with $p < 0.05$ and $p < 0.01$, respectively.

| Dataset | Num. of instances in best path | | |
|----------------|--------------------------------|----------------|---------|
| | (a) raw | (b) decomposed | (b)/(a) |
| <i>hosp</i> | 97 | 185 | 1.91 |
| <i>muni</i> | 106 | 186 | 1.75 |
| <i>bccwj</i> | 174 | 340 | 1.95 |
| <i>reuters</i> | 191 | 268 | 1.40 |

Table 7: The number of decomposed pre-editing instances in the best path of 10 sampled units.

best path of 10 randomly sampled units for each of the four datasets. First, we decomposed each of the sampled 568 instances into a sequence of primitive edit operations, because our editor might not strictly seek the primitiveness. Indeed, as shown in Table 7, this process increased the number of instances by from 1.40 to 1.95 times, resulting a total of 979 instances of primitive edit operations. We then manually created a typology of edit operations, by categorizing each instance, regarding surface-level differences of each pair of STs as clues. Table 8 shows the resulting typology with 53 types of edit operations that cover all of the analyzed pre-edit instances, with their frequency in each dataset. We observed an extended variety of edit operations in our collection, ranging from ones at surface-level, such as insertion/deletion of punctuation and word reordering, to various types of syntactic alternation.

The most frequent type across the datasets was **C01 (Alternative lexical choice)**, including edit operations such as replacing “一度” with “一回” (both mean *once*), and “習得する” (*acquire*) with “学ぶ” (*learn*). This type of edit operations would be automated by constructing lexical resources tailored to particular MT systems. We also identified several frequent types of edit operations that are

likely to be effective for improving MT quality. For example, **S05 (Phrase reordering)** and **S07 (Insertion/deletion of punctuation)** can help MT systems parse the input sentences correctly, which subsequently leads to better MT outputs.

Some types of edit operations were observed only in specific domains. For example, we observed **S15 (Use/disuse of clause-ending noun)** and **S20 (Use/disuse of nominal/verbal suffix)** only in the news domain (*bccwj* and *reuters*). Both types reflect the fact that the elliptic expressions often used in news articles could degrade the MT quality. Our method is also useful to unveil these kinds of domain-specific issues.

Last but not least, let us describe a less frequent type of edit operations, i.e., **S13 (Head-switching of verb phrase)**:

[before] 懸念を強め (*strengthen anxiety*)

[after] 強い懸念を抱き (*have strong anxiety*)

This type of edit operation has not been covered by existing controlled language rule sets in Japanese, such as (Ogura et al., 2010; Hartley et al., 2012; Miyata et al., 2015), nor even by a comprehensive typology of paraphrases.⁷ It is worth exploring to what extent these types of edit operations are effective in improving MT quality.

6 Conclusion and Future Work

In this paper, we have presented our human-in-the-loop protocol for collecting pre-edit instances. Using this protocol, we collected 12,287 pre-edit instances for four different datasets, demonstrating that most of the source sentences can be edited into machine-translatable ones. Human evaluation revealed that, for some datasets, English-translatable Japanese STs significantly improved the quality of translations into Chinese and Korean. We also built a typology comprising a wide range of edit operations, and found that alternating lexical choice was the most frequent one taken by our editor.

Based on this study, we plan to develop an automatic pre-editor. One approach to this is controlled language formulation by assessing the effectiveness of each type of edit operation (Bernth and Gdaniec, 2001; Miyata et al., 2015). Another is to build a statistical model. It is worth investigating data-driven methods based on our collection of pre-edit instances, although this data do not guar-

⁷<http://paraphrasing.org/paraphrase.html>

| ID | Type | Frequency | | | |
|-----|--|-----------|----|----|----|
| | | H | M | B | R |
| S01 | Division/synthesis of sentence(s) | 4 | 1 | 7 | 2 |
| S02 | Use of line break | 0 | 3 | 0 | 0 |
| S03 | Use of compound/complex sentence | 0 | 0 | 0 | 1 |
| S04 | Split of phrase | 0 | 0 | 1 | 0 |
| S05 | Phrase reordering | 24 | 6 | 22 | 13 |
| S06 | Insertion/deletion of subject | 0 | 2 | 2 | 2 |
| S07 | Insertion/deletion of punctuation | 24 | 5 | 27 | 27 |
| S08 | Change of scope of subject | 0 | 0 | 1 | 1 |
| S09 | Use of nominative case “ga” or topic marker “wa” | 0 | 1 | 3 | 2 |
| S10 | Change of marked element | 0 | 2 | 11 | 0 |
| S11 | Change of voice | 3 | 1 | 13 | 3 |
| S12 | Change of restrictive/continuous modification | 2 | 0 | 12 | 13 |
| S13 | Head-switching of verb phrase | 0 | 0 | 0 | 3 |
| S14 | Indication of conditional clause | 2 | 7 | 2 | 0 |
| S15 | Use/disuse of clause-ending noun | 0 | 1 | 3 | 5 |
| S16 | Change of subject in noun phrase | 0 | 0 | 1 | 0 |
| S17 | Use of noun phrase or verb phrase | 3 | 4 | 9 | 0 |
| S18 | Use/disuse of compound verb | 2 | 0 | 2 | 0 |
| S19 | Use/disuse of compound noun | 2 | 7 | 5 | 8 |
| S20 | Use/disuse of nominal/verbal suffix | 2 | 1 | 10 | 5 |
| S21 | Change of connective expression | 6 | 16 | 12 | 13 |
| S22 | Change of parallel expression | 2 | 3 | 1 | 0 |
| S23 | Change of apposition expression | 0 | 0 | 0 | 5 |
| S24 | Change of specification expression | 0 | 0 | 0 | 3 |
| S25 | Change of locative expression | 0 | 0 | 0 | 2 |
| S26 | Change of hearsay expression | 0 | 0 | 0 | 4 |
| S27 | Change of expression for indirect question | 0 | 0 | 0 | 1 |
| S28 | Change of <i>sahen</i> noun expression | 1 | 2 | 7 | 4 |
| S29 | Change of formal noun expression | 0 | 1 | 3 | 5 |
| S30 | Change of substantive verb expression | 1 | 0 | 0 | 1 |
| S31 | Change of <i>ni-to-naru</i> expression | 0 | 0 | 0 | 11 |
| C01 | Alternative lexical choice | 29 | 36 | 69 | 33 |
| C02 | Lexical elaboration | 5 | 3 | 2 | 1 |
| C03 | Lexical simplification | 0 | 5 | 0 | 0 |
| C04 | Change of reference expression | 0 | 0 | 0 | 1 |
| C05 | Use of redundant expression | 0 | 1 | 0 | 1 |
| F01 | Use of honorific expression | 19 | 11 | 14 | 4 |
| F02 | Change of tense | 0 | 3 | 1 | 2 |
| F03 | Change of conjunctive word | 4 | 4 | 0 | 1 |
| F04 | Change of auxiliary verb | 1 | 0 | 0 | 0 |
| F05 | Insertion/deletion of particle | 4 | 9 | 24 | 9 |
| F06 | Use of particle | 4 | 3 | 3 | 10 |
| F07 | Use of compound particle | 0 | 1 | 1 | 5 |
| T01 | Change of named entity | 0 | 0 | 3 | 6 |
| O01 | Orthographical change | 1 | 7 | 7 | 4 |
| O02 | Change of sentence-ending expression | 0 | 1 | 2 | 0 |
| O03 | Insertion/deletion/change of symbol | 0 | 6 | 0 | 0 |
| O04 | Insertion of omitted element | 0 | 0 | 3 | 2 |
| O05 | Specification of chunk with brackets | 0 | 5 | 3 | 1 |
| I01 | Change of content | 18 | 20 | 27 | 16 |
| I02 | Change of nuance | 0 | 7 | 17 | 6 |
| E01 | Grammatical errors | 3 | 1 | 4 | 6 |
| E02 | Other errors | 19 | 0 | 6 | 6 |

Table 8: Our typology of edit operations (H: *hosp*, M: *muni*, B: *bccwj*, R: *reuters*): The first letter of ID indicates seven major categories: **S** (Structure), **C** (Content word), **F** (Functional word), **T** (Terminology), **O** (Orthography), **I** (Information), and **E** (Edit that causes/resolves error in ST).

antee to improve MT quality as directly addressed by post-editing (Simard et al., 2007).

Acknowledgments

We are deeply grateful to the anonymous reviewers for their valuable comments on the earlier version of this paper. This work was partly supported by JSPS KAKENHI Grant Numbers 25730139 and 25240051. One of the corpora used in our study was created under a MIC program “Promotion of Global Communications Plan: Research, Development, and Social Demonstration of Multilingual Speech Translation Technology.”

References

- Aikawa, Takako, Lee Schwartz, Ronit King, Monica Corston-Oliver, and Carmen Lozano. 2007. Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In *Proc. of MT Summit*, pages 1–7.
- Bernth, Arendse and Claudia Gdaniec. 2001. MTranslatability. *Machine Translation*, 16(3):175–218.
- Hartley, Anthony, Midori Tatsumi, Hitoshi Isahara, Kyo Kageura, and Rei Miyata. 2012. Readability and translatability judgments for ‘Controlled Japanese’. In *Proc. of EAMT*, pages 237–244.
- Mirkin, Shachar, Sriram Venkatapathy, Marc Dymetman, and Ioan Calapodescu. 2013. SORT: An interactive source-rewriting tool for improved translation. In *Proc. of ACL: System Demonstrations*, pages 85–90.
- Miyata, Rei, Anthony Hartley, Cécile Paris, Midori Tatsumi, and Kyo Kageura. 2015. Japanese controlled language rules to improve machine translatability of municipal documents. In *Proc. of MT Summit*, pages 90–103.
- O’Brien, Sharon and Johann Roturier. 2007. How portable are controlled language rules? A comparison of two empirical MT studies. In *Proc. of MT Summit*, pages 345–352.
- Ó Broin, Ultan. 2009. Controlled authoring to improve localization. *Multilingual*, Oct./Nov.:12–14.
- Ogura, Eri, Mayo Kudo, and Hideo Yanagi. 2010. Simplified technical Japanese: Writing translation-ready Japanese documents. *IPSJ SIG Technical Reports*, 2010-DD-78(5):1–8. (in Japanese).
- Pym, Peter. 1988. Pre-editing and the use of simplified writing for MT: An engineer’s experience of operating an MT system. In *Proc. of Translating and the Computer 10*, pages 80–96.
- Resnik, Philip, Olivia Buzek, Chang Hu, Yakov Kronrod, Alex Quinn, and Benjamin B. Bederson. 2010. Improving translation via targeted paraphrasing. In *Proc. of EMNLP*, pages 127–137.
- Simard, Michel, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Proc. of NAACL-HLT*, pages 508–515.
- Uchimoto, Kiyotaka, Naoko Hayashida, Toru Ishida, and Hitoshi Isahara. 2006. Automatic detection and semi-automatic revision of non-machine-translatable parts of a sentence. In *Proc. of LREC*, pages 703–708.

Bootstrapping Quality Estimation in a live production environment

Joachim Van den Bogaert

CrossLang
Gent, Belgium
joachim@crosslang.com

Bram Vandewalle

CrossLang
Gent, Belgium
bram.vandewalle@crosslang.com

Roko Mijic

CrossLang
Gent, Belgium
roko.mijic@crosslang.com

Abstract

In this paper, we discuss how we bootstrapped Quality Estimation (QE) in a constrained industry setting. No post-edits were at our disposal and only a limited number of annotators was available to provide training data in the form of Post-Edit (PE) effort judgments. We used a minimal approach and applied a simplified annotation procedure. We used as few as 17 baseline features for QE training.

1 Introduction

In this paper, we discuss how we bootstrapped Quality Estimation (QE) – the process of scoring Machine Translation (MT) output without access to a reference translation – for 9 language pairs and 3 domains in a constrained industry setting. No post-edits were at our disposal and only a limited number of annotators was available to provide training data in the form of Post-Edit (PE) effort judgments. We used a minimal approach (Callison-Burch et al., 2009), by annotating only 800 segments per language pair and content type, and applying a simplified annotation procedure. We used as few as 17 baseline features (Specia et al., 2009b) for QE training.

As the project progressed, post-edits became available, allowing us to validate our approach and further develop the bootstrapped system, using off-the-shelf PE distance (TER) as training labels. We added syntactic and web-scale Language Model (LM) features (Kozlova et al., 2016), (Andor, et al., 2016) to improve a second

iteration of the QE system and trained on 80,000 PE distance labels to compare our results.

Finally, we roughly estimated the number of sentences needed for training a PE distance-based system that performs on par with a PE effort-based system.

2 Use case and related work

2.1 Use case

In Language Industry, Quality Estimation is used to filter out low-quality translations for post-editors, when they review Machine Translated texts (Specia et al., 2009b). This is important, because bad translations not only cause extra work (it is sometimes easier to translate from scratch (Specia, 2011)), they are also a source of frustration and negatively impact the image and acceptance of MT among translators (Wisniewski et al., 2013).

To alleviate these problems, we investigated the use of Quality Estimation for 9 language pairs (EN-DE, DE-EN, EN-FR, EN-RU, EN-ZH, EN-PT, EN-ES, EN-IT, EN-JP) and 3 domains (referred to as DOM1, DOM2 and DOM3). Since the MT engines were not cleared for use at the time the project began, no post-edits were available and a staged approach was required.

For production use, we are mainly interested in best practices (rather than in developing the best possible general-purpose QE system) and in deploying the system as quickly as possible with acceptable costs. This greatly differs from an academic setting, in which the exploration of Machine Learning algorithms and metrics, as well as the discovery of novel features are the main focus (see for example (Specia & Soricut, 2013)).

2.2 Related work

In industry, QE (also known as “Confidence Estimation (CE)” (Specia, 2011), (Blatz, et al., 2004) is most often used in sentence-based tasks, because all major translation environments use sentences as the basic units of work. For this reason, word-based (see for example (Blatz, et al., 2004), (Ueffing & Ney, 2005)) or document-based QE (see for example (Soricut & Echiabi, 2010)) were not considered, although they are useful in, respectively, the development of interactive MT systems, and document ranking for obtaining consistent high-quality output. The foundations of the work performed are described in (Callison-Burch et al., 2009), (Callison-Burch, et al., 2012) and (Specia et al., 2009b). We use their baseline system with the 17 features they describe.

3 Approach

Our approach differs in the way data collection is set up, and in the fact that we use PE effort judgments, although PE distance has been favored since the WMT 2013 campaign (Bojar, et al., 2013).

PE effort judgments were expressed according to the scores of (Callison-Burch, et al., 2012):

1. The MT output is incomprehensible, with little or no information transferred accurately. It cannot be edited, needs to be translated from scratch.
2. About 50-70% of the MT output needs to be edited. It requires a significant editing effort in order to reach publishable level.
3. About 25-50% of the MT output needs to be edited. It contains different errors and mistranslations that need to be corrected.
4. About 10-25% of the MT output needs to be edited. It is generally clear and intelligible.
5. The MT output is perfectly clear and intelligible. It is not necessarily a perfect translation but requires little or no editing.

The collection procedure outlined in WMT 2009 (Callison-Burch et al., 2009) was simplified as follows:

- By lack of post-edit data, neither high-quality targeted or hTER-optimized (Snover et al., 2006) post-edits were presented during annotation.
- No reference translation was presented – only the source sentence and MT output were displayed during annotation. Initial

experiments showed that scores were assigned in too narrow a band when reference translations were provided. This potentially hurts QE performance, so we decided not to show them.

- We did not measure intra-annotator agreement, since we were dealing with professional translators, who are expected to perform similar tasks on a regular basis. Note that we intend not to discard any data.
- The obtained data was weighted according to the scheme in (Callison-Burch, et al., 2012): more weight was given to judges with higher standard deviation from their own mean score to obtain a more even spread in the range [1, 5].

We used the following metrics to evaluate our data sets and QE systems:

- Fleiss’ coefficient (Fleiss, 1971), a generalization of Cohen’s kappa to multi-raters (Wisniewski et al., 2013) to measure the degree of agreement between annotators.
- Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), standard metrics for regression, quantifying the amount by which the estimator differs from the true score (Specia et al., 2009a) (Wisniewski et al., 2013)
- Pearson’s correlation, to express the linear correlation between predicted PE effort/PE distance and reference PE effort/PE distance.
- TER (Snover et al., 2006), to calculate the number of edits required to change a hypothesis translation into a reference translation.

Furthermore, we use our own proprietary software for feature extraction (based on (Eckart de Castilho & Gurevych, 2014)), and a LIBSVM epsilon-SVR with a Radial Basis Function Kernel, based on (Bethard et al., 2014).

Subsequent development of web-scale LM features is based on (Kozlova et al., 2016), the use of syntactic features is based on (Kozlova et al., 2016) and (Andor, et al., 2016).

4 Development of the baselines

4.1 MT Systems

The Machine Translation systems for which we develop QE, are based on Moses SMT (Koehn, et al., 2007), and on the work of (Neubig, 2013), and

| DOMAIN | DE-EN MAE/MRSE | | EN-DE MAE/MRSE | | EN-ZH MAE/MRSE | | EN-ES MAE/MRSE | | EN-PT MAE/MRSE | | EN-FR MAE/MRSE | | EN-IT MAE/MRSE | | ALL MAE/MRSE | |
|--------|-------------------|------|-------------------|------|-------------------|------|-------------------|------|-------------------|------|-------------------|------|-------------------|------|-----------------|------|
| | | | | | | | | | | | | | | | | |
| DOM1 | 0.65 | 0.88 | 0.68 | 0.88 | - | - | - | - | - | - | - | - | - | - | 0.73 | 0.97 |
| DOM2 | 0.54 | 0.86 | 0.94 | 1.16 | 0.79 | 1.06 | 0.63 | 0.98 | 0.77 | 0.99 | 0.54 | 0.76 | 0.62 | 0.87 | 0.76 | 1.03 |
| DOM3 | - | - | 0.80 | 1.05 | 0.68 | 0.95 | 0.54 | 0.85 | 0.86 | 1.10 | 0.63 | 0.95 | - | - | 0.79 | 1.03 |
| LANG | 0.63 | 0.90 | 0.80 | 1.03 | 0.70 | 0.97 | 0.52 | 0.83 | 0.76 | 1.02 | 0.55 | 0.80 | 0.62 | 0.87 | 0.77 | 1.04 |
| BULK | 0.77 | | | | | | | | 1.04 | | | | | | | |

Table 3: QE test results

(Bisazza et al., 2011). The systems use extensive normalization, segmentation and classification routines, as well as some syntactic features. Since the focus is on QE, we will not go into further detail, but we list the data set sizes (number of unique sentence pairs) to give a general idea of the MT systems’ potential output quality (see Table 1).

The domains consist of software-related materials, written in three distinctive styles. We will refer to them as DOM1, DOM2 and DOM3. DOM1 consists of solution descriptions, written by development and/or support staff, DOM2 relates to published documentation, DOM3 is intended for software training.

| DOMAIN | DOM1 | DOM2 | DOM3 |
|--------|-----------|------------|-----------|
| DE-EN | 2,613,489 | 22,375,900 | - |
| EN-DE | 2,971,501 | 13,838,326 | 1,154,653 |
| EN-ZH | - | 2,557,042 | 439,980 |
| EN-ES | - | 3,456,275 | 366,423 |
| EN-PT | - | 2,942,499 | 298,687 |
| EN-FR | - | 4,944,361 | 343,352 |
| EN-RU | - | 2,108,723 | 455,203 |
| EN-IT | - | 3,198,050 | - |
| EN-Jp | 878,036 | 4,915,823 | 533,053 |

Table 1: training set sizes MT systems

4.2 Data collection

The number of segments selected for each language pair is listed in Table 2. For DOM1 we only have 3 data sets and MT systems, but it is the only domain for which Post-Edits were available at the time of writing (see validation in section 5).

For each cell in the table, annotations from 3 translators were collected. Average inter-annotator agreement was at a level of 0.44 (Fleiss’ coefficient) and can be considered *fair* according to (Landis & Koch, 1977).

| DOMAIN | DOM1 | DOM2 | DOM3 | TOTAL |
|--------|------|------|------|-------|
| DE-EN | 800 | 800 | - | 1,600 |
| EN-DE | 800 | 800 | 800 | 2,400 |
| EN-ZH | - | 800 | 800 | 1,600 |
| EN-ES | - | 800 | 800 | 1,600 |
| EN-PT | - | 800 | 800 | 1,600 |
| EN-FR | - | 800 | 800 | 1,600 |
| EN-RU | - | 800 | 800 | 1,600 |
| EN-IT | - | 800 | - | 800 |
| EN-Jp | 800 | 800 | 800 | 2,400 |

Table 2: training set sizes (PE Effort) QE systems

4.3 Results

The MAE and MRSE for the resulting systems are listed in Table 3. We tried several combinations of the data to find the optimum set of models:

- for each data set, *language + domain-specific* models were trained (listed in the white columns)
- *language-specific* models (LANG row) were trained by combining all data available for each language pair.
- language agnostic *domain-specific* models were trained by aggregating all data for each domain separately (ALL column in grey).
- finally, a language-agnostic **BULK** model (BULK row in grey), with all available data was trained.

The **BULK** model and the *domain-specific* models perform roughly on par, but in almost all cases, they are outperformed by the *language-specific* and *language + domain-specific* models. Which is what we expected, but we wanted to know whether it would be operationally feasible to train one single model or one model per domain.

In terms of performance, it is not clear which strategy, *language-specific* or *language + domain-specific*, to select. From a systems management perspective though, having one *language-specific* system for each language pair reduces deployment complexity immensely, with only a minor decrease in performance as trade-off (except for the EN-DE language pair).

5 Validation of the approach

As stated in section 1, we fell back to the 2009 WMT protocol (Callison-Burch et al., 2009) by lack of PE data. We surmised that a prohibitive number of Post-Edits would be required to obtain acceptable QE performance, so only 800 segments (per domain and language-pair) were sent out for PE effort judgment (to 3 annotators) to remain within budget. If we assume – for the sake of simplicity – that annotating a sentence with a PE effort judgment and post-editing a sentence are

equally expensive, then we expect our bootstrapped *language + domain-specific* systems to outperform QE systems trained on three times as many PE distance labels (2,400 data points).

Figure 1 summarizes and extrapolates the number of data points it takes to obtain comparable correlations. The graphs clearly indicate that more than triple the data is required to get comparable QE performance. For EN-DE, we were able to obtain around 80,000 post-edits. Even with this relatively large data set, the baseline PE distance-based QE system does not achieve the quality we get from a PE effort-based system.

This corroborates our intuition that – starting with almost no data – it pays off to consider PE effort-based solutions when developing a baseline. Obviously, it would go too far to state that using PE effort should be the preferred, authoritative (Callison-Burch et al., 2009) approach, because there are too many intrinsic shortcomings to adopt it as a best practice. For example, the Pearson correlation we used to compare PE effort-based and PE-distance based QE, expresses the extent to which a predicted entity (PE effort or PE distance) has a linear relationship with *some* hidden variable. For all we know, this hidden variable may be *sentence length*, instead of Post-Edit quality. There is also the issue of *subjectivity* at the annotator side. PE distance eliminates subjectivity, and can thus be expected to yield more consistent results. We believe however, that the use of professional translators filtered out a lot of the noise that can be observed in the WMT campaigns.

Conversely, the extrapolation gives us an idea about how many sentence pairs are needed to build a system that performs on par with PE effort-based QE, using Post-Edits exclusively. This opens possibilities when training MT and QE systems in a data-rich (MT training data > 1M sentence pairs) environment. It would be interesting to investigate whether an optimum split can be achieved to divide the data into a larger part that

is used to train MT systems with, and a smaller part that can be used to generate *pseudo* post-edits (the PE distance between reference and MT-generated hypothesis would be measured). The aim would be to maximize QE quality while minimizing MT quality loss. With the available data set, the use of *real* Post-Edits versus *pseudo* Post-Edits could be compared to validate such approach.

| | SYSTEM | MAE | PEARSON CORRELATION |
|-------|-----------|---------------|---------------------|
| EN-IT | BASELINE | 0.628+/-0.029 | 0.460+/-0.050 |
| | +OOVs+WLM | 0.631+/-0.026 | 0.463+/-0.027 |
| EN-FR | BASELINE | 0.543+/-0.024 | 0.367+/-0.028 |
| | +OOVs+WLM | 0.549+/-0.017 | 0.354+/-0.009 |
| EN-PT | BASELINE | 0.766+/-0.012 | 0.416+/-0.010 |
| | +OOVs+WLM | 0.763+/-0.010 | 0.422+/-0.021 |
| DE-EN | BASELINE | 0.597+/-0.014 | 0.486+/-0.015 |
| | +OOVs+WLM | 0.597+/-0.012 | 0.484+/-0.032 |
| EN-RU | BASELINE | 0.624+/-0.015 | 0.335+/-0.030 |
| | +OOVs+WLM | 0.624+/-0.006 | 0.336+/-0.018 |
| EN-ES | BASELINE | 0.525+/-0.018 | 0.293+/-0.022 |
| | +OOVs+WLM | 0.522+/-0.018 | 0.304+/-0.012 |
| EN-JP | BASELINE | 0.699+/-0.012 | 0.526+/-0.013 |
| | +OOVs+WLM | 0.719+/-0.012 | 0.499+/-0.014 |
| EN-DE | BASELINE | 0.800+/-0.013 | 0.514+/-0.019 |
| | +OOVs+WLM | 0.794+/-0.009 | 0.520+/-0.006 |
| EN-ZH | BASELINE | 0.655+/-0.008 | 0.586+/-0.013 |
| | +OOVs+WLM | 0.657+/-0.007 | 0.586+/-0.003 |
| AVG. | BASELINE | 0.649+/-0.006 | 0.442+/-0.008 |
| | +OOVs+WLM | 0.651+/-0.005 | 0.441+/-0.006 |

Table 4: comparison with and without OOVs and Web-scale LM

6 Additional features

Having obtained acceptable performance with a basic feature set, we added three features/feature sets to improve our models: technical OOVs, web-scale Language Models (WLMs) and SyntaxNet features.

6.1 Technical OOVs

When applying QE to real-life data, we expect the presence of technical OOVs (Fishel & Sennrich, 2014) to hurt performance for the following reasons: (1) usually, technical OOVs are not modelled in the MT system’s translation and language model, instead they are normalized or treated as OOVs to be copied verbatim into the target. If this

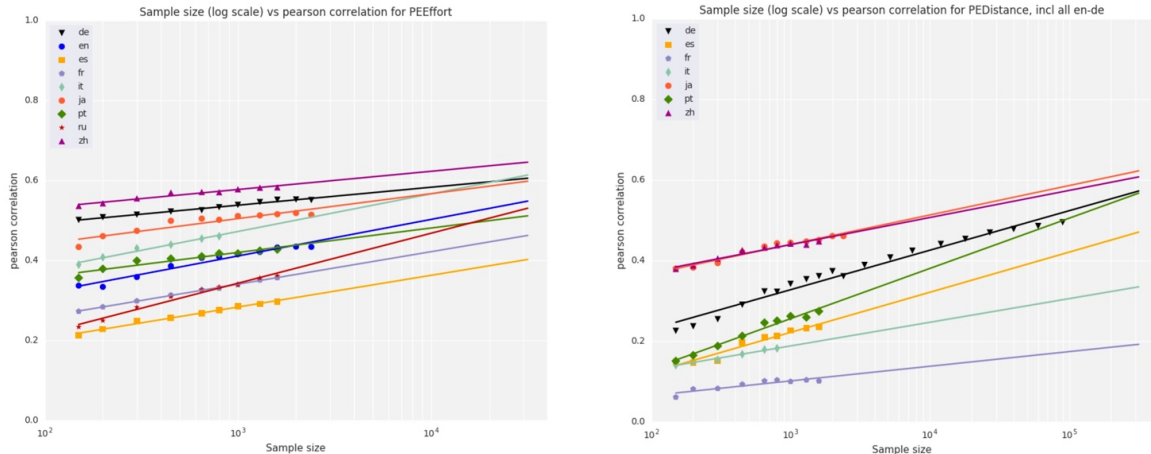


Figure 1: extrapolation of required PE distance labels for comparable performance

behaviour is not compensated for by the QE system, sentences with technical OOVs will unrightfully receive a penalty at lookup time; (2) in addition, technical OOVs, require a simple copy operation (if not resolved by the MT system), which makes the task of sentences containing OOVs easier, instead of more difficult

We use a custom-made classifier learnt from manually annotated data, and pre-processed with manually constructed rules (Kluegl et al., 2016), to annotate the training data.

6.2 Web-scale Language Models

We further experimented with Web-scale Language Models, as described in (Kozlova et al., 2016). We use public data (mostly Wikipedia) and collect around 48M sentences for English. The obtained gains are rather poor, probably because our language models are already quite big, and the extra out-of-domain data only adds little information.

6.3 SyntaxNet features

As a final experiment, we parsed our data with SyntaxNet (Andor, et al., 2016) and followed the approach outlined by (Kozlova et al., 2016). We use their tree-based features, as well as their features derived from Part-Of-Speech (POS) tags and dependency roles. Experiments were run on the EN-DE PE distance data set, because it was the only data set we had available at that time.

Our final results are listed in Table 5. The quality jump obtained (7,000 vs. 70,000), and the increasing difference between baseline (technical OOVs included for source and target) and best system, indicate that – in the long run – PE distance based QE remains worthwhile pursuing.

| SAMPLE SIZE | FEATURES SET | # | MAE | PEARSON CORRELATION |
|-------------|----------------|----|-----------------|---------------------|
| 700 | BASILINE | 19 | 0.269 +/- 0.003 | 0.258 +/- 0.017 |
| | + SYNTAX | 43 | 0.264 +/- 0.001 | 0.318 +/- 0.005 |
| | + SYNTAX + WLM | 45 | 0.267 +/- 0.002 | 0.309 +/- 0.012 |
| 7,000 | BASILINE | 19 | 0.241 +/- 0.001 | 0.432 +/- 0.005 |
| | + SYNTAX | 43 | 0.237 +/- 0.001 | 0.459 +/- 0.002 |
| | + SYNTAX + WLM | 45 | 0.236 +/- 0.001 | 0.460 +/- 0.004 |
| 70,000 | BASILINE | 19 | 0.229 +/- 0.001 | 0.504 +/- 0.002 |
| | + SYNTAX | 43 | 0.219 +/- 0.001 | 0.548 +/- 0.002 |
| | + SYNTAX + WLM | 45 | 0.217 +/- 0.001 | 0.556 +/- 0.002 |

Table 5: final results on the EN-DE PE distance data set

7 Discussion and future work

We have described the development of QE systems with no access to post-edit data. While mainly building on the work previously done in the QE field, our contribution consists of the de-

velopment of *a method to quickly build QE systems with minimal resources and a simplified annotation scheme*. We observed that using around 100k PE distance labels can produce a QE system that correlates equally strong with PE quality as a PE effort-based system trained on 800 sentence pairs. This is valuable information, as it allows for budget planning and opens opportunities to use *pseudo* Post-Edits instead of real Post-Edits.

In the future, we plan to investigate the use of such *pseudo* Post-Edits and describe a method to obtain an optimum trade-off between MT quality and PE quality when operating in data-rich environments. We will also further develop the syntax-based features, using the +40 parsers that are made available through the SyntaxNet project.

Acknowledgements

The authors wish to thank the reviewers for their helpful suggestions.

References

- Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., . . . Collins, M. (2016). Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*.
- Bethard, S., Ogren, P., & Becker, L. (2014). ClearTK 2.0: Design Patterns for Machine Learning in UIMA. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 3289-3293). Reykjavik, Iceland: European Language Resources Association (ELRA).
- Bisazza, A., Ruiz, N., & Federico, M. (2011). Fill-up versus interpolation methods for phrase-based SMT adaptation. *IWSLT*, (pp. 136-143).
- Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., . . . Ueffing, N. (2004). Confidence estimation for machine translation. *Proceedings of the 20th international conference on Computational Linguistics* (pp. 315-321). Association for Computational Linguistics.
- Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., . . . Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. *Proceedings of the Eighth Workshop on Statistical Machine Translation* (pp. 1-44). Sofia, Bulgaria: Association for Computational Linguistics.
- Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2007). (Meta-) evaluation of machine translation. *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 136-158). Association for Computational Linguistics.

- Callison-Burch, C., Koehn, P., Monz, C., & Schroeder, J. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In A. f. Linguistics (Ed.), *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, (pp. 1-28). Athens, Greece.
- Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., & Specia, L. (2012). Findings of the 2012 Workshop on Statistical Machine Translation. *Proceedings of the Seventh Workshop on Statistical Machine Translation* (pp. 10-51). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Eckart de Castilho, R., & Gurevych, I. (2014, August). A broad-coverage collection of portable NLP components for building shareable analysis pipelines. *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT* (pp. 1-11). Dublin, Ireland: Association for Computational Linguistics and Dublin City University.
- Fishel, M., & Sennrich, R. (2014). Handling Technical OOVs in SMT. *The Seventeenth Annual Conference of the European Association for Machine Translation (EAMT2014)*, (pp. 159-162).
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378.
- Kluegl, P., Toepfer, M., Beck, P.-D., Fette, G., & Puppe, F. (2016). UIMA Ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, 22(01), 1-40.
- Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., . . . Zens, R. (2007). Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177-180.
- Koponen, M. (2012). Comparing human perceptions of post-editing effort with post-editing operations. *Proceedings of the Seventh Workshop on Statistical Machine Translation* (pp. 181-190). Association for Computational Linguistics.
- Kozlova, A., Shmatova, M., & Frolov, A. (2016). YSDA Participation in the WMT'16 Quality Estimation Shared Task. *Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics*, 2, pp. 793-799. Berlin, Germany: Association for Computational Linguistics.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Neubig, G. (2013, August). Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers. *Proceedings of the ACL Demonstration Track*.
- Neubig, G., Watanabe, T., & Mori, S. (2012). Inducing a discriminative parser to optimize machine translation reordering. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 843-853). Association for Computational Linguistics.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of the Association for Machine Translation in the Americas, 200*, pp. 223-231.
- Soricut, R., & Echiabi, A. (2010). Trustrank: Inducing trust in automatic translations via ranking. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 612-621). Association for Computational Linguistics.
- Specia, L. (2011). Exploiting objective annotations for measuring translation post-editing effort. *Proceedings of the 15th Conference of the European Association for Machine Translation*, (pp. 73-80).
- Specia, L., & Farzindar, A. (2010). Estimating machine translation post-editing effort with HTER. *Proceedings of the Second Joint EM+/CNGL Workshop Bringing MT to the User: Research on Integrating MT in the Translation Industry (JEC 10)*, (pp. 33-41).
- Specia, L., & Soricut, R. (2013). Quality estimation for machine translation: preface. *Machine Translation*, 27(3-4), 167-170.
- Specia, L., Saunders, C., Turchi, M., Wang, Z., & Shawe-Taylor, J. (2009a). Improving the confidence of machine translation quality estimates. *Proceedings of the Twelfth Machine Translation Summit (MT Summit XII)*, (pp. 136-143).
- Specia, L., Shah, K., De Souza, J. G., & Cohn, T. (2013). QuEst - A translation quality estimation framework. *ACL (Conference System Demonstrations)* (pp. 79-84). ACL.
- Specia, L., Turchi, M., Cancedda, N., Dymetman, M., & Cristianini, N. (2009b). Estimating the sentence-level quality of machine translation systems. *13th Conference of the European Association for Machine Translation*, (pp. 28-37).
- Ueffing, N., & Ney, H. (2005). Application of word-level confidence measures in interactive statistical machine translation. *Proceedings of EAMT*, (pp. 262-270).
- Wisniewski, G., Singh, A. K., & Yvon, F. (2013). Quality estimation for machine translation: Some lessons learned. *Machine translation*, 27(3-4), 213-238.

MTradumàtica: Free Statistical Machine Translation Customisation for Translators

Gökhan Dođru

Universitat Autònoma de Barcelona
gokhan.dogru@e-campus.uab.cat

Adrià Martín-Mor

Universitat Autònoma de Barcelona
adria.martin@uab.cat

Sergio Ortiz-Rojas

Prompsit Language Engineering
sergio@prompsit.com

Abstract

MTradumàtica is a free, Moses-based web platform for training and using statistical machine translation systems with a user-friendly graphical interface. Its goal is to offer translators a free tool to customise their own statistical machine translation engines and enhance their productivity. In this paper, we aim to describe the features of MTradumàtica and its advantages for translators by focusing on its current capabilities and limitations from a user perspective.¹

1. Introduction

The working environment of modern translators has been changing drastically. While there are still some translators trying to adapt to the advent of computer-aided translation (CAT) tools, now there is a need to adapt to a new working environment which also includes machine translation (MT). However, MT systems are presented generally as black-box solutions in which translators cannot intervene, make

modifications or customisations. Hence, the translators are dependent on MT solutions provided by either their language service providers or huge corporations.

We see the availability of free statistical machine translation (SMT) systems like Moses as a unique opportunity to narrow the technology gap between human translators and MT technology, and therefore to increase the effective usage of this technology. For the last few years, building and training SMT systems by end users has been a complex task involving a number of computing skills which might prevent the adoption of the technology. Therefore, we think that whenever necessary tools (free, open and easy-to-use tools) are presented to the translators, this gap can be eliminated to some extent, and translators can be empowered and be prepared to be competitive in the sector. With this assumption in mind, we have developed a Moses-based web platform, MTradumàtica, within the scope of ProjecTA.

ProjecTA (www.projecta.tradumatica.net) is a Tradumàtica group research project (www.tradumatica.net) at the Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental at the Universitat Autònoma de Barcelona. It works from the basic assumption that translators have the appropriate profile to manage MT-related tasks, and that empowering translators in MT tasks is beneficial for

¹ This work was supported by the ProjecTA project, grant number FFI2013-46041-R [MINECO / FEDER, UE].

translation companies. The project was split in two phases: first, to explore how MT is used by the translation sector in Catalonia and Spain through a survey sent to 187 translation agencies; second, based on the survey responses, to develop software to bring MT about closer to translators. The conclusion of the first phase is that MT use among most translation companies in Catalonia and Spain is low. Hence, ProjecTA decided to focus the second phase in the development of a software that can eliminate some of the barriers to implementing MT systems in the translation industry. These considerations have led to the creation of MTradumàtica.

2. What is MTradumàtica?

MTradumàtica is a free, Moses-based web platform for training and using SMT systems with a graphical user interface. Users can create their own engine in a few steps by uploading sentence-aligned parallel files in the usual Moses text format, then use these files to train a translation model and a language model, and ultimately train an SMT engine. To put it simply, there are 5 steps: (1) Upload files (2) Create and manage monotexts (3) Build language models (LMs) (4) Create and manage bitexts (5) Train SMT translation models. The LM, in the context of SMT, is the statistical model of a natural language, while the translation model (TM) includes the translation probabilities derived from parallel corpora. Monotexts are the monolingual texts used to create the language model, while bitexts are aligned bilingual texts (for example, a technical text aligned sentence by sentence with its translation) used to create the translation model. These two types of texts provide the training data for SMT to operate on.

At the end of the training, users can use their engine to translate texts or documents within the website. This means that translators can use their own resources or open resources (such as corpora from the Opus collection <http://opus.lingfil.uu.se>) and customise their own engines according to their needs. As stated

above, MTradumàtica aims at empowering translators in the context of the local translation fabric, made up mainly of small companies. Although the corporate perspective typically confines translators to mere end-users of MT, MTradumàtica aims at allowing them to develop their own engines and use them within their own personal, low-scale workflows.

The current version of MTradumàtica is available from GitHub

(<http://github.com/tradumatica/mtradumatica>).

It comes with a semi-automated installation procedure that works on Linux (local and server) and relies on technologies such as Python and Docker, as well as the software usually coming along with the Moses SMT system and other pieces of software from the Apertium project.

3. The Advantages and Limitations of MTradumàtica for Translators

One of the assumptions of SMT is that building MT engines from domain-specific parallel corpora tends to increase the quality of the raw output and, therefore, productivity. Considering that professional translators generally work on specialised domains for long time and collect huge amount of parallel corpora in time (under the form of translation memories), they can build their own engines and use them on a project-based basis. Since this customisation is made on the web platform, translators can use any operating system, provided that it has a web browser and an internet connection. However, there are still some developments needed for MTradumàtica to be fully functional for translators. Considering that most translators work with computer-aided translation (CAT) tools, their parallel corpora are generally exported in Translation Memory Exchange (TMX) format. Nevertheless, in the current version, it is not possible to upload TMX files to the file manager of MTradumàtica. Despite the fact that converting TMX to a moses file format is an easy task, the addition of the TMX upload feature will make MTradumàtica more convenient for translators. Secondly, for the

same reason, the integration of MTradumàtica with CAT tools through an API key will allow the translators to use their SMT engine within their own work environment. Thirdly, automatic evaluation metrics such as BLEU are needed to be able to evaluate the quality of the SMT engine beforehand so as to decide whether its quality is high enough to be used for translation tasks. Fourthly, confidentiality is a very important issue for translators (since they enter into non-disclosure contracts with their clients). The platform shall provide private user space (an account with a username and password) and guarantee that the parallel corpora are not used by anyone else. Although these are the prioritised features from the point of view of translators, some other features such as concatenating and prioritising models through GUI, terminology management, integrated corpora management, automated pre and post-editing functionalities shall be added to MTradumàtica. Currently, a feature called *Inspect* is also available. This feature, partially functioning at the moment for demonstration purposes, should allow the user to query and examine the components of the engines already created, i.e., the TM and the LM.

4. Concluding remarks

This paper has shortly described the current state of the MTradumàtica platform and its further developments. There is a certain need for a free machine translation platform for translators to remain competitive in the translation sector. MTradumàtica attempts to ease integration with the workflow and to remove most of the technical barriers for the integration of MT in enterprises so that freelance translators and small companies can use it. MTradumàtica is available at the moment for testing purposes at www.m.tradumatica.net.

Using error annotation to evaluate machine translation and human post-editing in a business environment

Lucia Comparin

Universidade de Lisboa
Centro de Linguística
da Universidade de Lisboa
Unbabel, Lisboa, Portugal
lcompa@gmail.com

Sara Mendes

Universidade de Lisboa
Centro de Linguística
da Universidade de Lisboa
Faculdade de Letras da Universidade de Lisboa
s.mendes@campus.ul.pt

Abstract

Quality Assessment currently plays a key role in the field of Machine Translation (MT) and in the organization of the translation market. Besides allowing to rank the players providing MT services, accurately assessing the quality of translation results is also a valuable step to improve the performance of automatic systems. In this study, we present the results of a study involving an error annotation task of a machine translated corpus from English into Italian. The data obtained allowed us to identify frequent and critical errors, and to observe their prevalence at different stages of the translation process, a most valuable analysis to outline strategies to automatically detect and correct the most relevant and prevalent errors in MT results. Accomplishing this is a crucial future step towards being able to guarantee the quality of results and a cost-effective workflow to obtain them.

1 Introduction

Research in machine translation (henceforth MT) has increased in the last decades, and MT systems have been increasingly integrated as part of the workflow adopted by translation providers in the market. Despite the development and improvements in MT systems and the continuous research done in the field, the quality of the results is still variable and dependent on many aspects such as the MT system used and the type of texts translated. This makes post-editing a necessary step

when MT is part of the translation process adopted by a company. At the same time, the variability of results highlights the importance of evaluating the performance of MT systems. Error annotation, i.e. the identification and categorization of errors present in a text, is used to assess the results of a MT system in terms of quality. Assessing quality of machine translated texts through error annotation is useful not only to evaluate the quality of the results produced by a MT system, but also to outline strategies to improve them and reduce the number of errors in the output produced. Such strategies can lead to the definition of specifications to implement in the system, or rules to automatically correct errors in the post-editing stage.

In the work presented in this paper, we performed error annotation of machine translated texts in order to provide data for improving translation results. The study was carried out within Unbabel, a startup company that offers almost real-time translation services by combining MT and human post-editing. Taking into account that Unbabel's translation workflow involves human post-editing, being able to identify and characterize the errors human editors are confronted with, and to which extent they persist after a first edition is crucial to outline strategies that aim at improving translation results in a cost-effective way.

2 Related Work

Due to the increasing adoption of MT systems in the translation process and to the development of different MT systems, quality assessment and the evaluation of MT systems have become an important field of research.

Quality assessment can be either performed by humans or automatic systems. Typically, in the former case, a human annotator identifies errors in

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

translation results, categorizes them and provides an analysis for them as described in Daems et al. (2014) and Stymne and Ahrenberg (2012). The latter, on the contrary, are based on the comparison of MT results with a human translation that is considered a high-quality reference. The most widely used systems are BLEU (Papineni et al., 2002) and METEOR (Lavie and Denkowski, 2009).

Research done in the analysis and description of MT errors is extensive and mostly related to the annotation and analysis of all errors present in texts that were translated using a particular MT system, in order to improve its performance (e.g. Kirchoff and Yang (2007) and Vilar et al. (2006)). The classification of errors is usually based on error taxonomies such as those presented in Vilar et al. (2006), and Popovic and Burchardt (2011). As annotation can be used to assess the quality of a translation for different purposes and in different contexts, error taxonomies are adapted to the purpose of the research. When they are used to assess the service provided by a company, they can be customized as described in the framework presented by Lommel (2015) under the scope of the Quality Translation 21 project. In Costa et al. (2015) an error taxonomy is presented to classify translation errors from English into European Portuguese, and a linguistic motivation for the selection of categories is provided. While these studies contemplate the description and categorization of errors, Hermjakob et al. (2008) concentrated on error detection, studying named entity translation errors, and improving an on-the-fly NE transliterator that is integrated into a statistical machine translation system.

3 Methodology

In the study presented here, we considered the language pair English-Italian and performed human annotation of a corpus. The corpus consisted of text provided by Unbabel clients and included web content such as travel descriptions and Customer Service emails, which were translated from English into Italian using the Google Translator API. In the translation process adopted by Unbabel, texts are firstly translated by the MT system, and then edited online by a community of human translators. Depending on the content of the text and on its length, one or more post-editions of the same text is performed. The corpus considered in this work included texts of 100 to 700 words. The

motivation for using texts with this length lies on the fact that, in order for the annotation to be accurate, texts have to be long enough for the annotator to understand the content, but short enough so that the task is not too time-consuming and demanding.

In this work, we annotated the texts of the corpus both immediately after MT and after the first human post-edition. This allows us to calculate the amount of errors that are corrected in post-editing and to figure out which errors generated by MT systems go on unnoticed at the following stage. The information resulting from the type of work described in this paper can therefore be used to outline strategies to improve post-editing and guarantee high-quality translations (Comparin, 2016; Comparin and Mendes, 2017).

In order to perform the annotation of the corpus, we considered the error taxonomy used at Unbabel. Work already done in the area and the analysis of each category were the starting point to better define the task performed in this study and its specifications. Data collected in the annotation of machine translated texts and in that of edited texts were then compared and analyzed, setting the grounds for the design of strategies to address the issues in the post-editing stage, as proposed in Comparin and Mendes (2017).

4 Error Annotation

The documents and guidelines used as a basis in order to define the typology used in the annotation were the MQM framework (Lommel, 2015) and TAUS documents (www.taus.net). The former is a model developed in the Quality Translation 21 project, funded by the European Union Horizon 2020 research and innovation program, whose goal is to overcome language barriers to encourage flow of ideas, commerce and people within the EU. TAUS is a resource center offering support to translation service providers by making available different tools, such as software, metrics, and knowledge. A framework was developed in the scope of the QT21 project in order to define task-specific translation metrics, that help assessing the translation performed by a MT system or by a company.

The tool used in this study was created by Unbabel and used to assess the quality of the texts delivered to clients in different language pairs on a weekly basis. The tool shows the source text, the target text, the annotations, and the glossary terms.

When the annotator selects a word or a sequence of words in the target text, possible error types appear in a box and the relevant one can be selected. Additionally, the annotator can also assess the fluency of the entire text, using a scale of 0 to 5.

In order to design an error taxonomy suitable to an annotation task with the goals of the one discussed in this work, some prerequisites have to be considered. Taking into account the standards and the work already done in the annotation field not only to define the set of useful error types, but also to guarantee that annotation is accurate, first, all errors that can be generated in MT should be covered, but the number of error types should be limited, to avoid noise in data annotation and to make the annotation process affordable both in terms of time dedicated to the task and in terms of its cost. Secondly, error types should be clearly distinguished from one another.

4.1 Error Types and Penalty system

The 41 error types included in the taxonomy used at Unbabel and considered in this study are divided into 7 major categories: accuracy, fluency, style, terminology, wrong language variety, named entities, and formatting and encoding. Below we briefly define the aforementioned error categories considered in the typology.

ACCURACY: errors in this category concern the relationship between the source text and the target text and the extent to which the latter maintains the meaning and the information of the former

FLUENCY: errors in this category regard the quality of a text, assessing whether it is well-written and easy to read, and if it accomplishes its communication purpose in the target language

STYLE: issues concerning register and fluency

TERMINOLOGY: mistranslation of terminology

WRONG LANGUAGE VARIETY: use of a word or expression from a different language variety.

NAMED ENTITIES: wrong translation of proper nouns

FORMATTING AND ENCODING: issues concerning the segmentation of sentences and paragraphs

In addition to the categorization of errors, a penalty is also available to be associated to each

error annotated. By doing this, a numerical quality score can be calculated by the tool for each translation, and can be used as an indicator of its quality and of the improvements still to be made. Additionally, such a score is used in the industry to position a company in the market. The penalty system was set up based on the system used at Google LQE (Localization Quality Evaluation) and in the MQM. The errors annotated were divided according to their severity into minor, major and critical errors, following the definitions below.

Minor: Errors that do not change nor compromise the information provided in the source text. They do not prevent the reader of the target text to understand it in a clear way and they do not generate confusion or doubts. They can nonetheless affect fluency. The penalty associated to minor errors is 0.5 points.

Major: Errors that make the target text either confusing or ambiguous. They make it more difficult for the reader to clearly understand the text, although the target text conveys the message. In some cases, the meaning of the target text can slightly change, however general comprehension is guaranteed. The penalty associated to this type of error is 1 point.

Critical: Critical errors change the meaning of the source text. Not only they prevent the reader from understanding the information provided in the text, but also they can cause damage to the reputation of a company and carry health, safety or legal implications. The penalty associated to this type of error is 3 points.

4.2 Some remarks on the annotation performed in this study

Before discussing the data obtained in the annotation task, we would like to discuss a few aspects related to annotation and make a few notes regarding the task in this particular case. Human annotation can be a challenging task, as it is related to the annotators understanding and categorization of an error. In this study each annotation was performed by a single annotator, which made the definition of clear guidelines to help in the task a necessary step. In those cases in which a single error simultaneously involved different error types, the type that provided more information about the phenomenon at stake was preferred. For instance, when a conjunction was omitted, the error category *conjunctions* was selected instead of *omission*.

Additionally, due to technical constraints regarding the platform used at Unbabel - the annotation tool used does not allow the association of more than one error type to the same expression - , when one word or sequence of words contained more than one error, only the most relevant one was marked. Since data collected from annotation, in this specific case, were used to improve translation results through the definition of a set of rules for automatic post-edition and/or automatic checking of machine translated results, errors types involving grammar phenomena were preferred, such as *agreement, tense/mood/aspect, word order, sentence structure, prepositions, conjunctions, or determiners*. If the purpose of the annotation were to study spelling mistakes in MT, then *orthography* errors would be selected as more relevant.

Since in this work we concentrated on errors after MT and after the first post-edition, a high number of errors, and particularly of critical errors, was observed in the target text. Given this, and even if a penalty was assigned to each error during annotation, we do not discuss this aspect here as the high number of errors and the great impact they have on translation quality does not allow for clear and insightful distinctions in terms of severity for a great part of the annotated errors.

The guidelines defined in the MQM framework (Burchardt and Lommel, 2014), which highlight the fact that the annotator should be as precise as possible both in the selection of the text containing the error, and in the selection of the error type, were taken into account in this study, as long as the specifications of the annotation tool used at Unbabel allowed the annotator to do so, which was not always the case, as mentioned above.

5 Annotation Data

The errors annotated both after MT and after the first post-edition are presented in the tables below. In Table 1, absolute and relative frequency of annotated errors per error category in the typology is presented.

The data in Table 1 show that the number of errors in machine translated texts is high and not evenly distributed among the different error categories. This is certainly related to the fact that it was not possible to mark two errors in the same word or sequence of words, and, in such cases, the error with the greatest impact on the quality of the translation and particularly on the access to the

content of the text was marked, and thus the categories mentioned in section 4.2 were preferred.

With regard to the number of errors in the two stages considered in our study, MT and the first post-edition, there is an 85% error reduction between the two stages. However, the impact of human post-edition on error reduction is variable between different error categories: e.g while *fluency errors* lower their relative frequency from 77% to 49%, *accuracy errors* actually increase their relative frequency (the absolute number of errors decreases significantly in both cases, naturally: 90.2% for *fluency* and 76.7% for *accuracy*). The significant increase of the relative frequency of errors in error types more related to style and client specifications (e.g. *inconsistent register, repetitive style, or noncompliance with client's glossary and vocabulary*) is due to the fact that, in many cases in which an error belonging to these types occurred after MT, more severe errors were present in the translated texts, and were thus the ones marked. As the first post-edition tends to correct the most severe errors, those related to the creative use of language and style become in turn visible. Let us now consider the most frequently marked error categories in more detail, i.e. *accuracy errors* and *fluency errors*.

The error type with the highest number of errors annotated in machine translated texts is *determiners*, followed by *lexical selection, agreement, tense/mood/aspect, and word order*. Errors belonging to these error types, in the majority of the cases, do not allow the reader to understand the text clearly, and therefore have a major or critical impact on the quality of the translation. Two error types that have a lower number of errors but are still crucial for the quality of translation results are *sentence structure* and *prepositions*. Errors in sentence structure, in particular, have a great impact on translation, because they often result in a sentence that cannot be understood without knowledge of the source language and the sentence structures commonly used in it. Additionally, such errors require a major intervention of the editor, since the text has to be rewritten in the majority of the cases, which takes significantly more time than just changing a morpheme or a word. The time spent in the correction of errors involving prepositions is also considerable, because, when the wrong preposition is selected, the meaning of the text often cannot be fully and accurately

| Main error types | MT | | FIRST EDITION | |
|--------------------------------|------------|------------|---------------|------------|
| | abs. freq. | rel. freq. | abs. freq. | rel. freq. |
| Accuracy errors | 236 | 0.21 | 55 | 0.32 |
| Fluency errors | 848 | 0.77 | 83 | 0.49 |
| Style errors | 1 | 0 | 3 | 0.02 |
| Terminology errors | 0 | 0 | 14 | 0.08 |
| Wrong language variety errors | 0 | 0 | 0 | 0 |
| Named entities errors | 19 | 0.02 | 15 | 0.09 |
| Formatting and encoding errors | 0 | 0 | 0 | 0 |
| Total | 1104 | 1 | 170 | 1 |

Table 1: Absolute and relative frequency of annotated errors per error category after MT and first human edition

| Accuracy errors | MT | | FIRST EDITION | |
|---------------------------------|------------|------------|---------------|------------|
| | abs. freq. | rel. freq. | abs. freq. | rel. freq. |
| Mistranslation | | | | |
| Overly literal | 9 | 0.01 | 4 | 0.02 |
| False friend | 0 | 0 | 0 | 0 |
| Should not have been translated | 18 | 0.02 | 3 | 0.02 |
| Lexical selection | 165 | 0.15 | 37 | 0.22 |
| Omission | 6 | 0.01 | 0 | 0 |
| Untranslated | 27 | 0.02 | 9 | 0.05 |
| Addition | 11 | 0.01 | 2 | 0.01 |
| Total | 236 | 0.21 | 55 | 0.32 |

Table 2: Absolute and relative frequency of accuracy errors after MT and first human edition

| Fluency errors | MT | | FIRST EDITION | |
|-----------------------------------|------------|------------|---------------|------------|
| | abs. freq. | rel. freq. | abs. freq. | rel. freq. |
| Inconsistency | | | | |
| Word selection | 1 | 0 | 1 | 0.01 |
| Tense selection | 0 | 0 | 0 | 0 |
| Coherence | 2 | 0 | 1 | 0.01 |
| Duplication | 0 | 0 | 0 | 0 |
| Spelling | | | | |
| Orthography | 1 | 0 | 1 | 0.01 |
| Capitalization | 52 | 0.05 | 19 | 0.11 |
| Diacritics | 0 | 0 | 0 | 0 |
| Typography | | | | |
| Punctuation | 9 | 0.01 | 4 | 0.02 |
| Unpaired quote marks and brackets | 1 | 0 | 0 | 0 |
| Whitespace | 17 | 0.02 | 5 | 0.03 |
| Inconsistency in character use | 0 | 0 | 0 | 0 |
| Grammar | | | | |
| Function words | | | | |
| Prepositions | 70 | 0.06 | 10 | 0.06 |
| Conjunctions | 12 | 0.01 | 1 | 0.01 |
| Determiners | 237 | 0.21 | 19 | 0.11 |
| Word form | | | | |
| Part-of-speech | 30 | 0.03 | 1 | 0.01 |
| Agreement | 159 | 0.14 | 13 | 0.08 |
| Tense/mood/aspect | 101 | 0.09 | 3 | 0.02 |
| Word order | 106 | 0.10 | 4 | 0.02 |
| Sentence structure | 50 | 0.05 | 1 | 0.01 |
| Total | 848 | 0.77 | 83 | 0.49 |

Table 3: Absolute and relative frequency of fluency errors after MT and first human edition

understood just by considering the text produced by the MT system. Comparing these more frequent types of errors in the two stages of the translation process, we can identify two types of behavior: some of the most critical errors, such as *tense/mood/aspect*, *word order* and *sentence struc-*

ture are almost non-existent after the first human post-edition; on the other hand, errors that are in principle more straightforward to correct, such as *determiners* or *agreement* are visibly reduced, but their relative weight considering all the errors annotated after the first human post-edition is still

considerable. This observation is probably not independent from the fact that these are errors which are easier to be overseen by a human editor, as they often amount to a small variation in the form of the lexical items involved. Finally, some brief remarks regarding errors involving *prepositions* and *lexical selection*, which, respectively, show no reduction and an increase in their relative weight after the post-edition stage, when compared with what was the case after MT. These data make apparent that this type of error persists even after human edition, its weight in the overall amount of errors annotated remaining important by the crucial reduction of other types of error.

6 Results and final remarks

The error annotation presented in this work allowed us to analyze the most significant types of error occurring in machine translated texts from English into Italian using Google Translator API, and their prevalence after the first human post-edition. As expected, the comparison between the errors annotated at these two stages of the translation process is marked by a significant reduction in the absolute number of errors. This comparison also made apparent that there are certain types of error that continue to be present even after human edition. The amount of errors after MT and the prevalence of certain types of error make apparent the need for using the results and analysis of this annotation task to outline strategies to automatically tackle the shortcomings of MT systems and aid human post-edition, as we have proposed and evaluated in Comparin (2016) and Comparin and Mendes (2017).

References

- Burchardt, Aljoscha and Arle Lommel. 2014. *Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality*. URL <http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf>.
- Comparin, Lucia. 2016. *Quality in Machine Translation and human post-editing: error annotation and specifications*. MA dissertation, Faculdade de Letras da Universidade de Lisboa, Portugal.
- Comparin, Lucia and Sara Mendes. 2017. Error detection and error correction for improving quality in machine translation and human post-editing. *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2017*, Budapest, Hungary, 2017.
- Costa, Angela, Wang Ling, Tiago Luis, Rui Correia, and Luisa Coheur. 2015. A linguistically motivated taxonomy for Machine Translation error analysis. *Machine Translation*, 29(2): 127-161.
- Daems, Joke, Lieve Macken, and Sonia Vandepitte. 2014. On the origin of errors: A fine-grained analysis of MT and PE errors and their relationship. *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, Reykjavik, Iceland, May 26-31, 2014, pages 62-66.
- Hermjakob, Ulf, Kevin Knight, and Hal Daume III. 2008. Name Translation in Statistical Machine Translation - Learning When to Transliterate. *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, June 15-20, 2008, Columbus, Ohio, USA, pages 389-397.
- Kirchhoff, Katrin and Mei Yang. 2007. The University of Washington machine translation system for the IWSLT 2007 competition. *2007 International Workshop on Spoken Language Translation, IWSLT 2007*, Trento, Italy, October 15-16, 2007, pages 89-94.
- Lavie, Alon and Michael J. Denkowski. 2009. The Meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105-115.
- Lommel, Arle. 2015. *Multidimensional Quality Metrics (MQM) Definition*. URL <http://www.qt21.eu/mqm-definition/definition-2015-12-30.html>.
- Papineni, Kishore, Salim Roukos, Todd Ward and Weijing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, July 6-12, 2002, Philadelphia, PA, USA, pages 311-318.
- Popovic, Maja and Aljoscha Burchardt. 2011. Error Analysis of Machine Translation Output. *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, pages 265-272, Genoa, Italy. European Association for Machine Translation.
- Stymne, Sara and Lars Ahrenberg. 2012. On the practice of error analysis for machine translation evaluation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, Istanbul, Turkey, May 23-25, 2012, pages 1785-1790.
- Vilar, D., J. Xu, L. D'haro, and H. Ney. 2006. Error Analysis of Statistical Machine Translation Output. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA).

Empirical evaluation of NMT and PBSMT quality for large-scale translation production.

Dimitar Shterionov^α Pat Nagle^α Laura Casanellas^β Riccardo Superbo^β Tony O’Dowd^β

{dimitars, patn, laurac, riccardos, tonyod}@kantanmt.com

^α KantanLabs, INVENT Building, Dublin City University Campus, Dublin 9, Dublin, IRELAND

^β KantanMT, INVENT Building, Dublin City University Campus, Dublin 9, Dublin, IRELAND

Abstract

Neural Machine Translation (NMT) has recently gained substantial popularity not only in academia, but also in industry. In the present work, we compare the quality of Phrase-Based Statistical Machine Translation (PBSMT) and NMT solutions of a commercial platform for Custom Machine Translation (CMT) that are tailored to accommodate large-scale translation production. In a large-scale translation production line, there is a limited amount of time to train an end-to-end system (NMT or PBSMT). Our work focuses on the comparison between NMT systems trained under a time restriction of 4 days and PBSMT systems. To train both NMT and PBSMT engines for each language pair, we strictly use the same parallel corpora and show that, even if trained within this time limit, NMT quality surpasses substantially that of PBSMT.

Furthermore, we challenge the reliability of automatic quality evaluation metrics (in particular, BLEU) for NMT quality evaluation. We support our hypothesis with both analytical and empirical evidence.

1 Introduction

Recent research in MT based on Artificial Neural Networks – Neural Machine Translation (NMT) (Bahdanau et al., 2014; Cho et al., 2014; Sutskever et al., 2014) – has shown promising results and has gained popularity not only in

academia but also in industry. It promises to solve some of the drawbacks that SMT comes upon. Studies like those of Bentivogli et al. (2016), Wu et al. (2016) and Junczys-Dowmunt et al. (2016) indicate that the quality of NMT surpasses that of SMT, and a shift in the state of the art is imminent. Although several MT vendors, such as Google, Microsoft, Systran, KantanMT, offer NMT as part of their services, it is still uncertain to which extent NMT can replace SMT as core technology for large-scale translation projects. The main reasons are the computational (and financial) cost of NMT and the uncertainty in the actual quality: while NMT output is often very fluent, sometimes it lacks adequacy or is even completely wrong.

In this work, we compare Phrase-Based SMT (PBSMT) and NMT within a translation production line. We set a time limit for training NMT models of 4 days – sufficient for our NMT models to reach high quality without introducing overhead in the production line. We use quality evaluation metrics such as BLEU (Papineni et al., 2002), F-Measure (Melamed, 1995), and TER (Translation Error Rate) (Snover et al., 2006),¹ as well as human evaluation. We challenge the relevance of BLEU for scoring NMT models. Our hypothesis is that BLEU *underestimates* the quality of NMT models. We provide empirical as well as analytical evidence to support our hypothesis.

2 Related work

Since 2015, NMT systems have been clearly outdoing SMT. In the International Workshop on Spoken Language Translation (IWSLT) 2015 competition (Cettolo et al., 2015), an NMT system outper-

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

¹BLEU, F-Measure and TER are algorithms for quality evaluation of MT systems, typically used to estimate fluency, adequacy and extent of translation errors.

formed a number of PBSMT systems. Bentivogli et al. (2016) compare and analyse the overall translation quality as well as the translation errors of NMT and PBSMT systems for English→German based on data from the IWSLT 2015 competition (Cettolo et al., 2015). Their results show that NMT is better than all the four different SMT systems on all investigated criteria: (i) higher automatic scores (i.e., BLEU); (ii) lower morphologic, lexical and reordering (especially, verb reordering) errors and (iii) reduced post-editing effort.

Despite the thoroughness of their analysis and the significance of their results, Bentivogli et al. (2016) compare systems trained and tuned on different data – their NMT system is trained on parallel data of 120,000 tokens, whereas their standard PBSMT system is trained on parallel data of 117,000 tokens and 2.4 billion tokens of monolingual data. Our work compares PBSMT and NMT trained on exactly the same data; we scored our systems and performed side-by-side comparison (i.e., *AB tests*) on the same test sets as well.

SMT and NMT systems have also been extensively compared by Junczys-Dowmunt et al. (2016). The authors investigate the BLEU scores of multiple NMT and SMT systems for 10 languages and 30 language directions trained on the United Nations Parallel Corpus v 1.0 (Ziems et al., 2016). Their NMT systems outrank SMT for all but three cases: French→Spanish (the BLEU score for PBSMT is 1.16% higher than NMT), French→English (the BLEU score for the hierarchical system Hiero as implemented in Moses is 1.15% higher than their initial NMT system; after additional training, the BLEU score for NMT is 1.13% higher than Hiero) and Russian→English (the BLEU score for the hierarchical system is respectively 1.32% and 0.75% higher than the initial NMT system and the one with additional training). On an NVIDIA GTX 1080, their NMT systems were initially trained for 8 days; for the language pairs that include English, an additional training of 8 days (16 days in total) was performed.

One of the largest providers of MT services (both public and commercial) – Google – has recently presented their NMT (Google NMT or GNMT) approach and compared it to PBSMT (employing both BLEU scoring and human evaluation) as well as to human translation (Wu et al., 2016). The results they report, although quite disputed, provide once again empirical evidence

that the quality of NMT is generally higher than that of PBSMT. The GNMT systems follow a rather optimised implementation of the sequence-to-sequence model (Sutskever et al., 2014) with attention mechanism (Bahdanau et al., 2014) trained on 96 GPUs². Each model was trained for approximately 6 days, then refined for approximately 3 days (9 days in total). For training 36 million parallel sentences for English→German and 5 million parallel sentences for English→French were used.

Another comparison between NMT and other MT paradigms was presented by (Crego et al., 2016). Their work investigates the quality (scored in terms of BLEU as well as human evaluation) of NMT systems, PBSMT, rule-based MT and human translation (from both professional and non-professional translators); moreover, an error analysis is presented. Although their NMT systems outperform PBSMT and rule-based MT, they still do not reach human translation quality.

3 BLEU as a quality metric for (N)MT

The most widely used quality evaluation metric for MT systems, i.e., BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002), was one of the first metrics to report high correlation between MT quality and human judgment. BLEU measures the precision of an MT system computed through the comparison of the system’s output and a set of ideally correct, and usually human-generated reference translations. The BLEU algorithm compares the n-grams (typically, $n \in \{1, \dots, 4\}$) of a candidate translation with those of the corresponding reference and counts the number of matches. The more n-gram matches between a translation and the reference, the higher the score.

BLEU scores can be computed either at a document level or at a sentence level (Chen and Cherry, 2014). They range between 0 (or 0% – lowest quality = completely irrelevant to the reference) and 1 (or 100% – highest quality = same as the reference). The relevant factors for computing BLEU scores are: (i) **Translation length**: a correct translation matches the reference in length; (ii) **Translated words**: the words in a correct candidate translation match the words in the reference; (iii) **Word order**: the order of words in a correct candidate translation and in the reference is the same.

In PBSMT, phrase-level (n-gram) translations are arranged in a specific order that maximises

²The reported GPUs are NVIDIA Tesla K80.

the sentence-level translation likelihood. If an n-gram cannot be translated, usually the original text is transferred. PBSMT translations typically conform with BLEU according to *translation length*, *translated words* and *word order*, as they are both n-gram based.

NMT systems operate differently from PSMT. A typical encoder-decoder system (Sutskever et al., 2014; Cho et al., 2014) would generate a sentence translation based on the complete sequence of tokens from the source sentence, as well as all preceding translated tokens from the current sentence. NMT translations are not bound by the limits of n-grams. As such, NMT output may deviate from the reference according to *sentence length* and *word order* within the n-gram limit specified by the BLEU algorithm. Furthermore, to tackle out-of-vocabulary (OOV) issues and reduce vocabulary size, it is customary to build NMT systems on subword units (Sennrich et al., 2016) or even characters (Chung et al., 2016). This would provide the network with greater flexibility and allow it to extend beyond exact words or phrases from the training data. For this reason, NMT output, although representing a correct translation, may deviate significantly from the reference also according to *word choice* (see Example 3.1).

That is why, we believe that BLEU **underestimates** NMT systems. In Section 4, we empirically support our claim. We ought to note that we focus on *sentence-level* BLEU, which has the granularity that suits our sentence-by-sentence comparison.

Example 3.1 An NMT translation with 0% BLEU that is better than a PBSMT one with 58% BLEU.

Source (EN): *All dossiers must be individually analysed by the ministry responsible for the economy and scientific policy.*

Reference (DE): *Jeder Antrag wird von den Dienststellen des zuständigen Ministers für Wirtschaft und Wissenschaftspolitik individuell geprüft.*

PBSMT: *Alle Unterlagen müssen einzeln analysiert werden von den Dienststellen des zuständigen Ministers für Wirtschaft und Wissenschaftspolitik.* **BLEU:** 58%

NMT: *Alle Unterlagen müssen von dem für die Volkswirtschaft und die wissenschaftliche Politik zuständigen Ministerium einzeln analysiert werden.* **BLEU:** 0% \triangle

4 Comparing NMT to SMT output

4.1 SMT and NMT pipelines

For the present work, we employ KantanMT (<https://kantanmt.com/>) – a cloud-based MT platform which delivers MT services individu-

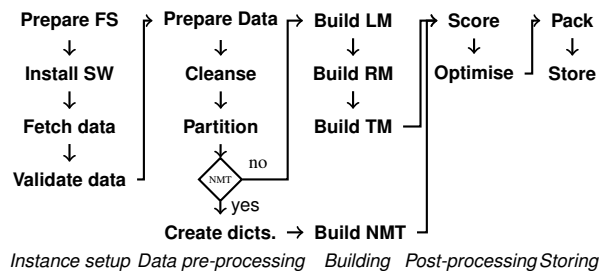


Figure 1: MT training pipeline.

ally to each user. A user can create, customise and exploit their own MT engine(s)³ within a secure environment. Typically, a user creates an engine from scratch; in case their data is not sufficient to train a performant engine, additional data or a pre-built engine can be retrieved from our data banks.

The training pipeline for both NMT and PBSMT engines follows the same architecture: 1. *Instance setup* – hardware is allocated, software is set up and data is downloaded; 2. *Data pre-processing* – data is converted to suitable format, cleansed and partitioned for training, testing and tuning; in the case of NMT, any duplicate sentence pair that appears in the source and the target sides of the parallel corpus (i.e., the training data) is removed; moreover, the required dictionaries are prepared; 3. *Building of models* – for PBSMT, a translation, a language and a recasing models are built; for NMT an encoder-decoder model is built; 4. *Engine post-processing* – the engine is evaluated, optimised and stored for future use. Figure 1 illustrates these steps. To train PBSMT models, our pipeline uses the Moses toolkit (Koehn et al., 2007) with default settings and lexicalised reordering model with distortion limit of 6 words. We use monolingual data extracted from the target side of the parallel corpus to build a 5-gram language model. For word alignment, we use fast_align (Dyer et al., 2013). Tuning is performed with MERT (Och and Ney, 2003) and a maximum of 25 iterations. For NMT, we employ OpenNMT (Klein et al., 2017). A single NMT model is trained on one NVIDIA G520 GPU with 4GB RAM. As a learning optimiser, we use ADAM (Kingma and Ba, 2014) with a learning ratio of 0.005. Within the scope of this study, we impose the following training limits: minimum number of training epochs is 3; maximum train-

³An MT engine refers to the package of models (translation, language and recasing models for PBSMT and encoder-decoder model for NMT) as well as to the required rules and dictionaries for pre- and post-processing.

ing time is four days; to consider a model fitted for evaluation, its validation perplexity should be below 3 at the end of the training. One exception, English→German, has a perplexity of 3.02 at the end of the fourth day; we ought to note also that the English→Chinese engine achieved perplexity of 2 on the first day.

Our decision to set a limit of four days is guided by economical and practical reasons. Our MT development process has a duration of six weeks. Training an engine for more than four days would disrupt the structure of this process and may impose further delays in a large-scale translation project. Furthermore, it is also financially inviable.

For data in Chinese, Japanese, Korean or Thai, our pipeline uses dictionaries based on character-by-character segmentation (Chung et al., 2016). For other languages, we use dictionaries built from word-subunits. These subunits are generated from the training data according to a byte pair encoding (BPE) (Sennrich et al., 2016) of 40,000 operations. We prepare the dictionaries from normalised (i.e., lower- and upper-cased) tokenised data.

4.2 Used data

We built five NMT and five PBSMT engines for the following language pairs: English→German (EN-DE), English→Chinese (EN-ZH-CN),⁴ English→Japanese (EN-JA), English→Italian (EN-IT) and English→Spanish (EN-ES). For each language pair, both the PBSMT and the NMT engines were built using strictly the same data set. By keeping identical train, test and tune data sets from one engine to another, we can give a more informative comparison of the SMT and NMT engines and their outputs. Details about the data used in our experiments are given in Table 1. The

| Lang. pair | Sent. count | Word count | Dict. size | Domain |
|------------|-------------|-------------|------------|-----------------|
| EN-DE | 8,820,562 | 110,150,238 | 859,167 | Legal/Medical |
| EN-ZH-CN | 6,522,064 | 84,426,931 | 956,864 | Legal/Technical |
| EN-JA | 8,545,366 | 87,252,129 | 676,244 | Legal/Technical |
| EN-IT | 2,756,185 | 35,295,535 | 765,930 | Medical |
| EN-ES | 3,681,332 | 44,917,583 | 752,089 | Legal |

Table 1: Details on the data used for experiments.

data comprises parallel translation memories in the Legal, Medical and Technical domains, acquired from the European Commission (DGT)⁵ and from Opus.⁶ Prior to training, the data was cleansed

⁴By Chinese, we mean Simplified Mandarin Chinese

⁵<https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory>

⁶<http://opus.lingfil.uu.se/>

and normalised, i.e., duplicates were removed. Untranslated segments and segments constructed of special characters were also removed, as they would not be relevant to the evaluation.

4.3 Evaluation

Quality evaluation metrics Table 2 shows the scores of the quality evaluation metrics we use (F-Measure, BLEU and TER) for both PBSMT and NMT engines. We also show the training time in hours; for the NMT engines, each model’s perplexity on the test set is also given.

| Lang. Pair | PBSMT | | | | NMT | | | | |
|------------|-----------|-------|-------|-----|-----------|-------|-------|------|-----|
| | F-Measure | BLEU | TER | T | F-Measure | BLEU | TER | P | T |
| EN-DE | 62.00 | 53.08 | 54.31 | 18 | 62.53 | 47.53 | 53.41 | 3.02 | 92 |
| EN-ZH-CN | 77.16 | 45.36 | 46.85 | 6 | 71.85 | 39.39 | 47.01 | 2.00 | 10 |
| EN-JA | 80.04 | 63.27 | 43.77 | 9 | 69.51 | 40.55 | 49.46 | 1.89 | 68 |
| EN-IT | 69.74 | 56.98 | 42.54 | 8 | 64.88 | 42.0 | 48.73 | 2.70 | 83 |
| EN-ES | 71.53 | 54.78 | 41.87 | 9 | 69.41 | 49.24 | 44.89 | 2.59 | 71 |

Table 2: Evaluation scores (in %), training time (T) in hours and perplexity (P) (only for NMT).

Side-by-side comparison We set up a side-by-side, or AB Test, project with our online quality evaluation tool. For the test, human evaluators compared 200 segments translated using the aforementioned PBSMT and NMT engines. This exercise was performed by 15 evaluators – three evaluators per language pair – all of whom were native speakers of the (target) language they evaluated. All evaluators were Translation Studies students recruited from five different universities in Europe, holding certificates of English proficiency or attending courses taught in English. All evaluators of one language pair had to compare the same segments translated by the two engines. The test was performed online. Each evaluator was instructed on how to access the platform and how to perform the test. Each evaluator was requested to evaluate all test sentences without taking any significant break. The sentences were presented on the screen as a triplet (*Source*, *PBSMT Translation*, *NMT Translation*) – denoted as (s, t_{NMT}, t_{PBSMT}). The order of the sentences t_{NMT} and t_{PBSMT} was randomised, i.e., t_{NMT} could be preceding t_{PBSMT} or vice versa. This would ensure that the evaluators do not get used to one style of translation and show preference towards it. The evaluator was instructed to first read the original sentence (s) in English, then the two translation candidates (t_{NMT} or t_{PBSMT}) and then decide which was of better quality or whether they were of equal quality (either good or bad). The test sets did not contain any

| | EN → ZH-CN | | | EN → JA | | | EN → DE | | | EN → IT | | | EN → ES | | |
|-------------|------------|-------|-----|---------|-------|-----|---------|-------|-----|---------|-------|-----|---------|-------|-----|
| | Same | PBSMT | NMT | Same | PBSMT | NMT | Same | PBSMT | NMT | Same | PBSMT | NMT | Same | PBSMT | NMT |
| Evaluator 1 | 41% | 20% | 39% | 21% | 19% | 60% | 19% | 27% | 54% | 25% | 19% | 56% | 12% | 28% | 60% |
| Evaluator 2 | 34% | 26% | 40% | 14% | 28% | 58% | 14% | 35% | 51% | 29% | 14% | 57% | 10% | 26% | 64% |
| Evaluator 3 | 37% | 25% | 38% | 27% | 16% | 57% | 6% | 40% | 54% | 19% | 25% | 56% | 7% | 31% | 62% |
| Average | 37% | 24% | 39% | 21% | 21% | 58% | 13% | 34% | 53% | 24% | 19% | 56% | 10% | 28% | 62% |

Table 3: Side-by-side PBSMT and NMT evaluation performed by human reviewers.

duplicates – i.e., training, testing and tuning data was normalised beforehand.

The results we gathered, summarised in Table 3, clearly contradict the scores presented in Table 2. We observe that all evaluators scored more of the translations that originate from an NMT engine better (i.e., being translations of higher linguistic quality and/or expressing more accurately the meaning of the source sentences) than their PBSMT alternatives. This (i) shows that NMT is better under the conditions specified in Section 4.1, and (ii) supports our claim that quality evaluation metrics are not reliable for NMT. It is, however, interesting to observe that for the EN-ZH-CN data, 37% of the translations are scored the same; in general, for this language pair, the NMT engine is not evaluated as high as the others. A closer investigation shows that this engine was trained quite quickly reaching a low perplexity that allowed the training process to terminate at an early stage. While further investigation for whether additional training will lead to improving these scores is required, we ought to stress the importance of how much time is devoted to training an NMT engine.

BLEU underestimation of NMT output quality

We use the data from our AB Test to analyse to what extent BLEU underestimates NMT quality as compared to human judgement.

For each language pair, we selected the set of triplets (s, t_{NMT}, t_{PBSMT}) for which the translation produced by the NMT engine was considered of better quality by all three evaluators. Let us denote their count as d^{NMT} . Then, from this set we counted the number of translations with a BLEU score lower than their PBSMT counterparts. Let us denote this number as d_{PBSMT}^{NMT} . We then computed the fraction $\frac{d_{PBSMT}^{NMT}}{d^{NMT}}$. We performed the same check for the PBSMT candidates that were considered of better quality by the three evaluators, i.e., we computed the fraction $\frac{d_{NMT}^{PBSMT}}{d^{PBSMT}}$. We present these scores as percentages in Table 4. We observe that the percentage of underestimated sentences for NMT is significantly higher than for PBSMT. It is interesting to highlight that two of the

| | EN-ZH-CN | EN-JP | EN-DE | EN-IT | EN-ES | Average |
|-----|----------|-------|-------|-------|-------|-----------|
| NMT | 40 | 59 | 55 | 34 | 53 | 48 |
| SMT | 12 | 0 | 9 | 9 | 0 | 6 |

Table 4: Underestimation of BLEU scores (%).

language pairs, EN-JA and EN-ES, do not have any underestimated scores for PBSMT, but they are respectively the highest and the third highest underestimated language pairs in the NMT case. On average, the underestimation of BLEU for our NMT engines and our test sentences amounts to 48%. That is, we can say that *on average*, 48% of the NMT translations with BLEU scores worse than for their PBSMT counterparts are judged by the human evaluators as better. We should also mention that, for the other quality evaluation metrics (i.e., F-Measure and TER), the results are rather similar. As it extends beyond our current research (which focuses on BLEU), further analysis will be addressed in future work.

5 Conclusions and future work

In this work, we analysed the NMT and PBSMT systems of a commercial MT platform. We trained five NMT and five PBSMT engines on the same data and under a time limitation that would allow for a large-scale translation development with no delays. We then compared the quality evaluation scores (F-Measure, TER and BLEU) of these engines with human evaluation. In all cases, the human reviewers, all native speakers of the evaluated language pairs, ranked the quality of the NMT engines higher than that of PBSMT. While these results are in agreement with previous research, we show that BLEU scores do not always conform with NMT quality. Rather, they underestimate NMT quality.

In the future, we plan to perform quality ranking of other language pairs, including challenging ones, e.g., Baltic languages. Furthermore, we intend to measure the quality of the NMT output in comparison to the quality of the PBSMT output to observe if the difference is significant and if it varies depending on the language pairs. Given the current differences in terms of setup and cost be-

tween PBSMT and NMT, this information is essential for MT users in a commercial environment.

Acknowledgements We would like to thank our external evaluators: Xiyi Fan, Ruopu Wang, Wan Nie, Ayumi Tanaka, Maki Iwamoto, Risako Hayakawa, Silvia Doehner, Daniela Naumann, Moritz Philipp, Annabella Ferola, Anna Ricciardelli, Paola Gentile, Celia Ruiz Arca, Clara Beltr, as well as University College London, Dublin City University, KU Leuven, University of Strasbourg, and University of Stuttgart.

References

- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR, Accepted for oral presentation at the International Conference on Learning Representations (ICLR) 2015*, abs/1409.0473.
- Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*.
- Cettolo, Mauro, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2015 evaluation campaign. *Proc. of IWSLT, Da Nang, Vietnam*.
- Chen, Boxing and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*.
- Cho, Kyunghyun, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of EMNLP 2014, Doha, Qatar, October*. Association for Computational Linguistics.
- Chung, Junyoung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the ACL, Berlin, Germany, August*.
- Crego, Josep Maria, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Ricciardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. 2016. Systran’s pure neural machine translation systems. *CoRR*, abs/1610.05540.
- Dyer, Chris, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, Atlanta, USA, June.
- Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions. *CoRR*, abs/1610.01108.
- Kingma, Diederik P. and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.
- Klein, G., Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.
- Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of ACL 2007, demonstration session, Prague, Czech Republic, June*.
- Melamed, I. Dan. 1995. Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Proceedings of the third Workshop on Very Large Corpora*, Cambridge, Massachusetts, USA.
- Och, Franz Josef and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1).
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, July.
- Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.
- Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA.
- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*.
- Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.
- Ziemski, Michal, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of LREC 2016, Portorož, Slovenia, May 23-28, 2016*.

Machine Translation at Booking.com: Journey and Lessons Learned

Pavel Levin
Booking.com
Amsterdam
pavel.levin
@booking.com

Nishikant Dhanuka
Booking.com
Amsterdam
nishikant.dhanuka
@booking.com

Maxim Khalilov
Booking.com
Amsterdam
maxim.khalilov
@booking.com

Abstract

We describe our recently developed neural machine translation (NMT) system and benchmark it against our own statistical machine translation (SMT) system as well as two other general purpose online engines (statistical and neural). We present automatic and human evaluation results of the translation output provided by each system. We also analyze the effect of sentence length on the quality of output for SMT and NMT systems.

1 Introduction

Booking.com is one of the biggest ecommerce companies in the world, offering content and serving customers in over 40 different languages. Because the need for translated content is growing very fast (in line with the overall Booking.com growth), machine translation is becoming a very attractive solution to complement the traditional human translation services.

One of the main use cases for translation at Booking.com is translating property descriptions (hotels, apartments, B&Bs, hostels, etc.) from English to any of the other supported languages. Integrating a machine translation solution would potentially dramatically increase the translation efficiency by increasing its speed and reducing the time it takes for a translated property description to appear online, as well as significantly cutting associated translation costs.

This work describes our production NMT system as well as an earlier version SMT system

for two important language pairs: English-German and English-French. We benchmark the two in-house systems against each other and against two general purpose online engines (statistical and neural). Further we look at how the performance of our NMT and SMT systems varies with the sentence length.

2 Related work

Despite being relatively young, neural machine translation (NMT) has been quickly gaining popularity over statistical machine translation (SMT) both in academic circles and in the industry (Jean et al., 2015; Crego et al., 2016). The main reasons for this are much simpler and more elegant training pipelines, ability to address—at least in theory—some of SMT’s fundamental limitations (Cho et al., 2014; Sutskever et al., 2014) and of course as of recently the quality performance (Bojar et al., 2016; Cettolo et al., 2016; Junczys-Dowmunt et al., 2016; Wu et al., 2016).

Although there is a lot of active development in NMT research (Neubig, 2017; Sennrich et al., 2016), there have not been many demonstrations of NMT usefulness in real world scenarios (some of the exceptions include Wu et al., 2016 and Crego et al., 2016)).

In addition, in Section 4.4, we offer some empirical data related to the ongoing discussion (Cho et al., 2014; Bahdanau et al., 2015) around the NMT performance as a function of sentence length.

3 Experimental settings

In this section, we describe configuration and design of statistical and neural MT engines in exploration and production, as well as the data our experiments were based on.

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

3.1 Data

Experiments were conducted using internal parallel data extracted from Booking.com translation memories that contain original property descriptions in English and their translations into German and French. Note that because translation coverage vary for German and French markets, amount of training data available for English-German and English-French differ.

Basic statistics of the tokenized training corpus can be found in Table 1. Note that ASL stands for average sentence length, M stands for million, K stands for thousand.

| Language | Sent. | Words | Voc. | ASL |
|----------------|-------|-------|------|------|
| English-German | | | | |
| German | 10.5M | 171M | 845K | 16.3 |
| English | | 174M | 583K | 16.5 |
| English-French | | | | |
| French | 11.3M | 193M | 588K | 17.7 |
| English | | 188M | 581K | 16.7 |

Table 1: Statistics of the training corpora.

The development corpus was 10K segments long for NMT training and contained 5K segments for SMT tuning.

3.2 SMT

The SMT system we used was based on the open-source *MOSES* toolkit (Koehn et al., 2007). We followed the guidelines, as detailed on the *MOSES* web page¹. Word alignment was estimated with *GIZA++* tool (Och, 2003). A 5-gram target language model was estimated using the *IRST* LM toolkit (Federico et al., 2008). The reordering method used in the *Moses*-based MT systems is *MSD* (Tillman, 2004), coupled with a distance-based reordering.

3.3 NMT

Our neural machine translation system is based on *OpenNMT* (Klein et al., 2017) implementation of the global attention Sequence-to-sequence (*seq2seq*) model on words level (Luong et al., 2015). In the last few years the family of *seq2seq* models has been gaining significant momentum in the machine translation world. The idea behind this class of models is to encode the source

¹<http://www.statmt.org/moses/>

sequence—usually as a fixed length vector—using some type of encoder and then to output the target sequence with a decoder, conditional on the encoded representation of the source. When trained jointly, the encoder and the decoder learn how to translate source to target (Sutskever et al., 2014).

In our system both the encoder and the decoder are long short-term memory (LSTM) recurrent neural nets (Hochreiter et al., 2017) with multiple hidden layers. The encoder LSTM reads each input sequence one token² at a time updating the internal representation of the sequence read so far. Those representations are essentially the LSTM hidden states. The final LSTM hidden state (after seeing the end of sequence $\langle /s \rangle$ token in the source) is then used to initialize the decoder LSTM whose task is to generate the output sentence, again one token at a time.

In addition to the simple recurrent neural net decoder we also used an attention mechanism because letting the decoder attend to relevant parts of the source input has been shown to dramatically improve translation quality (Bahdanau et al., 2014, Luong et al., 2015). The way attention works is as follows. At each time step of generating the output we assign a probability measure over the input tokens (“alignment weights”), which we use to take the weighted average of the input hidden states and feed the resulting “context” vector as an additional input to the decoder for the current time step. The alignment weights are computed by a shallow neural network which takes the current target LSTM hidden state and each source LSTM hidden states as inputs (Luong et al., 2015).

3.3.1 NMT Training

As is common (e.g. Sutskever et al., 2014; Luong et al., 2015) we use 4-layered LSTMs for both the encoder and the decoder with the vocabularies of 50K most common words for both languages (following Luong et al., 2015)³. All out-of-vocabulary words were encoded with a special $\langle \text{unk} \rangle$ symbol (following Sutskever et al., 2014 and Luong et al., 2015). Both the dimensionality of word embeddings and the LSTM hidden layer are of size 1000. Dropout (Srivastava et al., 2014) between the LSTM layers was set to 0.3. We ex-

²A token can be either a vocabulary word, a punctuation mark, beginning of the sentence $\langle s \rangle$, end of sentence $\langle /s \rangle$, blank $\langle \text{blank} \rangle$ or out-of-vocabulary word $\langle \text{unk} \rangle$.

³In our earlier experiments we tried using less than four layers but, as expected, got significantly worse results.

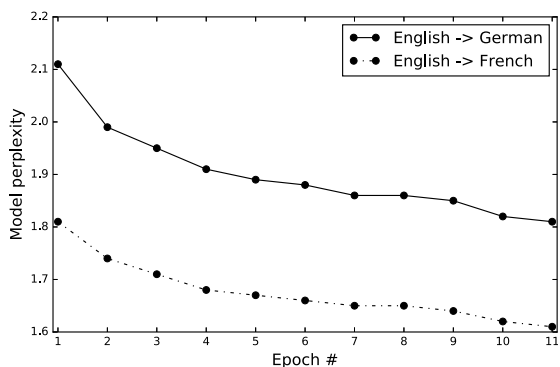


Figure 1: Model perplexity (measured on the validation set) as a function of training epoch.

cluded any sentences of length > 50 words. The total number of parameters the model has is just over 220 million which lets us train in batches of size 250 on a single NVIDIA Tesla K80 GPU. For a typical training corpus of size 10-11M sentence pairs, each epoch takes approximately 2 days.

The model parameters were fitted using normal stochastic gradient descent procedure, starting with learning rate 1, and halving it whenever the decoder perplexity on the validation set (see Figure 1) for the current epoch is not decreased. We also did BLEU score calculations on the validation set after each epoch. Our decision about when to stop was done on a case by case basis and was guided mainly by BLEU score improvements over previous epochs, manual analysis of a few hand-picked "sensitive" sentences⁴ and of course our product development time constraints. Depending on a particular language pair and corpus size we would usually stop after anywhere between 5 and 13 epochs.

Because of the closed vocabulary nature of our NMT system, the output translation may contain `<unk>` tokens for predicted out-of-vocabulary words. To get the final version of the translation, therefore, we follow a postprocessing step in which we look at the attention score distribution of the output `<unk>` token over the source words and copy the one with the maximal value. Because in our use case (hotel descriptions) those words are most commonly names of places, this heuristic of

⁴For example in some languages "The neighbourhood is very nice and safe" is often translated to mean "There is a safe installed in this very nice neighbourhood" during the early learning stage because the word *safe* is very often used to mean a *safe box* in our property descriptions.

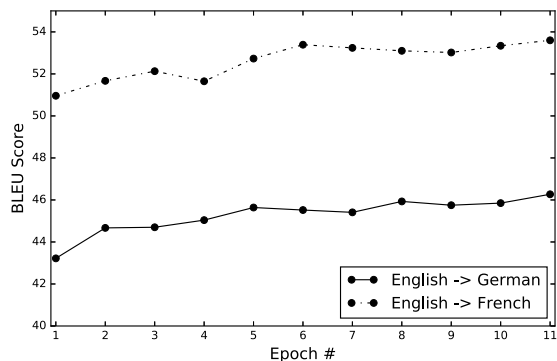


Figure 2: BLEU score (measured on the test set) as a function of training epoch.

copying the most probable word from the source usually works quite well in practice. Here is an example of a translated sentence with multiple out-of-vocabulary words:

| | |
|---|--|
| Source | <i>Offering a restaurant, Hodor Eco-lodge is located in Winterfell.</i> |
| Human Translation | <i>Das Hodor Eco-Lodge begrüßt Sie in Winterfell mit einem Restaurant.</i> |
| Raw Output | <i>Das <unk><unk> in <unk> bietet ein Restaurant.</i> |
| Output with <code><unk></code> replaced | <i>Das Hodor Eco-lodge in Winterfell bietet ein Restaurant.</i> |

4 Evaluation

We compared translation quality delivered by 4 MT systems: in-house *SMT* and *NMT* as described in the previous section, as well as statistical and neural online general purpose engines (*SGPMT* and *NGPMT*) trained on the general domain data.

4.1 Automatic evaluation

We used BLEU metric as the primary automatic metric of translation quality evaluation. BLEU (Papineni et al., 2002) shows the number of words shared between MT output and human-made reference, benefiting sequential words and penalizing very short translations. BLEU scores were calculated on the basis of truecased and detokenized test datasets of 10K segments and one reference translation. The evaluation conditions were case-sensitive and included punctuation marks.

In our analysis of the effect of the sentence length on machine translation quality (Section 4.4) we also use Word Error Rate (WER). WER is a variation of the word-level Levenshtein distance

measuring the distance between the target and the reference sentences by counting the insertions, deletions and substitutions necessary to go from one to the other.

4.2 Manual evaluation

We validated results of our findings with human *Adequacy-Fluency (AF)* evaluation applying a 4-level scale to both Adequacy and Fluency as described in the TAUS Adequacy/Fluency Guidelines⁵.

Evaluators (3 per language), which are native speakers of the target language, were provided with the original text in English and the MT hypotheses. They were asked to assess the quality of 150 randomly selected lines from the test corpus translated by the four MT systems under consideration. The evaluators were not aware of which system produced which hypothesis.

4.3 Evaluation results

Table 2 presents BLEU and AF scores of our benchmarking experiment. Figures 3 and 4 shows the human evaluation results.

| Translation | BLEU | Adequacy | Fluency |
|----------------|-------|----------|---------|
| English-German | | | |
| SMT | 35.24 | 3.62 | 3.15 |
| NMT | 45.64 | 3.90 | 3.78 |
| SGPMT | 27.63 | 3.57 | 3.37 |
| NGPMT | 31.45 | 3.65 | 3.57 |
| Human | – | 3.96 | 3.82 |
| English-French | | | |
| SMT | 35.80 | 3.40 | 3.28 |
| NMT | 52.73 | 3.67 | 3.40 |
| SGPMT | 30.25 | 3.32 | 3.31 |
| NGPMT | 32.18 | 3.78 | 3.41 |
| Human | – | 3.70 | 3.75 |

Table 2: Evaluation results.

We observed that:

- According to the BLEU scores, NMT consistently outperform all other engines with a significant margin;
- Both neural systems (NMT and NGPMT) consistently outperform their statistical coun-

terparts (SMT and SGPMT) according to both automatic and manual metrics;

- The performance of general purpose engines is worse than that of the in-house engines in case of English-German in terms of both BLEU and A/F scores, while in case of English-French, there is a mismatch between BLEU and adequacy score. In the latter case, NGPMT outperformed all other engines and surprisingly human translators in terms of adequacy, which may be an artifact of the small sample size, as well as the subjectivity of the metric itself.
- The fluency performance of the NMT engines is not far from human level for English-German, while for English-French adequacy delivered by both neural engines (in-house NMT and NGPMT) is approximately at the human translation level.

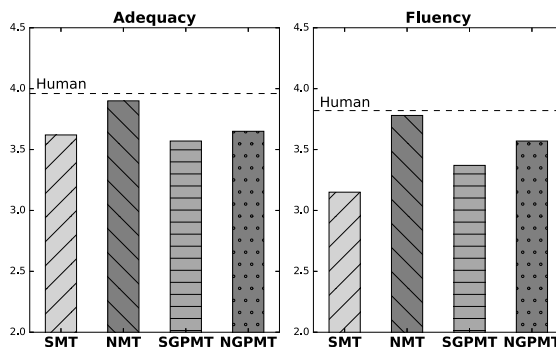


Figure 3: AF results for English-German for the four systems and a human translation benchmark.

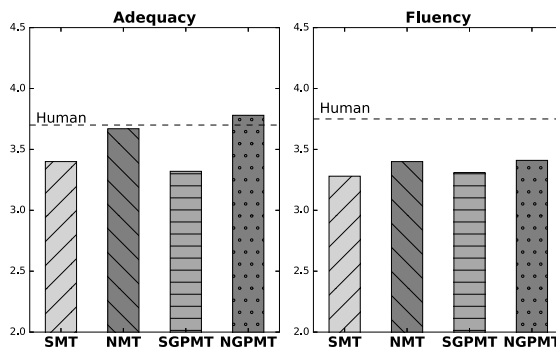


Figure 4: AF results for English-French for the four systems and a human translation benchmark.

⁵<https://www.taus.net/academy/best-practices/evaluate-best-practices/adequacy-fluency-guidelines>

4.4 Sentence length analysis

Multiple studies (Cho et al., 2014a) find that translation quality drops significantly when NMT translates long sentences. The primary cause being that, for longer sentences, the fixed-size vector representations of source sentences by encoder struggles to capture all cues for decoder to generate appropriate translations. Attention mechanism helps to combat this problem to a certain extent by selectively focusing on relevant parts of the source sentence while translating, instead of just relying on a fixed vector representation. There are other approaches as well, for example breaking long sentence into shorter phrases before translation (Pouget-Abadie et al., 2014). We were interested to see the correlation between sentence length and the machine translation quality in our data, particularly whether SMT outperforms NMT for longer sentences.

We segmented our tokenized test corpus into 10 bins according to lengths of the source sentences. Each bin contained roughly 1,000 sentences. We then ran BLEU score and negative word error rate evaluation separately on each of the 10 batches. Results are displayed in Figures 5-8.

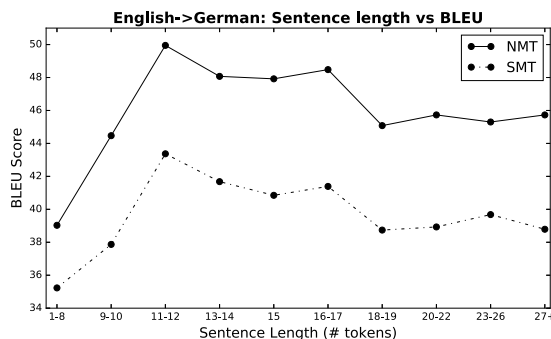


Figure 5: Sentence Length vs. Quality (BLEU) for SMT & NMT in English-German translation.

Our observation is two-fold:

- both systems, roughly followed the same trend. Quality was low for very small sentences i.e. 1-8 tokens, then increased with the length as the context helped in translation, but reached a peak soon around 11-17 tokens, and thereafter degraded for longer sentences;
- even for longer sentences though performance degraded, our NMT system outperformed SMT.

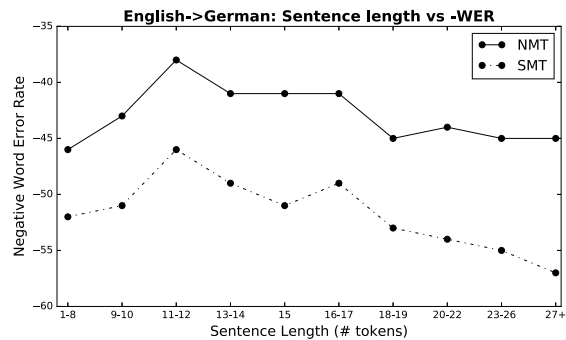


Figure 7: Sentence Length vs. Quality (-WER) for SMT & NMT in English-German translation.

We ran the same experiment on English to French translations, and observed very similar trends (See Figure 6).

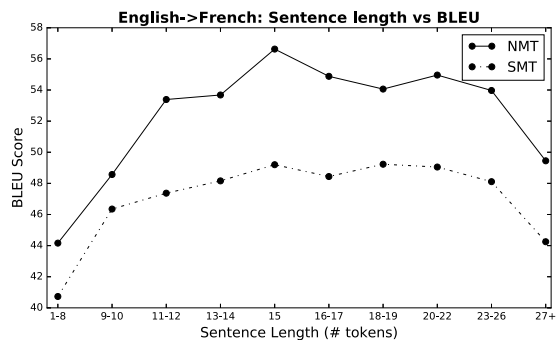


Figure 6: Sentence Length vs. Quality (BLEU) for SMT & NMT in English-French translation.

We used WER as a secondary metric to validate the results of BLEU analysis which could be biased for shorter sentences. We report negative WER to make this into a precision measure.

As can be seen in Figures 7 and 8, the results are very similar to those in Figures 5 and 6. This further corroborates our observations outside the constraints of the BLEU score.

5 Conclusions and future work

The main three findings of this study are: (1) neural MT technology consistently outperforms statistical; (2) in case of German, in-house NMT is also better than online general purpose engines in our application; (3) fluency of NMT is close to human translation level; and (4) in our application the relative performance of NMT against SMT does not degrade with increased sentence length.

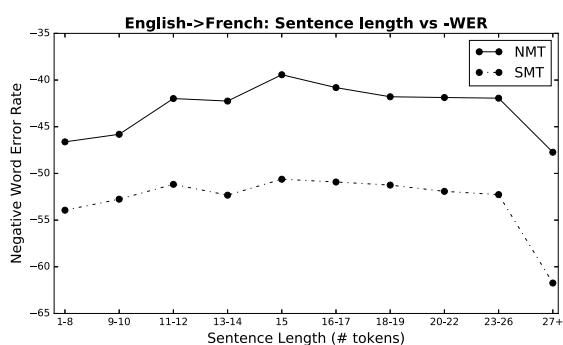


Figure 8: Sentence Length vs. Quality (-WER) for SMT & NMT in English-French translation.

Our future research directions include further improving our in-house NMT system in two important ways. The first one is the improved treatment of unknown and rare words which are particularly important to us because of a large number of named entities in our corpora, such as landmark or hotel names. The problem becomes even bigger with user generated content which may contain many misspellings and abbreviations. The second direction of research is improving our ability to identify business sensitive translation errors (e.g. “free” being translated to “available”).

References

- Bahdanau D., Cho K. and Y. Bengio. 2015. *Neural machine translation by jointly learning to align and translate.*, In Proceedings of ICLR’15, San Diego, CA, USA.
- Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., ... & Negri, M. 2016. *Findings of the 2016 Conference on Machine Translation (WMT16).*, In Proceedings of WMT at ACL 2016, Berlin, Germany.
- Cettolo M., Niehues J., Stuker S., Bentivogli L., Cattoni R. and M. Federico. 2016. *The IWSLT 2016 Evaluation Campaign.*, In Proceedings of IWSLT’16. Seattle, WA, USA.
- Cho K., Merriënboer B., Bahdanau D. and B. Yoshua. 2014. *On the properties of neural machine translation: Encoder-Decoder approaches.*, In Proceedings of SSST-8, Doha, Qatar.
- Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., ... & Enoue, S. 2016. *SYSTRAN’s Pure Neural Machine Translation Systems.*, Technical report.
- Federico M., Bertoldi N. and M. Cettolo. 2008. *IRSTLM: an open source toolkit for handling large scale language models.*, In Proceedings of Interspeech, Brisbane, Australia.
- Hochreiter, S. and J. Schmidhuber. 1997. *Long short-term memory.*, Neural computation, 9(8), 1735-1780.
- Jean S., K. Cho, R. Memisevic, and Y. Bengio. 2015. *On using very large target vocabulary for neural machine translation.*, In Proceedings of ACL’15, Beijing, China.
- Junczys-Dowmunt M., Dwojak T. and H. Hoang. 2016. *Is neural machine translation ready for deployment? A case study on 30 translation directions.*, CoRR, abs/1610.01108.
- Klein G., Kim Y., Deng Y., Senellart J. and A. Rush. 2017. *OpenNMT: Open-Source Toolkit for Neural Machine Translation.*, Technical report.
- Koehn, Ph., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. *Moses: open-source toolkit for statistical machine translation.*, In Proceedings of ACL’07, Prague, Czech Republic.
- Luong M., Pham H. and C. Manning. 2015. *Effective Approaches to Attention-based Neural Machine Translation.*, In Proceedings of ACL 2015, Beijing, China.
- F. Och. 2003. *Minimum error rate training in statistical machine translation.*, In Proceedings of ACL’03, Sapporo, Japan.
- Papineni, K., S. Roukos, T. Ward, and W. Zhu. 2002. *Bleu: a method for automatic evaluation of machine translation.*, In Proceedings of ACL’02, Philadelphia, PA, USA.
- Pouget-Abadie, J., Bahdanau, D., van Merriënboer, B., Cho, K. and Y. Bengio. 2014. *Overcoming the curse of sentence length for neural machine translation using automatic segmentation.*, In Proceedings of SSST-8, Doha, Qatar.
- Sennrich R., Haddow B. and A. Birch. 2016. *Edinburgh Neural Machine Translation Systems for WMT 16.*, In Proceedings of WMT’16. Berlin, Germany.
- Sutskever, I., Vinyals, O., & Le, Q. V. 2014. *Sequence to sequence learning with neural networks.*, Advances in neural information processing systems.
- C. Tillman. 2004. *A unigram orientation model for statistical machine translation.*, In Proceedings of HLT-NAACL’04, Boston, MA, USA.
- Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Klingner, J. 2016. *Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.*, Technical report.

MMT: New Open Source MT for the Translation Industry

Nicola Bertoldi¹, Roldano Cattoni¹, Mauro Cettolo¹, Amin Farajian¹, Marcello Federico¹,

Davide Caroselli², Luca Mastrostefano², Andrea Rossi², Marco Trombetti²

Ulrich Germann³, David Madl^{2,3}

1) Fondazione Bruno Kessler, Trento, Italy

2) Translated srl, Rome, Italy

3) University of Edinburgh, United Kingdom

Abstract

MMT is a new open source machine translation software specifically addressing the needs of the translation industry. In this paper we describe its overall architecture and provide details about its major components. We report performance results on a multi-domain benchmark based on public data, on two translation directions, by comparing MMT against state-of-the-art commercial and research phrase-based and neural MT systems.

1 Introduction

MMT aims to consolidate the current state-of-the-art technology into a single easy-to-use product, evolving it and keeping it open to integrate the new opportunities in machine intelligence, such as deep learning. MMT was designed and developed to overcome four technology barriers that have so far hindered the wide adoption of machine translation software by end-users and language service providers: (1) long training time before a MT system is ready to use; (2) difficulty to simultaneously handle multiple domains; (3) poor scalability with data and users; (4) complex installation and set-up. As we will describe in the next section, MMT on the contrary is very fast to train, it instantly adapts to a specific translation domain, it is designed to scale well with data and users, and, finally, it is very easy to install and configure.

This paper describes the current advanced prototype of MMT, a statistical phrase-based machine translation system, which already covers all the

above presented features and has being field-tested in real industrial settings. A comprehensive documentation of MMT, including installation manual, is available in the official website.¹

We also report experiments conducted on a public multi-domain benchmark covering technical translations from English to German and English to French.

2 Main Features of MMT

2.1 MMT Can Ingest New Data Instantly

MMT uses high-performance embedded databases² to store parallel and monolingual language data and associated statistics. Instead of pre-computing feature function scores, these are computed on the fly, at translation time, from raw statistics. Thanks to its implementation with databases, MMT is a fully incremental MT system, that can ingest new parallel data while in use, very quickly and without any interruption nor re-training.

2.2 MMT Can Adapt Itself to the Task

Input to the system can be augmented with a snippet of surrounding text. This context information is leveraged by MMT to adapt the translation process to a specific domain. Adaptation is performed on the fly by biasing the data sampling process underlying the computation of the feature functions towards training data that is close to the provided context. (see Sec. 3.2.5 below).

2.3 MMT Scales Easily

MMT is designed as a distributed multi-node architecture, with cloud deployment in mind. There-

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

¹<http://www.modernmt.eu>

²RocksDB: <https://github.com/facebook/rocksdb>

fore it can scale dynamically in response to current demand, simply by adding or removing MMT nodes in the cluster. Single-host deployment for small use cases is also possible.

2.4 MMT Is Easy to Set Up

MMT is distributed as a ready to install package either through Docker, or directly from binary files.³ In addition, instructions for installing MMT from source code are also available.

3 System Architecture

3.1 Distributed Infrastructure

MMT's distributed architecture is based on a Leader-Follower network where nodes form a cluster through the *Hazelcast* framework.⁴

Most inter-node communications are carried out through shared in-memory data structures, suitable for small data and thus employed for service communications and load balancing. When a node receives a translation request, it sends it to an Executor that transparently extracts jobs from its internal pool, chooses a worker node, and finally redirects the translation output to the original requesting node.

Bigger volumes of data are handled by the nodes in persistent messaging queues (*Kafka*⁵) and in an internal database (*Cassandra*⁶). The former is mostly used to distribute newly ingested resources, which any node may import during its life cycle; the latter to handle persistent application-internal data, like domains' and contributions' metadata.

Both the Followers and the Leader expose the same REST APIs; the Leader in addition hosts the messaging queue server and the internal database. To join the cluster, a worker node only needs to know the Leader IP.

All nodes in a cluster must have the same initial configuration. It is recommended to perform the initial training on a single node and share the resulting models to the others manually. Once a node has received this initial configuration it can join the cluster at any time, and will automatically receive any new updates through the above mentioned messaging channels.

³Currently we distribute binaries for Ubuntu.

⁴<https://hazelcast.com>

⁵<https://kafka.apache.org>

⁶<http://cassandra.apache.org/>

3.2 MT Worker Nodes

The core architecture of each node is composed by several interacting modules.

3.2.1 Tag Management

XML tags occurring in the input text are removed and a map between the tags and their positions is stored. According to the output and the word alignment provided by the decoder, the XML tags are re-introduced after applying a few consistency checks and heuristics.

3.2.2 Numerical Expression Management

Numerical expressions, like numbers, currencies, dates, etc., are transformed into format- and position-dependent placeholders, and a map between the actual numerical values and their placeholders is stored. The placeholders found in the output of the decoder are finally transformed back into their actual numerical values using the word alignment and a few heuristics to resolve possible ambiguities.

3.2.3 Tokenization and De-tokenization

The tokenizer and de-tokenizer, based on third-party software credited in the official documentation, support 45 languages through a unique entry point.

3.2.4 Central Vocabulary

Internally, words are represented by integer IDs managed by a joint vocabulary for source and target language that allows incremental updates.

3.2.5 Context Analyzer

The Context Analyser (CA) is in charge of identifying training data that best matches the provided input context. To this purpose, parallel data is sharded into chunks according to the customer, subject area, genre, etc. In a very loose use of the term, we refer to these shards as “domains”.

When queried, the Context Analyzer (CA) computes a ranked list of matching domains, with associated weights⁷ that indicate how closely they match the input text. The CA is built on top of the *Apache Lucene*⁸ framework, in particular *Lucene's Inverted Index* data structure. The *Inverted Index* is complemented with a filesystem-based data structure, called *Corpora Storage*, where all the original indexed data are stored: one file corresponds to one

⁷The weights are computed by means of the *tf-idf* metrics and the *Cosine Similarity*.

⁸<https://lucene.apache.org/core/>

data shard and new content can be appended to the corresponding storage file. In this way, the *Corpora Storage* always maintains the most updated version of the data. When required, the *Inverted Index* is synchronized with the *Corpora Storage* by re-indexing the changed domains and adding the new ones.⁹ This activity does not interfere with the look-up operations of the CA; *Lucene* allows concurrent reads and writes, always ensuring data availability and consistency.

3.2.6 Word Aligner

The Word Aligner (WA) performs many-to-many word-to-word alignment of sentence pairs.

The WA is built on top of *FastAlign* (Dyer et al., 2013); it computes two directional alignments, and symmetrizes them according to the *grow-diag-and-final* policy.¹⁰ The WA is multi-threaded and permits persistent storage and re-loading of the alignment models after training. It is able to align individual new sentence pairs without re-training the models. The WA is trained on all parallel data available at training time, irrespective of their domain.

3.2.7 Decoder

The decoder developed in MMT is an enhanced version of the phrase-based decoder implemented in *Moses* (Koehn et al., 2007). Differently from *Moses*, MMT generates and scores translation hypotheses *according to the context of the input sentence*. In particular, the decoder queries its models with the domain weights computed by the CA from the input context.

3.2.8 Translation Model

The MMT Translation Model (TM) is an enhanced re-implementation of the suffix array-based phrase table by Germann (2015). Its original implementation creates a phrase table at run-time by sampling sentences from the pool of word-aligned parallel data with a uniform distribution, extracting phrase pairs from them, and computing their scores on the fly. The new version provides two enhancements. First, instead of a suffix array, it relies on a DB-backed prefix index of the data pool, thus allowing for fast updates (i.e., insertions and deletions of word-aligned parallel data). Second, it keeps track of the domains from which phrase pairs are

extracted and performs *ranked sampling*: extracted phrases are ranked by their relevance (via the domain they were observed in). Translation scores are then obtained by going down the ranked list until a sufficient number of samples has been observed. Hence, by associating with all sentence pairs of each domain the corresponding weight, the TM selects and scores phrase pairs giving priority to the best-matching domain.

The TM scores are the forward and backward probabilities at lexical and phrase level; the phrase-level probabilities are weighted according to the domain weights.

3.2.9 Lexicalized Reordering Model

The same incremental DB-based implementation of the TM is also exploited by the Lexicalized Reordering Model. Similarly, its scores are computed on the fly exploiting the counts extracted from the sampled sentences and the corresponding word alignments, and some global counts stored in the DB. The scores are the forward and backward probabilities for monotone, swap, and discontinuous orientations.

3.2.10 Language Model

The MMT LM linearly combines a static background LM with a context-adaptive LM.

The static LM, implemented with the KenLM toolkit (Heafield et al., 2013), features 5-grams, interpolation of lower-order models, and the Kneser-Ney smoothing technique. It is trained on all monolingual target text regardless the domain information, and does not change over time.

The context-adaptable LM is an *internal mixture* LM (Federico and Bertoldi, 2001) using domain-specific counts extracted from the corresponding data shards and the weights of the CA.¹¹ The LM features 5-gram statistics, interpolation of lower-order models, and Linear Witten-Bell smoothing. Noteworthy, n -gram probabilities are not pre-estimated in the training phase, but computed on the fly, by exploiting domain-specific n -gram and global statistics, which are stored in a key-value DB.

3.2.11 Manager

The Manager controls the communication between all components to satisfy the translation and updating requests.

⁹For performance reasons, synchronization is subject to a time-out.

¹⁰<http://www.statmt.org/ Moses>.

¹¹For efficiency, only the LMs actually activated by the CA are included in the mixture.

3.3 Functionalities

From a functional perspective four phases can be identified, namely training, tuning, updating and translation.

3.3.1 Training

The training phase sets up MMT starting from a collection of bilingual and (possibly) monolingual corpora, which can be domain-specific or not-specialized. In particular, the DBs required by CA, LM and TM, are created, which respectively exploit only the source side, only the target side, or both sides of the training data. Texts are pre-processed by the corresponding modules.

3.3.2 Tuning

MMT implements a standard Minimum Error Rate Training procedure (Och, 2003) to optimize the decoder feature weights.

3.3.3 Updating

Once a system is trained, new bilingual data can be added to it,¹² either to an existing domain or establishing a new one. This operation is performed by updating the corresponding DBs of the CA, the TM and the LM. Such updates do not interfere with the translation process.

3.3.4 Translation

In a standard scenario, MMT translates one document as follows; it (i) processes and sends to the CA the whole document, considered as context for all its sentences, and gets the domain weights, (ii) pre-processes and sends all sentences to the available decoders, independently and in parallel, and gets their translations, and (iii) post-processes and returns all translations by re-creating the original document layout. More generally, however, MMT is able to translate any single sentence provided with some context, even made of a single word.

3.4 APIs

MMT system exposes APIs for its integration in third-party software. Plug-ins are under advanced construction to permit the integration of MMT in various commercial CAT tools.

¹²For instance, new data can be a translation memory of a new customers, or the post-edits of professional translators.

4 Evaluation

4.1 Points of Comparison

Although the main scope of the paper is the description of the components and features of the MMT system, an experimental comparison is proposed against a few popular MT engines. In particular, two phrase-based MT systems, Moses and the Google's web translation service, and two neural MT systems.

4.1.1 Moses

A Moses (Koehn et al., 2007) engine was trained on the concatenation of all the available training corpora. Word alignment models were trained with FastAlign (Dyer et al., 2013) and a 5-gram language model was estimated by means of the KenLM toolkit (Heafield et al., 2013). Feature weights were tuned with batch MIRA (Cherry and Foster, 2012) to maximize BLEU on the pooled dev sets. No adaptation was performed.

4.1.2 GT

The Google web translation service (GT), one of the most used engines by the translation industry, was accessed through its public API¹³ at the beginning of March 2017.

4.1.3 Neural MT Systems

We developed two neural MT systems using an in-house branch (Farajian et al., 2017) of the Nematius toolkit¹⁴ implementing the encoder-decoder-attention model architecture by (Bahdanau et al., 2014). This first system is a *generic NMT* (gNMT) system trained on all the pooled training data. Then, following common practice (Luong and Manning, 2015), *adapted NMT* (aNMT) systems were trained for each domain by tuning the generic NMT system to the training data of each domain.

4.2 Experiments

We present experiments carried out on two translation tasks involving a collection of eight domain-specific corpora and two translation direction, English-French and English-German. When comparing the four types of MT systems, we consider translation quality (BLEU), training time, tuning time, and translation speed (seconds per sentence).

¹³<https://www.googleapis.com/language/translate/v2>

¹⁴<https://github.com/rsennrich/nematius>

| | | English-French | | | English-German | | |
|-------|-----|----------------|------------|-------------|----------------|-------------|------------|
| | | segments | source | target | segments | source | target |
| train | dom | 1,332,972 | 17,581,131 | 19,297,282 | 1,004,214 | 15,772,744 | 14,427,002 |
| | gen | 4,255,604 | 92,363,974 | 101,236,914 | 4,165,505 | 104,489,832 | 98,381,272 |
| dev | dom | 3,527 | 46,640 | 52,484 | 3,073 | 37,023 | 35,187 |
| test | dom | 6,962 | 93,243 | 98,312 | 6,011 | 72,995 | 5,856 |
| | out | 4,503 | 104,831 | 111,050 | 5,168 | 111,331 | 106,443 |

Table 1: Statistics of training, dev and test sets for the English-French and English-German tasks: number of segments, source and target words. Figures refer to texts processed with the MMT modules.

| | | English-French | | | | | English-German | | | | |
|-------------|--|----------------|-------|-------|-------|-------|----------------|-------|-------|-------|-------|
| | | MMT | Moses | gNMT | aNMT | GT | MMT | Moses | gNMT | aNMT | GT |
| dom | | 62.48 | 61.78 | 49.23 | 63.00 | 43.62 | 48.27 | 48.51 | 37.41 | 48.95 | 31.37 |
| out | | 30.11 | 28.93 | 33.28 | – | 36.47 | 19.08 | 16.84 | 22.82 | – | 27.13 |
| training | | 1h | 10h | 100h | 100h | – | 1h | 10h | 100h | 100h | – |
| tuning | | 1h | 10h | – | 10h | – | 1h | 10h | – | 10h | – |
| translation | | 1s | 1s | 1s | 1s | 0.1s | 1s | 1s | 1s | 1s | 0.1s |

Table 2: Quality and speed performance of MMT and few competitor systems: BLEU scores on *dom* and *out* test sets for both English-French and English-German; overall time (order of magnitude in hours) to complete training and tuning ; the average time (order of magnitude in seconds) to translate one sentence.

4.2.1 Data

We consider eight publicly available parallel corpora as representatives of specific domains (*dom*): European Central Bank, Gnome, JRC-Acquis, KDE4, OpenOffice, PHP, Ubuntu, and UN documents.¹⁵ To increase the training data, two additional generic corpora (*gen*) were added to the pool, namely CommonCrawl¹⁶ and Europarl,¹⁷ which are not considered for the evaluation.

Each domain-specific corpus was randomly partitioned into training, development and test portions. Additional test data from WMT¹⁸ was prepared, in order to test the systems on out-of-domain data (*out*). Duplicate sentence pairs were removed from all dev and test sets. Statistics about training, dev and test sets are reported in Table 1.

4.2.2 Performance

Table 2 reports the translation quality performance (BLEU score), the overall computational cost for the compared systems to complete training and tuning, and the average time to translate one sentence in isolation. Time measures have to be taken with grain of salt because experiments were

not run under very comparable conditions. For instance, neural MT systems were run on PCs equipped with GPU cards, while MMT and Moses were run only on multi-core CPUs. Hence, the order of magnitude, which are definitely reliable, is reported.

4.2.3 Discussion

In the following, we try to point out strengths and drawbacks of MMT against the other competitors.

MMT vs Moses MMT and Moses perform similarly as expected in terms of translation quality, because both share the same phrase-base MT paradigm. MMT performs better than Moses in the out-of-domain condition thanks to its adaptability feature (+1.18 and +2.24 gains). While translation speed is comparable, training and tuning of MMT is one order of magnitude faster.

MMT vs gNMT The BLEU scores on the out-of-domain condition (*out*) confirms that NMT has a better generalization capability than MMT (-3.17 and -3.74 losses), while MMT performs largely better when translating domain specific data (+13.25 and +10.86 gains). The training time is largely in favour of MMT, hundreds of hours for gNMT versus few hours for MMT. Translation speeds are actually comparable.

¹⁵UN corpus is used only for English-French. All corpora are available in <http://opus.lingfil.uu.se>

¹⁶<http://www.statmt.org/wmt15/translation-task.html>

¹⁷<http://www.statmt.org/europarl/>

¹⁸*newstest2014* and *newsdiscuss2015* for English-French, and *newstest2015* and *newstest2016* for English-German.

MMT vs aNMT After adaptation of gNMT to each specific domain, aNMT systems perform on par with MMT on the in domain condition (dom). It is worth noticing, that under this condition distinct domain-specific NMT systems have to be tuned and translation should be run in a supervised way, by dispatching each test to the appropriate system. As a difference, MMT requires one system and does not require any domain labels at test time. The extra time needed to tune the aNMT systems on each domain is tens of hours.

MMT vs GT The comparison of MMT against Google Translate, show that the latter performs significantly better on the out of domain test (-6.35 and -8.05 losses), very likely due to the much larger training data available to the commercial system. On the contrary MMT perform largely better than GT on the in domain condition (+18.86 and +16.09 gains). With respect to translation speed, GT is significantly faster than MMT.

5 Conclusion

MMT aims to develop an innovative solution for the translation industry, by providing both better MT quality for post-editing as well as a better integration of MT with commercial CAT tools. MMT actually targets two use cases: (i) the enterprise use case, in which a language service provider or localisation department of a large company installs MMT to manage its translation workflow, and (ii) the translator use case, in which single translators install the MMT plugin in their favorite CAT tool and use MMT as their preferred source of suggestions/matches for their daily workflow.

For both scenarios MMT can provide machine translation technology that instantly adapts to the document to be translated and that quickly learns from the users' data – e.g. translation memories– and their post-editing work.

In this paper, we have presented an advanced phrase-based MT prototype of MMT, which shows competitive performance against similar approaches. In order to improve the generalization capability of MMT in operating conditions with a severe domain mismatch between testing and training data, work is in progress to integrate also neural MT in the final MMT release, which is planned for the end of 2017.

Acknowledgements

This work has been supported by the EC-funded project ModernMT (grant no. 645487).

References

- Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. “Neural machine translation by jointly learning to align and translate.” *arXiv preprint arXiv:1409.0473*.
- Cherry, Colin and George Foster. 2012. “Batch tuning strategies for statistical machine translation.” *Proc. NAACL-HLT*, 427–436. Montreal, Canada.
- Dyer, Chris, Victor Chahuneau, and Noah A. Smith. 2013. “A simple, fast, and effective reparameterization of IBM Model 2.” *Proc. of NAACL*, 644–648. Atlanta, GA, USA.
- Farajian, M. Amin, Marco Turchi, Matteo Negri, Nicola Bertoldi, and Marcello Federico. 2017. “Neural vs. Phrase-Based Machine Translation in a Multi-Domain Scenario.” *Proc. of EACL*. Valencia, Spain.
- Federico, Marcello and Nicola Bertoldi. 2001. “Broadcast news lm adaptation using contemporary texts.” *Proc. of Eurospeech*, 239–242.
- Germann, Ulrich. 2015. “Sampling phrase tables for the Moses statistical machine translation system.” *The Prague Bulletin of Mathematical Linguistics*, 104:39–50.
- Heafield, Kenneth, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. “Scalable modified kneser-ney language model estimation.” *Proc. of ACL (Volume 2: Short Papers)*, 690–696. Sofia, Bulgaria.
- Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. “Moses: Open source toolkit for statistical machine translation.” *Proc. of ACL (Interactive Poster and Demonstration Sessions)*, 177–180. Prague, Czech Republic.
- Luong, Minh-Thang and Christopher D Manning. 2015. “Stanford Neural Machine Translation Systems for Spoken Language Domains.” *Proc. of IWSLT*, 76–79. Da Nang, Vietnam.
- Och, Franz Josef. 2003. “Minimum error rate training in statistical machine translation.” *Proc. of ACL (Volume 1)*, 160–167. Sapporo, Japan.