Proceedings of MT Summit XVI,

Vol.2 Commercial MT Users and Translators Track

Masaru Yamada (Kansai University)
&
Mark Seligman (Spoken Translation, Inc.), Eds.

Co-hosted by

International Association for Machine Translation
http://www.eamt.org/iamt.php

Asia-Pacific Association for Machine Translation
http://www.aamt.info

Graduate School of Informatics, Nagoya University
http://www.is.nagoya-u.ac.jp/index_en.html

©2017 The Authors. These articles are licensed under a Creative Commons 3.0 license, no derivative works, attribution, CC-BY-ND.
Introduction

The Commercial MT Users and Translators Track at MT Summit XVI, to be held in Japan for the first time in 24 years, features twenty presentations in diverse fields of research from worldwide organizations including academic institutes, enterprises, and individuals in the translation and language technology industry. This Summit is the first since the practical deployment of neural machine translation (NMT), so many of the presentations involve related AI-driven MT technologies. Other studies go beyond traditional post-editing and efficiency scenarios to address the adoption of state-of-the-art MT across the industrial spectrum: topics include MT use cases in crisis scenarios or educational environments; terminology management and QA in systems combining customized MT engines; and many more. Some additional examples:

- Quality evaluation of NMT and comparison with SMT
- Detailed investigation of post-editing errors and efficacy
- Dictation of translation

Many presentations will document the acceptance and integration of neural machine translation technology and its application in real-life scenarios.

Commercial MT Users and Translators Track Co-Chairs
Masaru Yamada
Mark Seligman
**Commercial MT Users and Translators Track Co-chairs**

Masaru Yamada  
Kansai University  
Mark Seligman  
Spoken Translation, Inc.

**Commercial MT Users and Translators Track Committee**

Jan Alexandersson  
Srinivas Bangalore  
Laurent Besacier  
Michael Carl  
Mike Dillinger  
Kurt Eberle  
Satoshi Enoue  
Raymond Flournoy  
Anthony Hartley  
Ron Kaplan  
Jeffrey Killman  
Isabel Lacruz  
Yves Lepage  
Rei Miyata  
Joss Moorkens  
Ricardo Muñoz  
Masaaki Nagata  
Sharon O'Brien  
Akiko Sakamoto  
Xiaodong Shi  
Michel Simard  
Midori Tatsumi  
Carlos Teixeira  
Kirti Vashee  
Yuji Yamamoto
Contents

Page
1 Zero-shot translation for low-resource Indian languages
   Giulia Mattoni, Pat Nagle, Carlos Collantes and Dimitar Shterionov
11 Feature-rich NMT and SMT post-edited corpora for productivity and evaluation
   tasks with a subset of MQM-annotated data
   Kim Harris, Lucia Specia and Aljoscha Burchardt
13 Usability of web-based MT post-editing environments for screen reader users
   Silvia Rodríguez Vázquez, Sharon O'Brien and Dónal Fitzpatrick
26 Live presentations to a multilingual audience: personal universal translator
   Chris Wendt
27 Towards a full-scale neural machine translation in production:
   the Booking.com use case
   Pavel Levin, Nishikant Dhanuka, Talaat Khalil,
   Fedor Kovalev and Maxim Khalilov
38 The Interact Project and Crisis MT
   Sharon O'Brien, Chao-Hong Liu, Andy Way, João Graça, André Martins,
   Helena Moniz, Ellie Kemp and Rebecca Petras
49 A Case Study of Machine Translation in Financial Sentiment Analysis
   Chong Zhang
59 A New Methodology to Maximize the Strength of SMT and NMT
   Yu Gong and Demin Yan
67 Rule-based MT and UTX Glossary Management – Honda's Case Dealing with
   Thousands of Technical Terms
   Saemi Hirayama and Yuji Yamamoto
79 A detailed investigation of Bias Errors in Post-editing of MT output
   Silvio Picinini and Nicola Ueffing
91 Terminology-based post-editing of neural MT using the structured glossary data
   format, UTX
   Yuji Yamamoto
109 Harvesting Polysemous Terms from e-commerce data to enhance QA
   Silvio Picinini
116 Translation Dictation vs. Post-editing with Cloud-based Voice Recognition: A Pilot Experiment
   Julián Zapata, Sheila Castilho and Joss Moorkens

130 Will Neural MT be a Breakthrough in Terms of Post-Editing Productivity in English-to-Japanese Technical Translation?
   Tsunao Mikasa and Nobuko Kasahara

142 The Impact of MT Quality Estimation on Post-Editing Effort
   Carlos S. C. Teixeira and Sharon O'Brien

154 Utilizing Neural MT Engines in Industrial Translation
   Toru Shishido

166 Comparative Evaluation of NMT with Established SMT Programs
   Lena Marg, Naoko Miyazaki, Elaine O'Curran and Tanja Schmidt

179 Journey around Neural Machine Translation quality
   Marco Ganci

206 A Reception Study of Machine Translated Subtitles for MOOCs
   Ke Hu, Sharon O'Brien and Dorothy Kenny

214 TraMOOC - Translation for Massive Open Online Courses: Recent Developments
Zero-Shot Translation for Indian Languages with Sparse Data

Giulia Mattoni  
giuliam@kantanmt.com
Pat Nagle  
patn@kantanmt.com
Carlos Collantes  
carlosc@kantanmt.com
Dimitar Shterionov  
dimitars@kantanmt.com
KantanMT.com, INVENT Building, Dublin City University Campus, Dublin 9, Dublin, IRELAND

Abstract

Neural Machine Translation (NMT) is a recently-emerged paradigm for Machine Translation (MT) that has shown promising results as well as a great potential to solve challenging MT tasks. One such a task is how to provide good MT for languages with sparse training data. In this paper we investigate a Zero Shot Translation (ZST) approach for such language combinations. ZST is a multilingual translation mechanism which uses a single NMT engine to translate between multiple languages, even such languages for which no direct parallel data was provided during training.

After assessing ZST feasibility, by training a proof-of-concept engine ZST on French→English and Italian→English data, we focus on languages with sparse training data. In particular, we address the Tamil→Hindi language pair. Our analysis shows the potential and effectiveness of ZST in such scenarios.

To train and translate with ZST engines, we extend the training and translation pipelines of a commercial MT provider – KantanMT – with ZST capabilities, making this technology available to all users of the platform.

1 Introduction

Nowadays Machine Translation (MT) is an essential tool for the translation industry. The most used MT paradigms are Phrase-based Statistical Machine Translation (PBSMT) (Koehn et al., 2007) and Neural MT (Bahdanau et al., 2014; Sutskever et al., 2014; Cho et al., 2014). While PBSMT has been the state-of-the-art both in academia and industry for the last decade, recently NMT has showed great potential and in many cases has surpassed PBSMT (Bentivogli et al., 2016; Junczys-Dowmunt et al., 2016; Chung et al., 2016; Shterionov et al., 2017).

NMT, similar to PBSMT, is a data-driven MT paradigm, making it strongly dependent on the parallel data used for training. That is, the translation quality of an NMT system correlates with the quality and quantity of the training corpora. Freely accessible parallel corpora are available from various providers, such as: Opus1, DGT-EC (European Commission)2 and Linden/Clarin repository.3 Within the industry, MT

1http://opus.lingfil.uu.se/
3https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-625F-0
systems are typically built with proprietary data – i.e., data with restricted access, provided by a translation vendor and tailored towards specific translation task(s) mainly because of data confidentiality requirements and/or because the data includes terminology and style that are specific for the translation task. Often, to build a custom MT system (i.e., an MT system that is customised according to a translation vendor’s requirements) that can produce high-quality translations, proprietary and non-proprietary data are concatenated.

For some language pairs, however, there is not enough available parallel data (proprietary or non-proprietary) to build MT systems of high quality to meet users’ requirements. This specifically applies to minority or low-resource languages – languages that have a low population density, are under-taught or have limited written resources or are endangered – as well as to language pairs with sparse training data – language pairs for which there have not been documented (human) translations that can be used as training data. Sparsity of training data for language pairs such as, e.g. Tamil or Hindi, which by itself are not low-resource languages, is a phenomenon that hinders the MT industry.

Aiming to overcome the data sparsity within the spectrum of the Indian languages, this paper investigates a zero-shot translation (ZST) (Johnson et al., 2016) strategy for the Hindi and Tamil languages. We built ZST engines on available parallel data to and from English (we use English as an intermediate or a pivot language).

To determine the viability and potential of ZST we first build a proof-of-concept (POC) ZST engine for high-resource languages (English, Italian, Spanish and French). Second, we build a ZST engine for Hindi and Tamil as well as Hindi and Tamil, using parallel corpora with English, Hindi and Tamil data as well as for English, Hindi and Tamil data to prove the applicability of ZST for sparse-data language pairs.

To determine the quality of these engines we compare their outputs with the results of (i) one-to-one NMT engines for the same language combinations and (ii) via a pivoting language. In particular, case (ii) boils down to using two different NMT engines – one that translates from the investigated source language into English and another that translates from English to the investigated target language.

We use the KantanMT\(^4\) platform to train and translate ZST engines. KantanMT is a custom Machine Translation (MT) platform that allows its users to build custom MT systems covering more than 75 languages. The analysis of the resulting quality is performed through comparison of quality evaluation metrics and the A/B testing interface of the KantanMT platform (KantanLQR\(^TM\)). As a provider of commercial MT solutions, the KantanMT platform is designed and tailored to train and deploy one-to-one translation engines (Phrase-based Statistical Machine Translation and NMT). The research and development of ZST engines imposes certain architectural requirements to the KantanMT platform. In this work, we also discuss the changes that such a platform requires in order to accommodate ZST technology.

The main contribution of this paper is two-fold: on the one hand it is the insights that we draw from our analysis of ZST as a means to tackle the problem of sparse data; on the other hand, we extend the pipeline of a commercial MT – KantanMT.com – making ZST available to KantanMT users.

This paper is organised as follows. In Section 2 we present relevant background and motivate our work; in Section 3 we discuss our MT training and translation pipeline, the changes we have done in order to accommodate ZST technology and we outline the data used to build our ZST engines; Section 4 is devoted to the analysis of the

\(^4\)www.kantanmt.com
translation capabilities of these engines; we conclude our work and present our future plans in Section 5.

2 Background and Motivation

2.1 Zero-shot Translation

Zero-shot translation (ZST) (Johnson et al., 2016) is an approach to train a single NMT engine to translate between multiple languages. Such a multilingual engine can translate from a source to a target language without having seen explicit parallel corpora for that specific language pair during training. ZST exploits transfer learning to overcome the need of building one-to-one translation engines. According to (Johnson et al., 2016), an NMT engine can be trained as a multilingual ZST engine$^5$ by simply augmenting the training data with a token before each segment stating the target language. In particular, a sentence $S^{L_1}$ in language $L_1$ aligned to a sentence $S^{L_2}$ in language $L_2$ will be augmented with a token $<2L_2>$. Following their findings we exploit a similar approach to augment each segment of the parallel training corpora with a token to indicate the target language. Moreover, we extend this data processing step to handle different tokenisation rules for each language correctly. We add one more token to indicate the language of origin$^6$ of the specific sentence that will be used during tokenisation.

In the work of (Johnson et al., 2016; Ha et al., 2016) a single shared attention mechanism and a single ‘universal’ encoder-decoder across all languages is used. Firat et al. (2016) also present a multilingual approach that uses a shared attention mechanism. However, they use multiple encoders/decoders for each source and target language. Aiming at smallest possible alterations of our training and translation pipelines we focus on the single encoder/decoder model with shared attention. Such an architecture does not impose any changes to our platform (i.e. KantanMT), except in the preprocessing (both before training and before translation) step.

In (Johnson et al., 2016), the authors prove that mixing language pairs with little and large available data into a single multilingual NMT model produces a considerable translation quality improvement of the low resource language. This translation capabilities are due to the fact that all the parameters of the multilingual model are implicitly shared by all the language pairs. The analysis on multilingual NMT and zero-shot (or zero-resource) translation, given by (Firat et al., 2016), investigates multiple strategies for multi-way, multilingual translation engines. They show that an NMT engine trained on parallel data without data between two languages translates very poorly for these two languages. In contrast, adding pseudo-parallel data for these two languages to fine-tune the engine improves significantly the quality. They also, investigate a more basic multilingual NMT engine – trained on two parallel corpora (with or without a fine-tuning corpus) and is focused to translate between two of these language pairs, in contrast to (Johnson et al., 2016) where the focus is on translating a plethora of languages with one engine.

Motivated by the promising results documented in the aforementioned publications, our main objective is to demonstrate that ZST is particularly beneficial when it comes to MT for language combinations with sparse parallel corpora. We aim to translate one particular language pair (Tamil→Hindi) with a single encoder-decoder with shared attention mechanism NMT engine while using English→Tamil and English→Hindi as

$^5$In the remaining of this paper we refer to multilingual NMT engines that are trained according to the Zero Shot Approach as *ZST engines.*

$^6$We refer to the language of a specific sentence as its *language of origin* to differentiate between source and target languages.
well as a small set of Tamil→Hindi data.

2.2 Indian languages

Research, conducted on MT for Indian languages, mainly focuses on to- and from-
English translation (Sindhu and Sagar, 2016; Antony, 2013). In the survey of Antony
(2013) of MT systems for Indian languages there is only one Tamil-Hindi system. Even
exploiting data-driven MT paradigms (such as PBSMT or NMT) that ease the creation
and exploitation of MT systems even by non-linguistically informed users, the lack of
parallel data is what restricts high-quality MT systems to be built. Ramasamy et al.
(2012) present an English-Tamil PBSMT engine as well as a corpus of circa 200000
parallel sentences. Post et al. (2012) present parallel corpora for six Indian languages
and English. Bojar et al. (2014) discuss the HindEnCorp dataset which constitutes of
approximately 300000 parallel sentences. Another source for data are platforms like
Opus and EMILLE. These resources, however are not sufficient (both quantity-wise
as well as quality-wise) to build an efficient, domain oriented one-to-one MT engine
between two Indian languages.

The aforementioned issues impose a translation gap between Indian languages. We
exploiting ZST methodology in order to reduce this gap. We use various available
parallel corpora, which we cleansed and organised, to training our ZST engines.

3 Zero Shot Translation Engines

3.1 Pipeline

The KantanMT platform has two main pipelines: one to train an MT engine and a
second one to translate text with a selected MT engine. Figure 1 illustrates these
pipelines.

![Diagram of training and translation pipelines](image)

Figure 1: Abstract representation of the training and translation pipelines. Blue boxes
indicate processing steps that are common for both pipelines. The input of the training
pipeline is source and target data; the input of the translation pipeline is text to
translate.

While their core processing mechanisms are different, as shown in Figure 1 they
both use the same tokenisation step, as well as word segmentation. In practice, in the
latter step a dictionary is used that is created in the Build dictionaries step in the
training pipeline; this dictionary is then stored and reused during translation again in
the word segmentation step.⁸

---

⁷We refer the interested reader to the (Antony, 2013) for more information on the system.
⁸We present more details about word segmentation and dictionaries in Section 4.
In order to support both training and translating with a ZST engine, it is necessary to adapt these common steps (i.e., the tokenisation and the word segmentation) such that they meet the following requirements:

1. Training and test data are augmented with ZST tokens as defined in Section 2.
2. Different languages require different tokenisation rules which needs to be accommodated in the tokenisation step. That is, the training and test data sets would contain sentences in different languages (see Section 2). The tokeniser would required to know their language of origin and tokenise them according to language-specific rules.
3. Any ZST token is not affected by any consecutive preprocessing step.
4. During both training and translation the output of the neural network does not contain any ZST token.

To meet the first requirement the user needs to introduce ZST tokens for the source and the target data. The target data needs to be augmented with one ZST token, which indicates the language of origin of the data. E.g., if the target data is in English, each sentence needs the prefix \(zst\_en\) (if a locale is specified, e.g., British English, the prefix is \(zst\_en\_gb\)). The source data, however, needs two ZST tokens: one to indicate the language of origin and another to indicate the target language. These have the same form as mentioned above with the first ZST token referring to the language of origin and the second one indicates the target language. Example 3.1 illustrates the source and target data, augmented with ZST tokens.

Example 3.1

Source (English, original): It helps for detachment of umbilical cord.
Source (English, with ZST tokens): \(zst\_en\) \(zst\_hi\) It helps for detachment of umbilical cord.
Target (Hindi, original): आपको ईमेल एलर्ट के लिए सबस्क्रिब किया गया है।
Target (Hindi, with a ZST token): \(zst\_hi\) आपको ईमेल एलर्ट के लिए सबस्क्रिब किया गया है।

In order to meet requirements 2, 3 and 4, we modified the Tokenisation step as well as the Word segmentation step in our pipelines. The Tokenisation step is adapted to read the first from the two ZST tokens from each sentence of the source data and the only one ZST token from each sentence of the target data and extract the language and locale codes. Then it removes these ZST tokens. Next, each sentence will be tokenised according to tokenisation rules specific for the language and locale codes extracted from the ZST token.

The Word segmentation step, which is prior to the Build NMT step (in the training pipeline) or to the Translation step (in the translation pipeline) will split each word into subword units (Sennrich et al., 2016). During this step the ZST tokens may become segmented which would negatively impact the training of the network. We augment the Word segmentation with an extra step to recover any segmented ZST token.

Example 3.2 shows the form of the source and the target data prior to training. The @@ symbols are used as a delimiter for the word segmentation.

Example 3.2

Source (English, original): You have been subscribed to email alerts.
Source (English, tokenized, word-segmented): \(zst\_hi\) You have been subscribed to email alerts.
3.2 Engines

With the adapted pipelines we can now easily build ZST engines and use them to translate between language pairs for which parallel data was not provided. In particular, given parallel data set between languages \(L_1\) and \(L_2\) as well as between \(L_2\) and \(L_3\) we can build a ZST engine that translates a text in \(L_1\) into \(L_3\).

**Example 3.3** Consider we have available parallel data between English (EN) and Tamil (TA) and between English and Hindi (HI). We use TA and HI data both as source and as target (aligned correctly with their EN counterpart), and the same for the EN data (aligned correctly with the TA and HI) and train a ZST engine:

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Tamil</td>
</tr>
<tr>
<td>Tamil</td>
<td>English</td>
</tr>
<tr>
<td>English</td>
<td>Hindi</td>
</tr>
<tr>
<td>Hindi</td>
<td>English</td>
</tr>
</tbody>
</table>

This engine would allow us to translate from TA to HI, but also the other way round – from HI to TA. Moreover, it would translate from EN to HI or TA (and vice-versa) as well as from EN to EN.

Example 3.3 shows how we use the available parallel data both as source and as target, aligned correctly, in order to train a basic ZST engine. In general, given data for \(N\) languages all aligned with 1 other language (in Example 3.3 that is English) we can build a ZST engine to translate between all of the \(N \times (2 + N)\) source and target options, including (as in Example 3.3) translating between the same language.

The reason that a ZST engine requires the data to be used both as source and as target is that the neural network will learn to map unseen language pairs through their
Table 1: Summary of the data used to build ZST and one-to-one engines.

<table>
<thead>
<tr>
<th>Engine Name</th>
<th>Languages:</th>
<th>Number of Sentences</th>
<th>Source Words:</th>
<th>Target Words:</th>
<th>Used to translate:</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZST1</td>
<td>EN-FR, EN-IT</td>
<td>798 999</td>
<td>3 844 982</td>
<td>3 475 693</td>
<td>FR→IT</td>
<td>Legal</td>
</tr>
<tr>
<td>Pivot1</td>
<td>EN→FR</td>
<td>198 999</td>
<td>3 399 530</td>
<td>3 502 284</td>
<td>IT→EN</td>
<td>Legal</td>
</tr>
<tr>
<td>Pivot2</td>
<td>IT→EN</td>
<td>198 999</td>
<td>3 844 982</td>
<td>3 475 693</td>
<td>FR→IT</td>
<td>Legal</td>
</tr>
<tr>
<td>ZST2</td>
<td>EN→TA, EN→HI</td>
<td>1 009 892</td>
<td>15 284 069</td>
<td>15 284 069</td>
<td>TA→HI</td>
<td>General</td>
</tr>
<tr>
<td>Pivot3</td>
<td>TA→EN</td>
<td>168 871</td>
<td>2 759 734</td>
<td>3 960 123</td>
<td>TA→EN</td>
<td>General</td>
</tr>
<tr>
<td>Pivot4</td>
<td>EN→HI</td>
<td>268 317</td>
<td>3 338 686</td>
<td>3 620 144</td>
<td>EN→TA</td>
<td>General</td>
</tr>
<tr>
<td>one-to-one1</td>
<td>TA→HI</td>
<td>41 739</td>
<td>365 571</td>
<td>546 584</td>
<td>TA→HI</td>
<td>Technical</td>
</tr>
</tbody>
</table>

In the scope of this work we build ZST engines with English, French, Italian data, as well as with English, Tamil and Hindi data. First, we build a proof-of-concept ZST engine on English-French, English-Italian data; we use this engine to translate between French and Italian. To test the performance of this engine we also build two One-to-one engines: one from French to English and another from English to Italian. We refer the latter engines as Pivot engines and use them in a sequence to derive an Italian translation, starting from a French text.

Next, we focus on the Indian languages and build two ZST engines - one on English-Tamil and English-Hindi data and a second one on the same English-Tamil and English-Hindi data as well as Tamil-Hindi data. Then we build three one-to-one engines: one Tamil-Hindi, one Tamil-English and a third one English-Hindi all using the same data as for the ZST engines.

Table 3.2 enumerates the available data and the engines we trained.

In Section 4 we present and discuss our findings from comparing the translation quality of these engines.

4 Experiments

We perform our analysis on the MT engines – ZST or one-to-one – enumerated in Table 3.2.

NMT setup. Our training and translation pipelines are based on the OpenNMT toolkit\(^{10}\) (Klein et al., 2017) version 0.7. As learning optimizer we use ADAM (Kingma and Ba, 2014) with learning rate 0.0005. We train our networks for at least 5\(^{11}\) epochs on NVIDIA G520 GPUs with 4GB RAM (each model is trained on a single GPU). The maximum batch size is 50. The maximum input length used for training is 150.

Dictionaries. Each NMT engine is trained on two dictionaries – one for the source and one for the target data. For ZST engines, we use the concatenated source or target training data to build a source or target dictionary. The dictionaries are composed of word segments in order to increase the vocabulary capabilities of the network and avoid out-of-vocabulary (OOV) problems. We use byte pair encoding (BPE) Sennrich et al. (2016) of 40 000 operations to build the word segments.\(^{12}\) We prepare the dictionaries from normal-cased (i.e., lower- and upper-cased) tokenised data.

\(^{9}\)For more details we refer the interested reader to (Johnson et al., 2016).

\(^{10}\)http://opennmt.net/

\(^{11}\)We present and analyse results of engines with the same number of epochs as to make the comparison fair.

\(^{12}\)For data in Chinese, Japanese, Korean or Thai, our pipelines use dictionaries based on character-by-character segmentation (Chung et al., 2016). That is, each word segment in the dictionary is a single character. BPE is used for all other languages, including Tamil and Hindi.
Table 2: Evaluation of our Indian engines. * - the higher the better; ** - the lower the better.

<table>
<thead>
<tr>
<th>Engine</th>
<th>BLEU</th>
<th>F-Measure</th>
<th>Perplexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZST$_2$</td>
<td>0.21</td>
<td>3.26</td>
<td>17.12</td>
</tr>
<tr>
<td>ZST$_3$</td>
<td>9.78</td>
<td>26.40</td>
<td>21.91</td>
</tr>
<tr>
<td>one-to-one$_1$</td>
<td>8.20</td>
<td>22.16</td>
<td>78.96</td>
</tr>
<tr>
<td>Pivot$_3$ + Pivot$_4$</td>
<td>0.16</td>
<td>16.94</td>
<td>24.85</td>
</tr>
</tbody>
</table>

**Result analysis.** We began our experiments using a ZST engine consisting of Legal domain data acquired from the European Commission – DGT, which is freely available for use. We decided on a POC engine consisting of English↔French and English↔Italian parallel data sets. We also constructed two one-to-one engines for the same language pairs as the ZST (i.e., Pivot$_1$ and Pivot$_2$, see Table 3.2). We started by running 50 sentences of legal domain content that the engines had not seen during training. The translation test set content was in French and needed to be translated into Italian. First, the ZST engine translated the content from French to Italian. Next the same French legal content was translated through the French↔English engine (Pivot$_1$); then we used the output from this engine as input for the English↔Italian engine (Pivot$_2$).

We then evaluated both Italian outputs produced by the ZST$_1$ and Pivot$_2$ engines running an A/B testing with KantanLQR, KantanMT’s quality evaluation platform. A native Italian speaker with French fluency ranked the translations. The result of the A/B test was conclusively in favour of ZST, with our reviewer choosing 58 percent of the test segments from this engine as better quality than that of the pivot engines. With this result from our POC engine with high resource languages we began experimenting with low resource languages, in particular English↔Tamil and English↔Hindi.

The initial translation tests for our ZST Indian language engine were not as promising as we had hoped from the results of our POC engines. The output was not a complete translation to Hindi but a combination of all 3 input languages of English, Tamil and Hindi. From this result we concluded that we would need more parallel data in both language pairs and possibly aligned data for Tamil↔Hindi to help bridge the sparse data gap. We augmented our test data (the statistics of our data to built the ZST$_3$ engine, shown in Table 3.2).

We use BLEU (Papineni et al., 2002) and F-Measure (Melamed et al., 2003) to assess the quality of the Indian engines. We also report the perplexity of the engine scored after training is finished. To test whether indeed ZST can improve on one-to-one or pivot engines, we use the same test data set. It contains 500 sentences that are from the same domain of the one-to-one engine (one-to-one$_1$ in Table 3.2). Our results are summarised in Table 4.

While the enlisted scores for the given test set are in general very low, we observe that the best scores are achieved by the ZST$_3$ engine – the ZST engine which combines parallel data in different languages and a small set of Tamil↔Hindi data – the BLEU and F-Measure scores for the ZST$_3$ engine are the highest.

Furthermore, these results confirm that a ZST engine with parallel data for the languages of interest can significantly boost the translation capabilities (compare the scores of ZST$_2$ and ZST$_3$).

We ought to note that while these engines may not produce high-quality Tamil↔Hindi translations (according to the evaluation metrics reported in Table 4) they show that ZST has a potential and deserves further investigation. Our direct ef-
forts are in bringing a Tamil→Hindi engine together with other Indian languages to industry standards.

5 Conclusions and Future Work

In this paper, we present our first Zero Shot Translation engines for languages with sparse training data. We observed that while ZST produces good quality output for high resource languages (with good training data), it is not performing as good for the Tamil→Hindi language pair that we used as our main use case. However, our ZST engine that combines multiple-source data and Tamil→Hindi performs better than the rest of the Indian engines.

Our results showed that further experiments on zero shot translation are needed. First, we will focus on data analysis in order to understand which data combinations are useful for ZST and which are not. Next, we intend to test ZST for other language combinations in order to evaluate which language families or specific languages could benefit the most from such a translation approach.

In addition, with this work we adapted the training and translation pipelines of a commercial MT provider to support ZST engines. In the future we aim to further improve these pipelines and provide more and better ZST services to the users.

References


Feature-rich NMT and SMT post-edited corpora for productivity and evaluation tasks with a subset of MQM-annotated data

Kim Harris
Lucia Specia and Aljoscha Burchardt

Abstract

This presentation will discuss the creation and practical use of a large data set created through an unprecedented large-scale collaboration between MT R&D and translation experts. It contains post-edited and annotated industry data for four morphologically rich language pairs (EN-DE, EN-CS, DE-EN, EN-LV). A subset of “almost perfect” sentences also contains MQM error annotations for further detailed analysis and profiling for recurring error patterns. The post edits were performed by professional translators and the data is freely available for further use. The data used for post-editing comprised 20,000 to 45,000 sentences of industry data (IT, life sciences) depending on the language pair. The post-editing of all four language pairs was performed using PET (Aziz, W. et al.). Several crucial and novel data points were taken during the post-editing: time logging, keystroke logging quality evaluation of the post-editing effort by the translator upon completion of the post-editing. The recording of this information during the post-editing phase allows for specific features and novel combinations of features to be used for a variety of research- and user-oriented purposes, including establishing the actual post-editing effort by translators based on time and keystrokes and comparing these results to the perceived level of quality of the post-edited sentence, establishing correlations between certain characteristics such as sentence length and post-edit time, or post-edit time and human quality evaluation. The datasets also measure post-editing productivity and are used to detect error patterns in the MT output. This would allow users of MT to adequately assess a) the use of MT in general, b) the actual productivity gains achieved in two different systems or across languages, domains and other data subsets such as long sentences or sentences containing certain grammatical constructs or terminology. For two language pairs identical sets of source sentences comprising 30,000 sentences respectively were post-edited for NMT and SMT output, allowing for a variety of innovative comparisons to be done on the results of the two given the unique data points that were collected during post-editing. In addition, the creation of MQM-annotated subsets of these post-edits for typical industry domains provide
information about error patterns and support feature-oriented quality estimation and evaluation currently unknown to MT quality evaluation and estimation and can be used to improve the MT output.
Usability of web-based MT post-editing environments for screen reader users

Silvia Rodríguez Vázquez, Sharon O’Brien, Dónal Fitzpatrick
silvia.rodriguez@unige.ch · sharon.obrien@dcu.ie · donal.fitzpatrick@dcu.ie

Motivation

Advocacy for TEnT accessible design

But why ?
Potential social impact

- The inaccessible design of popular TEnTs prevents qualified translators with visual and motor impairments from accessing the job market

“Translation tools: help or hindrance?” (Owton & Mileto 2011)

- Translator-Computer Interaction based on:
  - Keyboard-only input
  - Text-to-speech and/or text-to-Braille output
- Other interaction modes: not practical, too time consuming
  - Use of mouse simulation commands
  - Scripting
  - Collaboration with sighted assistant/colleague

Recent research interest on user-centred factors in translation technology design and evaluation

- Usability-UX
  - Involvement of end users at design stage (Bota et al. 2013)
  - Usability of FOSS CAT (Veiga Díaz & García González 2015)
  - CAT usability modelling (Krüger 2016)
  - User Interface needs of post-editors (Moorkens & O’Brien 2017)

- Multimodal TEnT
  - Mobile post-editing app (Torres Hostench et al. 2017)
  - Interactive Translation Dictation (Zapata 2016)

- Ergonomics (Teixeira 2015)
Motivation

Request for Proposal (RFP)
“Computer-Assisted Translation (CAT) Tool for facilitating the provision of reference and translation services”

February 2017

Food and Agriculture Organization of the United Nations (FAO)

Accessibility as part of evaluation criteria

- **STILL**: Scarcity of translation technology research focusing on end-users with special needs

  - Exploratory Single Case Studies (Rodríguez Vázquez & Mileto, 2016)
    - Blind user interaction with different versions of SDL Studio
  - Questionnaire for blind and visually impaired translators (Rodríguez Vázquez & Mileto, 2016)
    - Low levels of satisfaction with current state-of-the-art desktop CAT
      - Poor interaction CAT-AT (assistive technology)
      - Lack of comprehensive technical support
      - User guides: incomplete + inaccessible
    - *Fluency Now*: Most popular MT-integrated TEnT among users, not necessarily among LSP

<table>
<thead>
<tr>
<th>n) Ergonomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>0120 – <em>Available keyboard shortcuts</em>: Some keyboard shortcuts are available.</td>
</tr>
<tr>
<td>0121 – <em>Customisable shortcuts</em>: The keyboard shortcuts are customisable.</td>
</tr>
<tr>
<td>0122 – <em>Interface Customisation</em>: Whether users are able to customise the interface. Please specify how and to what extent (e.g. size, location, arrangement, background colours of windows, fonts and letter size of menus and of the text displayed in the editor, contents and location of toolbars, etc.) this can be achieved. The software should work on dual screens; in particular, it should be possible to undock panes, if any, and move them to a second screen.</td>
</tr>
<tr>
<td>0123 – <em>Learning Curve</em>: As we deal with a number of external translators/revisers experienced with existing Cat Tools, we expect a low learning curve for rapid adoption of a new CAT tool.</td>
</tr>
<tr>
<td>0124 – <em>Accessibility</em>: accessibility features are available for people with disabilities.</td>
</tr>
<tr>
<td>0125 - OCR features and speech recognition: OCR features exist and some speech recognition software is compatible with the software.</td>
</tr>
</tbody>
</table>

No research work found on accessibility of translation tools and MT/post-editing
Study

**Goal:** Explore the potential of web-based MT-integrated TEnT as a more suitable solution for blind translators

**Selection Criteria**
- Integration of MT
- Free access
- All main components, including post-editing environment, are web-based
- The basic accessibility requirements to enable exploration of the following pages are met: sign up, log in, project creation, post-editing environment

**Tools chosen for study:**
- Memsource Cloud
- Matecat

**Method**

- **Classic usability study approach**
  - Task + questions about user experience
  - Summative evaluation
  - Remote, asynchronous usability evaluation (Petrie et al. 2006, Murphy et al. 2016)

- **Snowball sampling**
  - The Round Table mailing list (approx. 150 subscribers)
    - [http://lists.screenreview.org/listinfo.cgi/theroundtable-screenreview.org](http://lists.screenreview.org/listinfo.cgi/theroundtable-screenreview.org)

**INSTRUCTIONS**

1. Conduct a simple **post-editing exercise** with each tool
3. Fill in a **post-task questionnaire** after each exercise
   - Based on Computer System Usability Questionnaire (CSUQ) (Lewis 1995)
Participants - Profile

16 blind translators agreed to participate (consent form)

11 tested at least 1 tool

9 tested both tools

- **Age**: 18-24 (N=2), 25-34 (N=6), 35-44 (N=3)
- **Nationality**: Austria (N=3), Germany (N=2), Italy (N=2), Canada (N=1), Egypt (N=1), Poland (N=1), UK (N=1)
- **Education**: Translation background; university degree (BA/MA) (completed N=9; ongoing N=2)
- **Current occupation**: translator (N=6), public administration (N=1), web analyst (N=1), transcription service manager (N=1)
- **Computer skills** (*self-assessment, 5-point scale*): Adequate (N=1), Good (N=5), Excellent (N=5)

Participants – Use of user agents

<table>
<thead>
<tr>
<th>Operating System</th>
<th>Windows</th>
<th>Windows</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Browser</strong></td>
<td>Google Chrome (N=3) Mozilla Firefox (N=8) IE (N=1)</td>
<td>Google Chrome</td>
</tr>
<tr>
<td><em>(2 participants used 2 different browsers)</em></td>
<td><em>(3 participants used 2 different screen readers)</em></td>
<td></td>
</tr>
<tr>
<td><strong>Assistive technology</strong></td>
<td>Screen reader only (N=2), screen reader &amp; Braille refreshable display (N=8), per tool</td>
<td>Screen reader: 8 participants used JAWS, 4 participants used NVDA</td>
</tr>
</tbody>
</table>
## CSUQ – Measurement of usability

<table>
<thead>
<tr>
<th>ITEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Overall, I am satisfied with how easy it is to use this system</td>
</tr>
<tr>
<td>2 It was simple to use this system</td>
</tr>
<tr>
<td>3 I can effectively complete my work using this system</td>
</tr>
<tr>
<td>4 I am able to complete my work quickly using this system</td>
</tr>
<tr>
<td>5 I am able to efficiently complete my work using this system</td>
</tr>
<tr>
<td>6 I feel comfortable using this system</td>
</tr>
<tr>
<td>7 It was easy to learn to use this system</td>
</tr>
<tr>
<td>8 I believe I can become productive quickly using this system</td>
</tr>
<tr>
<td>9 I felt confident using the system</td>
</tr>
<tr>
<td>10 The system gives error messages that clearly tell me how to fix problems</td>
</tr>
<tr>
<td>11 Whenever I make a mistake using the system, I recover easily and quickly</td>
</tr>
<tr>
<td>12 The information (such as online help, messages, and other documentation) provided with this system is clear</td>
</tr>
<tr>
<td>13 It is easy to find the information I needed</td>
</tr>
<tr>
<td>14 The information provided with the system is easy to understand</td>
</tr>
<tr>
<td>15 The information is effective in helping me complete the tasks and scenarios</td>
</tr>
<tr>
<td>16 The organization of information on the system screens is clear</td>
</tr>
<tr>
<td>17 I found the various functions in this system were well integrated</td>
</tr>
<tr>
<td>18 This system has all the functions and capabilities I expect it to have</td>
</tr>
<tr>
<td>19 Overall, I am satisfied with this system</td>
</tr>
</tbody>
</table>

### CSUQ Scores (I)

#### Overall scores

1. Strongly disagree  
2 3 4 5 6 7. Strongly agree

**MateCat (CSUQ scores)**

- **System Usefulness**
  - 7
  - 6
  - 5
  - 4
  - 3
  - 2
  - 1

- **Information Quality**
  - 7
  - 6
  - 5
  - 4
  - 3
  - 2
  - 1

- **Overall Satisfaction**
  - 7
  - 6
  - 5
  - 4
  - 3
  - 2
  - 1

- **Fit for Purpose**

**Memsource (CSUQ scores)**

- **System Usefulness**
  - 7
  - 6
  - 5
  - 4
  - 3
  - 2
  - 1

- **Information Quality**
  - 7
  - 6
  - 5
  - 4
  - 3
  - 2
  - 1

- **Overall Satisfaction**
  - 7
  - 6
  - 5
  - 4
  - 3
  - 2
  - 1

- **Fit for Purpose**
### CSUQ Scores (II)

#### Overall scores

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>System usefulness</td>
<td>1.64</td>
<td>0.635</td>
<td>3.23</td>
<td>1.020</td>
<td>3.25</td>
<td>0.707</td>
<td>2.37</td>
<td>1.134</td>
</tr>
<tr>
<td>Information quality</td>
<td>4.00</td>
<td>0.316</td>
<td>4.21</td>
<td>0.476</td>
<td>5.19</td>
<td>0.441</td>
<td>4.20</td>
<td>0.514</td>
</tr>
</tbody>
</table>

**p-value (t-test)**

- System usefulness: <0.001
- Information quality: 0.051
- Fit for purpose: 0.081
- Overall: <0.001

### CSUQ Scores (III)

#### If we look closer, per item (highlights)

<table>
<thead>
<tr>
<th>System usefulness</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>7. It was easy to learn to use this system</td>
<td>3.11</td>
<td>2.315</td>
<td>1.89</td>
<td>1.536</td>
</tr>
<tr>
<td>8. I believe I can become productive quickly using this system</td>
<td>4.40</td>
<td>2.118</td>
<td>3.60</td>
<td>2.458</td>
</tr>
</tbody>
</table>

**p-value (t-test)**

- 7. It was easy to learn to use this system: 0.225
- 8. I believe I can become productive quickly using this system: 0.086

#### Confidence in having successfully completed the task

- 1 (80%, N=8)
- 3 (10%, N=1)
- 5 (10%, N=1)
- 1 (20%, N=2)
- 4 (10%, N=1)
- 5 (10%, N=1)
- 6 (20%, N=2)
- 7 (40%, N=4)
Frustration Experiences

Summary

• Most problematic steps during the translation exercise
  (“What were you trying to do?”)

<table>
<thead>
<tr>
<th>Step</th>
<th>#</th>
<th>%</th>
<th>(\bar{x}, in min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create a new project</td>
<td>9</td>
<td>31.03%</td>
<td>20’</td>
</tr>
<tr>
<td>Edit target segment (general)</td>
<td>5</td>
<td>17.24%</td>
<td>37’</td>
</tr>
<tr>
<td>Set up the project</td>
<td>5</td>
<td>17.24%</td>
<td>8’</td>
</tr>
<tr>
<td>Edit MT suggestions/post-edit</td>
<td>2</td>
<td>6.90%</td>
<td>15’</td>
</tr>
<tr>
<td>Upload source file</td>
<td>2</td>
<td>6.90%</td>
<td>30’</td>
</tr>
<tr>
<td>Navigate through main menu</td>
<td>2</td>
<td>6.90%</td>
<td>3’</td>
</tr>
<tr>
<td>Sign up</td>
<td>2</td>
<td>6.90%</td>
<td>6’</td>
</tr>
<tr>
<td>Read translated segments</td>
<td>1</td>
<td>3.45%</td>
<td>2’</td>
</tr>
<tr>
<td>Export the target file</td>
<td>1</td>
<td>3.45%</td>
<td>5’</td>
</tr>
</tbody>
</table>

Technical problem encountered
  (“What happened?”)

<table>
<thead>
<tr>
<th>Problem</th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non labelled buttons/fields</td>
<td>10</td>
<td>29.41%</td>
</tr>
<tr>
<td>Button not working</td>
<td>6</td>
<td>17.65%</td>
</tr>
<tr>
<td>Not possible to read own translated text</td>
<td>5</td>
<td>14.71%</td>
</tr>
<tr>
<td>Not possible to post-edit</td>
<td>5</td>
<td>14.71%</td>
</tr>
<tr>
<td>Lack of content structure</td>
<td>3</td>
<td>8.82%</td>
</tr>
<tr>
<td>Lack of information &amp; feedback</td>
<td>3</td>
<td>8.82%</td>
</tr>
<tr>
<td>Cursor got stuck in edit field</td>
<td>1</td>
<td>2.94%</td>
</tr>
<tr>
<td>Not possible to export</td>
<td>1</td>
<td>2.94%</td>
</tr>
</tbody>
</table>

Solution or coping strategy
  (“How did you solve the problem?”)

<table>
<thead>
<tr>
<th>Solution</th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>I was unable to solve it</td>
<td>13</td>
<td>44.83%</td>
</tr>
<tr>
<td>I figured out a way to fix it myself without help</td>
<td>8</td>
<td>27.59%</td>
</tr>
<tr>
<td>I ignored the problem or found an alternative solution</td>
<td>6</td>
<td>20.69%</td>
</tr>
<tr>
<td>I knew how to solve it because it has happened before</td>
<td>1</td>
<td>3.45%</td>
</tr>
<tr>
<td>I asked someone for help.</td>
<td>1</td>
<td>3.45%</td>
</tr>
</tbody>
</table>
Frustration Experiences

**MT/Post-editing**

<table>
<thead>
<tr>
<th></th>
<th>#</th>
<th>%</th>
<th>Time lost ((\bar{x}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edit target segment (general)</td>
<td>5</td>
<td>17.24%</td>
<td>37’</td>
</tr>
<tr>
<td>Edit MT suggestions/post-edit</td>
<td>2</td>
<td>6.90%</td>
<td>15’</td>
</tr>
<tr>
<td>Read translated segments</td>
<td>1</td>
<td>3.45%</td>
<td>2’</td>
</tr>
</tbody>
</table>

- Considered as important (N=2) or **very important (N=6) steps to complete the translation task**
- Related-problems encountered considered as frustrating (N=2) or **very frustrating (N=6)**

**P01:** “I could not edit the MT suggestions effectively. I could view the suggestions, but the only way to edit them that I could find was to copy them into the edit field; however, when I did that, the edit field still appeared to be empty and I couldn’t edit the text I had just copied and pasted. When I decided to simply write the translation myself, I couldn’t read what I had just typed in either; my braille display and screen reader showed an empty edit field.”

**P11:** “I entered Web Editor. Then, not without difficulties, I found my way to the target segment column. And then I started to write in it. The problem is, however, that NVDA would report what I have just written, but I went back with my edit field cursor, it only read “blank”[…] As long as I am not in full control of target-text editing, I cannot complete even a single segment of my translation.”

**P05:** “It wasn’t marked up as being an edit field, the target segment was just a line of text. Therefore I couldn’t find how to edit this.”

---

**Frustration Experiences**

**Summary**

- Most **problematic steps** during the translation exercise

<table>
<thead>
<tr>
<th></th>
<th>#</th>
<th>%</th>
<th>(\bar{x}), in min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edit MT suggestions/post-edit</td>
<td>6</td>
<td>22.22%</td>
<td>9’</td>
</tr>
<tr>
<td>Sign up and login</td>
<td>6</td>
<td>22.22%</td>
<td>10’</td>
</tr>
<tr>
<td>Upload source file</td>
<td>3</td>
<td>11.11%</td>
<td>8’</td>
</tr>
<tr>
<td>Revise translation</td>
<td>3</td>
<td>11.11%</td>
<td>3’</td>
</tr>
<tr>
<td>Edit target segment (general)</td>
<td>2</td>
<td>7.41%</td>
<td>13’</td>
</tr>
<tr>
<td>Export the target file</td>
<td>2</td>
<td>7.41%</td>
<td>23’</td>
</tr>
<tr>
<td>Set up the project</td>
<td>2</td>
<td>7.41%</td>
<td>3’</td>
</tr>
<tr>
<td>Navigate through main menu</td>
<td>1</td>
<td>3.70%</td>
<td>10’</td>
</tr>
<tr>
<td>Copy source to target</td>
<td>1</td>
<td>3.70%</td>
<td>15’</td>
</tr>
<tr>
<td>Check MT/TM metadata</td>
<td>1</td>
<td>3.70%</td>
<td>5’</td>
</tr>
</tbody>
</table>

(“What were you trying to do?”)
Frustration Experiences
Summary

Technical problem encountered
(“What happened?”)

<table>
<thead>
<tr>
<th></th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screen reader failure</td>
<td>6</td>
<td>21.43%</td>
</tr>
<tr>
<td>Button not working</td>
<td>5</td>
<td>17.86%</td>
</tr>
<tr>
<td>Not possible to post-edit</td>
<td>3</td>
<td>10.71%</td>
</tr>
<tr>
<td>Not possible to sign up</td>
<td>3</td>
<td>10.71%</td>
</tr>
<tr>
<td>Lack of information &amp; feedback</td>
<td>3</td>
<td>10.71%</td>
</tr>
<tr>
<td>Lack of structure</td>
<td>2</td>
<td>7.14%</td>
</tr>
<tr>
<td>Not possible to locate access to editor</td>
<td>2</td>
<td>7.14%</td>
</tr>
<tr>
<td>Not possible to export</td>
<td>1</td>
<td>3.57%</td>
</tr>
<tr>
<td>Not possible to read long segments</td>
<td>1</td>
<td>3.57%</td>
</tr>
<tr>
<td>Manual search/find of segments</td>
<td>1</td>
<td>3.57%</td>
</tr>
<tr>
<td>Difficulty editing text</td>
<td>1</td>
<td>3.57%</td>
</tr>
</tbody>
</table>

Solution or coping strategy
(“How did you solve the problem?”)

<table>
<thead>
<tr>
<th></th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>I figured out a way to fix it myself without help</td>
<td>10</td>
<td>37.04%</td>
</tr>
<tr>
<td>I was unable to solve it</td>
<td>7</td>
<td>25.93%</td>
</tr>
<tr>
<td>I ignored the problem or found an alternative solution</td>
<td>6</td>
<td>22.22%</td>
</tr>
<tr>
<td>I asked someone for help.</td>
<td>2</td>
<td>7.41%</td>
</tr>
<tr>
<td>I tried again</td>
<td>1</td>
<td>3.70%</td>
</tr>
<tr>
<td>I restarted the program</td>
<td>1</td>
<td>3.70%</td>
</tr>
</tbody>
</table>

Frustration Experiences
MT/Post-editing

<table>
<thead>
<tr>
<th></th>
<th>#</th>
<th>%</th>
<th>Time lost (x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edit MT suggestions/post-edit</td>
<td>6</td>
<td>22.22%</td>
<td>9’</td>
</tr>
<tr>
<td>Revise translation</td>
<td>3</td>
<td>11.11%</td>
<td>3’</td>
</tr>
<tr>
<td>Edit target segment (general)</td>
<td>2</td>
<td>7.41%</td>
<td>13’</td>
</tr>
<tr>
<td>Copy source to target</td>
<td>1</td>
<td>3.70%</td>
<td>15’</td>
</tr>
<tr>
<td>Check MT/TM metadata</td>
<td>1</td>
<td>3.70%</td>
<td>5’</td>
</tr>
</tbody>
</table>

- Considered as important (N=7) or very important (N=6) steps to complete the translation task
- Variability observed in levels of frustration related to problems encountered

P01: “Starting at the 4th segment, Jaws started behaving oddly while I was trying to read and edit the translation - speech output did not only read everything out loud twice, it also randomly read parts of the following lines.”

“I discovered that this only happened when the tags in the target segment hadn’t been put in place yet; once I had selected ‘Guess Tags’ this was no longer an issue. [...] Checking the translation via Braille display worked well, though.”

P07: “While I was revising certain (longer) segments, I was no longer able to read the end of the segment, neither using speech output nor with my Braille display.”

P15: “MateCat had automatically inserted the MT suggestion. But below the translation it indicated a symbol mismatch. When reading the translation, I noticed that there were strange symbols in the middle of the sentence. When I tried to move the cursor to these symbols to delete them, MateCat crashed, and I had to restart it. This happened several times.”
Overall research indicators

- None of the tools tested could be professionally used by blind translators in their current form
  - **BUT:** MateCat could be fully accessible only with minor changes

- **Blind translators are more resourceful than we thought!**
  - Advanced IT competence (use of multiple AT and browsers), so they can easily adapt
  - But want to be treated as their sighted peers

- **We need to look for designed-for-all solutions**
  - Tools for blind translators only; e.g. EasyTrans (Al-Bassam et al. 2016): not the preferred approach by real end users!

Future Work

- **In-depth analysis of qualitative data gathered**
  - Levels of frustration; correlation with time lost
  - Technical difficulties logged could provide insights for TEnT developers about what aspects to test (“accessibility check list”)
    - Send report to TEnT providers

- **Observation study with selected participants**
  - Interaction with more advanced TEnT features

- **Parallel usability study with sighted translators**
  - Comparison of CSUQ scores
  - Comparison of user preferences regarding information quality and user interface
Thank you
References (I)


References (II)


Live presentations to a multilingual audience: personal universal translator

Chris Wendt

Abstract

The fact that just about everybody carries an internet-connected smartphone enables us to break the language barrier for in-person meetings, presentations, talks and lectures. Using the smartphone in a smart way to connect the audience with the speaker allows all audience members to follow along and participate, regardless of language. Presentations often include specialized vocabulary, like people’s names, names of products, company specific acronyms and abbreviations. This is not a challenge for text translation, but in speech translation unknown words are mapped to the phonetically closest known word, which can have a catastrophic effect in translation. Customization of the speech recognition system helps here. We are showing a new and very convenient method to customize the speech recognition system, thus providing a useful automatic interpretation and translation of the speech. It uses the slide deck’s content, slides and notes, to customize the SR system at the start of the session, allowing the speaker to use the terms used in the deck, adding the specific terms to the SR system’s standard vocabulary. The system displays the transcript in the speaker’s language, or a language of choice on the presentation screen, and each audience member may follow in their own language, on their own device. The system extends to the audience. The audience member can ask a question in any of the supported language, speaking or typing into the mobile device. All audience members can read the question in their own language, as well as the answer of the presenter. This makes microphone-runners practically unnecessary. A benefit for audience members with hearing loss is the availability of full transcripts of what everybody is saying. We’ll show a few examples situations where this has proven useful. This session provides a view into conducting truly multilingual presentations with full audience participation regardless of language and hearing abilities.
Toward a full-scale neural machine translation in production: the Booking.com use case

Pavel Levin pavel.levin@booking.com
Nishikant Dhanuka nishikant.dhanuka@booking.com
Talaat Khalil talaat.khalil@booking.com
Fedor Kovalev fedor.kovalev@booking.com
Maxim Khalilov maxim.khalilov@booking.com

Abstract
While some remarkable progress has been made in neural machine translation (NMT) research, there have not been many reports on its development and evaluation in practice. This paper tries to fill this gap by presenting some of our findings from building an in-house travel domain NMT system in a large scale E-commerce setting. The three major topics that we cover are optimization and training (including different optimization strategies and corpus sizes), handling real-world content and evaluating results.

1 Introduction
Booking.com is one of the largest online companies in the world operating in 43 different languages, connecting millions of daily visitors to 1.4 million bookable accommodations while offering both parties multilingual support and information every step of the way. Given the company’s fast growth and a rising need for more high quality translated content, machine translation (MT) is becoming an increasingly attractive option to automate this difficult task.

Our experiments [9] consistently show the superiority of neural machine translation (NMT) systems over the more traditional statistical ones, even when we benchmark them against the well-established and tested general purpose systems. Therefore our recent focus has been on tailoring and improving our own in-house NMT systems to make them practical and effective for us. This work highlights some of the main learnings on our journey and should be of interest to anyone looking to deploy a custom NMT system.

In particular we focus on the following three major topics:

- Optimization and training

At Booking.com we have collected tens of millions of travel domain specific human-translated parallel sentences, which in theory allows us to train very flexible models with hundreds of millions of parameters. However learning such system can be computationally expensive which often translates to unacceptably long product development iteration cycles. To address this we first analyze how convergence is affected by different optimization techniques (Section 2.2), including in a multi-GPU environment. Second, we look at how the quality of a trained system improves as a function of the training corpus size (Section 2.3).
• Handling real-world content
  
  Real world text comes with many challenges which have to be addressed. Section 3 presents some practical considerations for dealing with named entities and rare words.

• Quality evaluations
  
  When building an MT system with customer-facing output, setting up a good quality evaluation loop can be one of the most important aspects. In this part we show how in addition to the BLEU metric [12], the de facto standard for automatic MT scoring, we employ human evaluation of translation adequacy and fluency. We take a close look at how the two approaches correlate. Further, we share our experience developing our business sensitivity framework, which helps us proofread the final translation identifying particularly pernicious errors.

2 Optimization and training

2.1 Model architecture
  
  The core of our translation pipeline is based on OpenNMT [7], which is a Lua written framework for training encoder-decoder neural architectures. Usually, both the encoder and the decoder recurrent neural networks (RNNs), in our case typically long short-term memory (LSTM) units [5], each with 4 layers. We always use (global) attention layer with input feeding to help the model learn faster by keeping a “memory” of past alignment decisions [10]. For European languages we use “case features” (see Section 3.1) as additional input variables from the “cases” embedding space [14]. The main word embeddings are concatenated with the case embeddings to form the inputs to the encoder. At each layer of the encoder the RNNs are bi-directional [13]. Both the encoder and the decoder use residual connections between layers [4] as well as the dropout rate of 0.3 [16].

2.2 Optimization and model fitting

2.2.1 Single-GPU environment
  
  To optimize the training of our NMT system in single-GPU environments, we evaluated different algorithms primarily based on their speed of convergence and translation output quality. The dataset used was English-German property descriptions with one million parallel sentences. We conducted experiments with four well-known optimizers: stochastic gradient descent (SGD) with learning rate decay, Adam [6], Adagrad [3] and Adadelta [18]. Our SGD decay strategy is based on a combination of the perplexity score and epoch number, meaning we decay current learning rate by a multiplicative factor of 0.7 if current epoch’s validation perplexity does not decrease, and after each epoch after the 9th epoch. Our initial learning parameters for SGD, Adam, Adagrad and Adadelta are 1.0, 0.0002, 0.1, and 1.0 respectively. We ran the model for 20 epochs and used both perplexity per epoch and BLEU score after every five epochs on the validation set of 10,000 sentences to measure the performance. Our results are summarized in Table 1 and Figure 1.

As can be seen in Table 1 and Figure 1, we observed that initially Adam converged faster as expected because it applies momentum on a per parameter basis, but SGD took over as soon as decay started and outperformed Adam thereafter. The perplexity reached by SGD in the 9th epoch was already achieved by Adam in the 6th. But from the 10th epoch onward, as soon as SGD learning rate starts decaying indefinitely, Adam’s perplexity is consistently worse than that of SGD. However, there was no decrease in perplexity from 15th till 20th epoch, so SGD already converged by epoch 15. We also observed that Adagrad performed very poorly on our model. Adadelta was much better than Adagrad but still slightly behind Adam and SGD.
Table 1: Performance of different optimizers on training English-German translation model reported every 5 epochs. Each experiment was conducted in a single NVIDIA Tesla K80 GPU.

We further validated our results using BLEU scores every 5 epochs. The results were mostly consistent with what we observe by looking at perplexity. In terms of time taken per epoch, SGD was the fastest. Adam was about 10% slower in comparison.

Figure 1: Model convergence. The subplot on the left shows model convergence for three different optimizers: SGD, Adam and Adadelta. Adagrad in our setting did so poorly that it would not fit in the plot (its validation at epoch 20 was above 12). The right subplot compares the convergence of SGD on a single GPU to those of SGD run on an 8-GPU cluster using synchronous and asynchronous parameter updates.

2.2.2 Multi-GPU environment

Next we experimented with the use of multiple GPUs by using data parallelism technique which trains batches in parallel on different GPUs. On a single GPU our model takes 6h11m per epoch on average, and we usually see it converging around 15th epoch, which means training a model on only 1 million sentences takes about 4 days. 15 epochs on a corpus of size 10M could easily translate to around 40 days. In an attempt to speed up our development cycle, we ran some experiments with synchronous and asynchronous SGD (with decay) on a cluster of 2, 4, 6 and 8 GPUs. The main difference between these two approaches is that in synchronous mode all gradients are accumulated and parameter updates are synchronized, while in asynchronous each GPU calculates its own gradient and communicates with the “master copy” of parameters independently and asynchronously. This master copy of parameters is stored on a single dedicated GPU which is not used for training. To achieve a faster convergence through better parameter initialization, only one GPU works for the first 6,000 iterations in async SGD.

As can be seen in Figure 2, average time per epoch came down as we added more hardware:

---

1Reported estimates do not account for any time related to model checkpointing.
from 6h11m to 1h23m for sync and to only 56 minutes for async. Note that with 2 GPUs, async takes almost the same time as non-parallel SGD (around 6 hours) while sync is much faster at 3h31m. The reason for that is that 2-GPU async is almost equivalent to a single GPU model as async blocks one GPU completely to store the master copy of parameters and is not used for training. Because async mode skips the overhead of parameter synchronization, it was expected that it would be faster than sync, so we also looked at the quality as measured by perplexity. During the first epoch sync perplexity is much worse than that of async due to only 1 GPU working in async for first 6,000 iterations resulting in better parameter initialization (this cannot be seen in Figure 1 which has been cropped for better visibility; sync has first epoch perplexity of 9.61, compared to 5.61 for async and 3.68 for single-GPU SGD). However, for all remaining epochs their scores are very similar. Single-GPU SGD, on the other hand, performed noticeably better in the first half of the training, but gets quite similar to multi-GPU models eventually (although still marginally better). Overall we are very happy with async’s performance as it is able to reduce the training time by about 85%.

2.3 The importance of corpus size

In order see how much benefit we get from an increased corpus size, we compared models trained on 1M, 2.5M, 5M, 7.5M and 10M sentences. For fair comparison we report the learning curves as a function of number of iterations (training time) and not the epoch number. Figure 3 shows our findings.

Essentially there were no major surprises. It appears that given enough iteration the model with more distinct sentences will have a higher BLEU score. Notice how in the beginning smaller datasets are actually winning, but given enough training time the model is starting to take full advantage of more data. The largest corpus size of 10M does not have the best performance at the end of 90M iterations, however as we shall see in Section 4.3 this is in fact not true and according to human evaluations 10M gives the best results which are simply not captured by the BLEU metric.

3 Handling real-world content

3.1 Tokenization and case features

In our final models we use byte-pair encoding (BPE) tokenization procedure [15]. BPE is a compression technique which was recently adapted to find optimal tokens for sequence composition in sequence-to-sequence learning tasks. In theory the technique should find a perfect compromise between using word-level translation (and dealing with out-of-vocabulary entities)
Figure 3: Performance (measured by BLEU score) of a model trained on different corpus sizes, reported every 10M iterations.

and character-level translation (and dealing with much longer sequences of tokens). The procedure is very straightforward. We start with a set of tokens which is the list of acceptable characters and iteratively grow it, at each step adding a concatenation of two items already in the list which is the most frequent in our corpus. The number of iterations can be viewed as the algorithm’s only hyperparameter. We can either apply BPE to the source and the target sentences separately, or we can apply them to the combined corpus. Based on our experiment (see Table 2) we decided ended up with the joint version.

Table 2: Comparison of the BLEU scores of identically trained models with different BPE configurations, as well as the baseline with a vocabulary of 50,000 most common words (see [9] for more details on the baseline model). All experiments were run on 1M corpus. We found 70,000 tokens (70k) jointly trained BPE to have the highest validation BLEU score. Because we saw a strong pattern which made it clear that separately trained BPE 90k model was not going to win, we decided to not run that experiment as it is also the most expensive one.

Apart from applying BPE tokenization we also use case features preprocessing. This allows us to map the same words and word pieces spelled with different cases to the same embeddings while also passing the casing information separately. For example raw terms book, Book and BOOK would all be mapped to the same token book, but would have different accompanying case feature values. Case features get their own embeddings which get combined with token embeddings during the translation [14]. In theory this greatly increases the encoding and decoding efficiency of the system, which we also observed in practice through much better
performance over not using case features.

<table>
<thead>
<tr>
<th>Raw source</th>
<th>Offering a restaurant with WiFi, Hodor Ecolodge is located in Winterfell.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokenized source</td>
<td>offeringC aL restaurantL withL wiC fiC N, hoC dorL ecolodgeL isL locatedL inL winterC fellL</td>
</tr>
<tr>
<td>Tokenized Output</td>
<td>dieC hoC dorC ecolodgeC inC winterC fellC bietetL einL restaurantL mitL wlanU N</td>
</tr>
<tr>
<td>De-tokenized output</td>
<td>Die Hodor Ecolodge in Winterfell bietet ein Restaurant mit WLAN.</td>
</tr>
</tbody>
</table>

Table 3: A typical sentence describing an accommodation translated from English to German. Before being fed into the encoder, the sentence is first tokenized using byte-pair encodings. Notice how the words “Hodor” and “Winterfell” which never occurred in our training corpus are split into pieces which are understood by the encoder. The symbol ■ indicates no space between two neighboring word pieces. The superscripts are case features (C: true case, L: lower case, U: all capitals, N: non-alphabetic).

3.2 Handling named entities

Text in the travel domain contains a large amount of entities. There is almost always some destination involved, a property name, distances, times, etc. Although many NMT researchers report results on end-to-end neural networks [1, 2, 17], we often found RNN encoder-decoder architecture insufficient to produce acceptable results, mainly due to mishandled named entities. This section outlines our approach to processing such entities which drastically improves the translation output quality.

As an example, mistranslated distances constitute one of the most common error types when NMT is applied naively on raw text, even with very large corpus sizes (over 10M parallel sentences). Interestingly NMT often correctly converts between kilometers and miles for commonly occurring distances (e.g. 5km, 10 miles); however, the number of distance-related mistakes in our validation set is too large to be left untreated. Another common type of error is related to times and dates (12 vs 24 hour clock times, different date formats).

<table>
<thead>
<tr>
<th>Source sentence</th>
<th>Winterfell Railway Station can be reached in a 55-minute car ride.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure NMT translation</td>
<td>Den Bahnhof Winterfell erreichen Sie nach einer 55-minuten Autofahrt.</td>
</tr>
<tr>
<td>NMT with distance placeholders</td>
<td>Den Bahnhof Winterfell erreichen Sie nach einer 55-minuten Autofahrt.</td>
</tr>
</tbody>
</table>

Table 4: Translation of a sentence involving a distance using a BPE-based NMT model and an identically trained model with placeholder preprocessing. These types of errors are critical, however they are not adequately reflected in the BLEU score or decoder perplexity change.

In most such cases we used a set of manually created templates to search for entities and replace them with special placeholders. As our team does not understand most of the languages
that we build MT systems for, we get some help from our in-house language specialists (translators). The template refinement cycle goes as follows. We come up with a set of reasonable regular expressions to identify named entities of a certain type in both languages and run them on our parallel corpus. Then we take the set of sentences where the numbers of recognized entities differs between the source and the target. We then look at the breakdown of most common entities in either language which did not have corresponding parallel counterparts, and refine our regular expressions accordingly. At translation (prediction) stage, we preprocess the input to replace all named entities with corresponding placeholders, run the translation, then substitute back the named entities parsed according to the target language format. This simple approach dramatically improves the translation output quality for sentences which involve problematic named entities.

4 Quality evaluation

Unlike simple classification or regression tasks, sequence learning problems are much more difficult to evaluate. The problem comes from the fact that there can be many possible solutions and it is hard (and often impossible) to compare the model output to all valid “true values”. To assess the quality of translations automatically, a useful heuristic is the so-called BLEU score [12] which roughly measures the degree of word overlap between the model translation and a human translation. BLEU score is attractive because it is completely automatic given translated sentences and corresponding model predicted sentences. However, multiple problems have been noted in using BLEU score alone. As a purely counting-based metric, BLEU will favor translation which have more common words and n-grams with the reference translation, regardless of the sentence grammar. It would also penalize models which rephrase the sentence in a way which uses different words from the reference sentence, while preserving its meaning.

In this section we first describe how we leverage our in-house linguistic expertise to score our models in a relevant way (Section 4.1). Then we analyze how BLEU score correlates with human metrics (Section 4.3).

4.1 Human evaluation loop

Our main human evaluation is based on adequacy/fluency methodology which, as the name suggests, is based on two criteria: adequacy and fluency. Adequacy shows to what degree the meaning of the source sentence is preserved, while fluency scores how grammatically well-formed (from the native speaker’s perspective) the translated segment sounds. Each sentence is scored by two independent professional translators from English to German (native German speakers). For the experiments in Section 4.3 we chose 200 randomly selected sentences and translators with at least one year of experience professionally translating Booking.com content.

Additionally we use human evaluators to score the quality of entity handling (as described in Section 3.2). For that task each sentence known to contain a specific entity type is given a binary score of whether or not the entity is translated correctly. We found having a separate evaluation specific to entities in addition to adequacy and fluency is important as it helps us to decide on tokenization procedure, entity handling procedures, etc.

4.2 Business sensitivity analysis

One important shortcoming of the BLEU score is that it says nothing about the so-called “business sensitive” errors. For example, the cost of mistranslating “Parking is available” to mean “There is free parking” is much greater than a minor grammatical error in the output. Typically

\[2\]https://www.taus.net/academy/best-practices/evaluate-best-practices/ adequacy-fluency-guidelines
Table 5: Business-sensitive translation errors analysis for English-German pair for the “parking availability” aspect.

(a) Performance of English and German components of our BSF framework measured with a hold-out set of 500 examples.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EN</td>
<td>DE</td>
<td>EN</td>
</tr>
<tr>
<td>Free parking</td>
<td>0.97</td>
<td>0.96</td>
<td>0.93</td>
</tr>
<tr>
<td>Non-free parking</td>
<td>0.79</td>
<td>0.83</td>
<td>0.89</td>
</tr>
<tr>
<td>Not about parking</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>Average</td>
<td>0.97</td>
<td>0.96</td>
<td>0.97</td>
</tr>
</tbody>
</table>

(b) The result of applying BSF to our English/German corpus, expressed in matches normalized by the total English volumes. For example out of all English sentences which BSF annotated as “Free parking”, 99.4% also get predicted as “Free parking” in German, while 0.5% of those get identified as “Non-free parking” and 0.1% as not about parking at all.

<table>
<thead>
<tr>
<th>English prediction</th>
<th>Free parking</th>
<th>Non-free parking</th>
<th>Not about parking</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>99.4%</td>
<td>0.5%</td>
<td>0.1%</td>
</tr>
<tr>
<td>Non-free parking</td>
<td>5.1%</td>
<td>94.6%</td>
<td>0.3%</td>
</tr>
<tr>
<td>Not about parking</td>
<td>&lt; 0.1%</td>
<td>&lt; 0.1%</td>
<td>99.9%</td>
</tr>
</tbody>
</table>

Non-free parking can either be a sentence about clearly paid parking, or it can be something ambiguous as “There is parking available nearby”.

4.3 BLEU score vs human-based metrics

While BLEU score is very convenient to use because it can be computed automatically, the main metrics we really trust are human-based (see Section 4.1). Here we look at how the BLEU scores from our English-German corpus size experiment of Section 2.3 are correlated with adequacy/fluency metrics.
5 Conclusion

We have presented our approach to developing a large scale NMT system, specifically focusing on practical considerations. We presented the performance of different optimization strategies for model training in single- and multi-GPU environments. We found that a combination of Adam and SGD with learning rate decay works the best on a single GPU, and asynchronous SGD parallelization is a great strategy to dramatically speed up the training. We presented the advantages of BPE tokenization for machine translation and argued in favor of preprocessing named entities for better quality translation. Finally, we presented our approach of dealing with critical translation mistakes through our business sensitivity framework and argue that despite being the main metric in research, BLEU score alone can be a poor way of tracing MT system improvement.

In the future we are going to continue running optimization related experiments, particularly around better strategies for taking advantage of multiple GPUs. In order to leverage our massive monolingual corpora that are not translated, we are also focusing more on the research topics of model pre-training and similar techniques. Other important research topics to us are domain adaptation and user-generated content.

Acknowledgements

We would like to thank our Language Specialists for providing invaluable human feedback and to Darina Kozlova for her important advice on human evaluation and for patiently coordinating all that work.
References


The INTERACT Project & Crisis MT

Sharon O’Brien, Chao-Hong Liu, Andy Way (DCU)
João Graça, André Martins, Helena Moniz (Unbabel)
Ellie Kemp, Rebecca Petras (Translators without Borders)

Overview of Interact Project

• International Network on Crisis Translation
• EU-funded Marie Curie Networking Project (with research outputs)
• Based on the main premise that:

  – In today’s age of globalisation, communication during a crisis must be multilingual and multilingual crisis communication is enabled through translation
Structure

• Brief overview of project
• The role of and challenges for MT in crises
• Previous work
• What we plan to do for MT

What do we mean by ‘Crisis’?

• “An event that is expected to lead to a dangerous situation, whether it is an emergency or a disaster”, Lighthouse Readiness Group

• Project focus:
  – Written Translation
  – Health Content
Partners in the INTERACT Project

- University of Auckland
- Arizona State University
- Translators Without Borders
- Unbabel
- Cochrane
- Microsoft
- Dublin City University (Coordinator)

Research Work Packages

- Crisis MT
  - Simplification
  - Training Citizen Translators
- Policy
- Ethics

INTERACT/Crisis MT  sharon.obrien@dcu.ie

@CrisisMT
Focus For Today

Why is (Machine) Translation Important in a Crisis?

- Clear, accurate, timely information is essential in a crisis
  - e.g. Seeger 2006; Fischer 2008; World Health Organisation 2012; Infoasaid 2012; Santos Hernández and Morrow 2013

- Greater cooperation between humanitarian agencies and linguistic volunteers is required
  (Harvard Humanitarian Initiative 2011)

- but...

- Little to no recognition of the fact that those in need of information may not speak the dominant ‘response’ language
Nature of Crises

• May have sudden onset
• Unpredictable language combinations and information needs
• Unpredictable duration
• Highly stressful
• Lives are at risk

• But also consider: the 4Rs of Disaster Management:
  – Risk
  – Response
  – Recovery
  – Resilience

  i.e. Translation is not just required during the response stage

Examples of Previous & Ongoing Work

• MT in response to the Haiti Earthquake (Lewis 2010
  – “Cookbook” for SMT in crisis situations (Lewis et al. 2011)
• DARPA’s Lorelei (Low Resource Languages for Emergent Incidents) project
• TWB’s Rule-Based MT systems
  – English-Kurmanji, English-Sorani via Apertium
Challenges for Crisis MT

• Unknown time of occurrence (no training in advance?)
• Unknown language pairs
• (Often) low resource languages
• (Often) no/little parallel data for affected languages
• (Sometimes) highly specialised (communicating risk of disease, nuclear threat etc.)
• (Sometimes) no/low power and Internet connections
• Text to Speech may be required too

Pivoting as Potential Solution?

• E.g. Arabic-speaking refugees arrive on a Greek Island. Responders speak Greek and limited English. Refugees speak Arabic, some have limited English, and no Greek
• For emergency response purposes, e.g. ascertaining the state of health of refugees, – we need Arabic <> Greek translation urgently
• We do not have sufficient translators/interpreters
• We have Arabic-to-English and English-to-Greek Engines, but no Arabic-to-Greek engines
Pivoting as Potential Solution

• Our questions: Could we use:
  – Arabic < > English < > Greek in this situation?
  – What is the quality like?
  – What impacts the quality?

Approaches to Pivoting

• Naïve Approach (Utiyama & Isahara 2007)
  – Translate from A to B
  – Use B as source to translate to C
Approaches to Pivoting

• Interpolated Direct Engine (Wu and Wang 2007)
  – Applies only to PBMT
  – A-to-C phrase table is derived from available trained A-to-B and B-to-C models
  – Combined with language model available for C to build direct engine from A-to-C

• Neural Interlingua Approach (Johnson et al. 2016)
  – ‘One” Neural Network is trained with A-to-B and B-to-C sentence pairs
  – NN is used to translate A-to-C even though no A-to-C sentence pairs are used in the training
  – Performance improved if small amount of A-to-C sentences are used in training
Our Pivot Triplets

1. Greek < > English < > Arabic
2. German < > English < > Arabic
3. French < > English < > Swahili

The task will involve identifying relevant training resources, building sample crisis MT engines for health content, and evaluating the results using standard MT evaluation techniques (automatic and human evaluation metrics).

The Human Factor

• (Citizen) Translators
• End Users, e.g. First Responders

• How do we train these people to work with MT in crises?
The Human Factor

• Development of training materials for Citizen Translators who volunteer in Crisis Scenarios

• One specific focus will be post-editing
  – E.g. Through the Unbabel crowdsourcing platform and TWB’s and Cochrane’s Networks

• Train, evaluate, re-design, train, evaluate...
• Peer review via the crowd (e.g. using Unbabel’s online MQM annotation tool)
• Train the trainer
• Training in support resources, e.g. Slándáil, ReliefWeb, TWB harvest of terminology from Sphere Handbook, etc.

The Human Factor

• Focus on end users too:
  – Ethical and informed use of MT for crisis scenarios
  – Use of Quality Estimation for triaging MT output so unreliable translation is not even seen by post-editors/end users
To Conclude

- Follow our progress on Twitter: @CrisisTranslation
Abstract

The European research project Social Sentiment Indices powered by X-Scores (SSIX) intends to allow Small and Medium-sized Enterprises (SMEs) to take advantage of social media sentiment data for the finance domain. The project aims to overcome language barriers and realize a financial sentiment platform capable of scoring textual data in different languages.

Our approach to achieve this goal takes maximum advantage of human translation while keeping costs low by incorporating machine translation. In the long run, we intend to provide a tool that helps SMEs to expand into new markets by analyzing multilingual social contents.

In this paper, we investigate how sentiment is preserved after machine translation. We built a sentiment gold standard corpus in English annotated by native financial experts, and then we translated the gold standard corpus into a target corpus (German) using one human translator and three machine translation engines (Microsoft, Google, and Google Neural Network) which are integrated in Geofluent to allow pre-/post-processing. We then conducted two experiments. One meant to evaluate the overall translation quality using the BLEU algorithm. The other intended to investigate which machine translation engines produce translations that preserve sentiment best.

Results suggest that sentiment transfer can be successful through machine translation if using Google and Google Neural Network in Geofluent. This is a crucial step towards achieving a multilingual sentiment platform in the domain of finance. Next, we plan to integrate language-specific processing rules to further enhance the performance of machine translation.
1. Background

Over the past two years, Lionbridge has been involved as a leading industrial partner in the European funded **SSIX project** (Social Sentiment Index, 2015 - 2018). During the project (which will be completed in February 2018), we have developed a platform for detecting opinions about stocks, companies and their products as expressed in social media and other media sources. For example, we can extract content from Twitter, StockTwits, news, company blogs, etc and analyze sentiment associated to each content.

In Lionbridge, we conceive the SSIX platform as a supporting tool for our sales representatives. Our goal is to make it easier to detect the following aspects:

- What are the needs of our customers
- What prospects may be entering within our areas of expertise
- What are the weak and strong points of our competitors

We consider such knowledge as strategic to trigger appropriate action in real time. For example, we can track customers’ needs on social media and adjust our services accordingly in real time; we can detect events that are relevant to our interests and deal with them strategically.

In the past, a sales representative would need to search different sources in an accessible locale to find relevant discussion of new products or market updates. This was done in the past manually to a large extent. Such manual approach may not be ideal for many reasons: it is prone to missing information, slow in response time, and expensive in terms of human labor.

Now the SSIX platform offers the possibility to partially automate the search. It allows search terms and media channels to be defined, and it notifies users of changes amongst public opinion. It allows us to see what people say about products and companies in real time. Furthermore, this is not restricted to a specific language and locale. Thanks to the integrated technology of Lionbridge GeoFluent (GeoFluent, Lionbridge Inc.), we can overcome the language barrier and provide financial sentiment analysis across languages.

2. Introduction

One of the primary targets of the SSIX project is sentiment analysis in the financial domain across multiple languages. The work has started with English, where a three-way validated sentiment gold standard has been developed and has been used to train the sentiment classifier. The work on English can rely on several available resources, such as text normalization tools, polarity lexica and distributed word representations that allow the development of a sentiment classifier for English to be based on pre-existing resources.

The work started with building a three-way validated sentiment gold standard corpus for English (Hürlimann et Al., 2016). Three experts in the domain of finance annotated the English corpus manually, and their sentiment scores were reconciled for consistency. This gold standard corpus was used to train and test the SSIX sentiment classifier.
Addressing languages different from English, however, is a more complex issue that raises a series of questions. Resources for other languages may neither be as readily available, nor as good in quality. This raises the question whether it is possible/sufficient to rely on the resources we have for English to address sentiment classification for other languages. Suppose, as it is in fact the case, that we want to develop a sentiment classifier for German when we already have a working version for English. Is there a way to capitalize on the resources developed for English to create a classifier for German?

To answer this question, we suggest at least three approaches:

1. Create a gold standard corpus for German from the ground up, manually annotate and cross review it, and then train the new classifier on it. We call this the Native approach.

2. Take the English sentiment gold standard corpus, translate it (either manually or automatically) to German, and train the German classifier on it. We call this the Derived approach.

3. Use machine translation to convert the German input to English, and feed the English translations to the English classifier. We call this the Direct Translation approach.

The three approaches obviously differ in quality, efficiency and costs. Each approach has its advantages and disadvantages, which are briefly outlined below.

2.1. The Native Approach

Building a new Gold Standard corpus from scratch, as in the Native approach, is expensive, but potentially very rewarding. The most prominent benefit is that no translation is taking place and the native expert judgments are on “first hand” data. Creating such a gold standard is both costly and time-consuming, as we need more than one annotator (at least 3) to agree on the sentiment of each piece of text in order to ensure good quality data. Considering that the sample should contain several thousands of tweets and that a domain like Finance needs judgments made by specialists, the cost may quickly skyrocket. On the other hand, the only variable in the Native Approach is the agreement of the annotators, provided their individual domain knowledge and familiarity with the exchange media (tweets) does not lead to vastly different sentiment scores for the same data. Due to the conditions of its design and implementation, we could assume that once available, such a gold standard would be the standard against which any other approach should be benchmarked.

2.2. The Derived Approach

In this approach, instead of building a new corpus and annotating it manually, we use the already existing English language gold standard and translate it to German. This approach presupposes that a statement with positive sentiment in English remains positive in German and vice-versa for negative judgments. Several translation methods are available: It can either be done manually, via machine translation, or in a hybrid way, using computer aided translation tools or post-translation review by human translators. We can also take advantage of the fact that only some words are sentiment-bearing thus targeting these words in context for optimal translation and ignoring the rest.
If we use human translation, the task of creating a translated GS will be cheaper than the creation of a native GS, in the sense that one domain expert will probably be enough, where previously three were needed. Certainly, the cost and time decrease drastically when using machine translation, but the resulting data, especially in a technical domain such as finance, may be of lower quality. Machine translation could, for instance, systematically map an English term to a German term which is synonymous in some other domain, but which is not relevant to the financial domain.

A human-reviewed machine translation is surely the safest approach if one wants to speed up the process and keep costs limited. This may actually reveal error patterns in the translation that can be fixed in post-processing.

2.3. The Direct Translation Approach

Instead of training a new classifier on German data, we translate the German input text to English and feed it to the English classifier. Clearly, translation here can mean only machine translation, as we will be dealing with large amounts of input data to be processed in real time. This approach can also add further costs as machine translation on large amounts of data comes at a cost.

The translation-based approaches in 2 and 3 face a number of issues related to the domain and the specificity of the text involved. Spelling errors, uncommon abbreviations and rhetorical text are all extra challenges that need to be tackled.

Input normalization and output optimization are strategies that can be pursued to improve the quality and accuracy of the translation. First, we may remove elements like repeated characters or delete unknown strings. During post-analysis of translated material, we can map common MT mistakes to the desired output, for instance, terms that need a specific translation in the domain of reference. There is a large range of operations that can be performed – some language-specific, some more general. In this respect, GeoFluent [2] is specifically designed not only to support automatic translation but also in preparing the input and correcting the output of the translation process (pre- and post-processing of the data).

3. Setup

The work discussed in this paper is a contribution to the Derived and Direct Translation approaches.

Within the scope of the SSIX project, we built a sentiment gold standard corpus for English, annotated by native experts from the domain of finance (Hürlimann et Al., 2016). The gold standard corpus was translated into a target corpus in German by a domain expert. At the same time, it was also translated into German by three machine translation engines. These are Microsoft, Google, and Google Neural Network, which are integrated in Lionbridge GeoFluent [2]. We used GeoFluent to introduce pre-/post-editing, such as DO-NOT-TRANSLATE rules to tackle special financial terms and text normalization rules.

In SSIX, we intend to take maximum advantage of human translation while keeping the cost low by incorporating the machine translation component. Our objective is to use manually translated data as a benchmark and examine machine translation outputs: their quality and preservation of sentiment in the financial domain.

A crucial prerequisite for our approach is that the sentiment of the gold standard corpus can be transferred to the target corpus after translation. If the sentiment is lost after translation,
either by human or by machine, we cannot use our previous research results, i.e. the English sentiment classifier, and implement either the Derived approach or the Direct Translation approach. The only viable option left would be the Native approach, which is bound to have high costs. As a result, to meet the prerequisite and make decisions for further actions, we must investigate the impact of machine translation on the sentiment quality of the gold standard corpus. We have conducted two experiments to study how machine translation influences sentiment, as discussed below.

4. Experiment 1

The first experiment was designed to find out the quality of each machine translation engine. In this experiment, we selected a sample of 700 English tweets from Twitter and StockTwits relative to the financial domain. This data set was selected for its clarity in expressing sentiment. For example, textual data that did not offer valuable information such as containing only URLs was filtered out to reduce noise.

During the experiment, this sample was translated into German simultaneously by one human translator and the three machine translation engines mentioned above, namely Microsoft, Google, and Google Neural Network, as integrated in Lionbridge GeoFluent. The human translator is a native speaker of German and a domain expert in finance.

To evaluate translation quality for the three machine translation engines, we calculated their BLEU scores (Koehn et al., 2007; for source code see References). Using human translation as the reference, the three machine translations were each compared to the human translation to see how close they are to the professional human translation.¹

The results are summarized in the table below. They suggest that Google and Google Neural Network performed better than Microsoft on 1-gram, and Microsoft performed better than Google and Google Neural Network on 2-grams, 3-grams, and 4-grams.

<table>
<thead>
<tr>
<th>Engine</th>
<th>1-gram</th>
<th>2-grams</th>
<th>3-grams</th>
<th>4-grams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>0.901470798</td>
<td>0.865873923</td>
<td>0.786125067</td>
<td>0.684824095</td>
</tr>
<tr>
<td>Google</td>
<td>0.963509145</td>
<td>0.846959705</td>
<td>0.728174371</td>
<td>0.605465403</td>
</tr>
<tr>
<td>Google Neural Network</td>
<td>0.963340387</td>
<td>0.846025029</td>
<td>0.727096883</td>
<td>0.604167208</td>
</tr>
</tbody>
</table>

Table 1. BLEU score for machine translations

The 1-gram is used to assess how much information is retained after translation. Clearly Microsoft has lost more information than both Google and Google Neural Network. Among 2-grams, 3-grams, and 4-grams calculations, 4-grams is believed to be the most correlated with judgements made by native speakers of the target languages (Papineni, K., et al., 2002).

¹ We understand that BLEU score is meant to evaluate translations on a corpus level. However, due to time and resource limitations, at this stage we can only investigate the current data sample size. We consider expanding our data size and reduplicating this experiment in order to confirm our results in future.
Our results suggest that Microsoft produced the most similar translations to human translator. Google and Google Neural Network performed more poorly in comparison.

However, we must notice that the BLEU algorithm was not sufficient for our purposes because it only evaluates translation quality in the respect of approximating human translation. Since the purpose of SSIX is to build a sentiment platform, we consider the quality of translation is the best when there is minimal discrepancy in sentiment between the original texts and the translations. Using our criterion, we need to explore the sentiment preservation. That is why we conducted Experiment 2.

5. Experiment 2

4.1 Experiment Design

For Experiment 2, we selected a subset of the previous sample (N = 200). We had to reduce the size of our sample because Experiment 2 required much more human resources than Experiment 1. To keep the time and expense cost under control, we chose a subset of the previous sample.

This experiment was designed to investigate whether translations (regardless of whether they came from human translators or machine engines) can maintain the sentiment from the original texts. As the first step, we recruited two German financial domain experts and they assigned sentiment to all four translations. The experts were kept away from the original English texts and their sentiment.

The sentiment scores assigned by the domain experts ranged from 1 to 10, 1 being the most negative, and 10 being the most positive. If the assigned pair of scores for a certain line of text diverged from each other for more than 2 points (including 2), we asked a third domain expert to evaluate the text again and chose the more appropriate sentiment score from the two alternatives.

For example, the human translator translated a certain tweet into German: "Der miterlebte Fortschritt ist echt atemberaubend." - Stifel Analyst, nachdem er Teslas Fabrik zum vierten Mal gesehen hat $TSLA$ https://t.co/nD7KECoM6V

Its original English tweet is: The progress witnessed is truly stunning." - Stifel analyst after seeing Teslas factory for the fourth time $TSLA$ https://t.co/nD7KECoM6V

One of our domain experts assigned the German translation a sentiment score of 3, and the other assigned it a 10. Since there was a big gap between the two scores, the third domain expert evaluated the translation, and chose 10 from the pair of 3 and 10. As a result, the sentiment score for this tweet is 10.

4.2 Results and Discussions

After the data were evaluated and reconciled in the above way, we performed some statistical analysis on the results. We used a mixed linear regression model, which was implemented with the lmer4.0 package in R (Federico et al., 2014; Guzman et al., 2012). Compared with a linear regression model, a mixed effects model can explicitly model individual character-
istics. In our design, we used the item as a random intercept to capture the variance of each translated item to maximize the differences we could find between compared sets.

We are mainly concerned with the following two questions:

- Do human translations preserve sentiment?
- Does machine translation preserve sentiment?

To answer the first question, we need to compare the sentiment of the English gold standard corpus with the sentiment of human translation. If there was no significant difference between the sentiment scores of English gold standard and human translation, we would know the sentiment did not change too much; if a significant difference was found, then the sentiment is already lost in human translations.

After calculating our data set, results showed that there was no significant difference between the sentiment of English gold standard and human translation (Figure 1). In other words, the difference between gold standard sentiment (mean = 5.674) and human translation sentiment (mean = 5.536) was not large enough for us to draw the conclusion that they are different on a statistical level. This proves that human translation can preserve sentiment from the original texts. The results are what we desire to see because human translation is believed to be more reliable than machine translation. If human translation could not preserve sentiment, it is unlikely that machine translation can.

Next, we try to answer the second question and assess the performance of machine translation engines on sentiment preservation. We compared the sentiment of the English gold standard with the sentiment of machine translations. Our results suggested that there were significant differences between the three pairs, i.e. English gold standard vs. Microsoft, English gold standard vs. Google, and English gold standard vs Google Neural Network (Table 2).

<table>
<thead>
<tr>
<th>Engine</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>t = -3.574</td>
<td>p &lt; .001</td>
</tr>
<tr>
<td>Google</td>
<td>t = 2.038</td>
<td>p &lt; .05</td>
</tr>
<tr>
<td>Google Neural Network</td>
<td>t = 3.101</td>
<td>p &lt; .01</td>
</tr>
</tbody>
</table>

Figure 1. Sentiment Comparison: Gold standard vs. Human
Table 2. Results for Sentiment Comparison (Gold standard vs. Machine)

The visualization of the result can be found in Figure 2. Here Microsoft shows stronger diversion from the original sentiment in the gold standard, and Google produced the sentiment that was the closest to the original.

We also notice that compared to the gold standard sentiment mean, both human and machine translations have sentiment with lower means. At least two factors attribute to this fact. One is that translations have “neutralized” sentiment, drawing its mean closer to the grand mean (i.e. 5.5) because translations always lose information to an extent. The other is due to our domain experts. We used different groups of domain experts for annotating sentiment of English and German data, who are English and German native speakers respectively. Our German annotator could be more conservative or negative in assigning sentiment scores.

Figure 2. Sentiment Comparison: Gold standard vs. Machine

These results indicate that translations generated by machine engines are not of the desired high quality and look to be at risk of losing or distorting sentiment. However, they do not imply that machine translation is without merit. Since we have established that human translation is successful in preserving sentiment, we can use human translation as the benchmark to compare machine translations. If the sentiment assigned to a given machine translation engine does not deviate significantly from that of human translation, we can conclude that the engine has produced sentiment scores comparable to human translation.

The three comparisons discussed above showed that there are significant differences between the sentiment of human translation and Microsoft, which indicates that the Microsoft engine did not produce translations whose sentiment was alike to human translation (Table 3). The visualization is provided in Figure 3.

---

2 The * on top of the bars indicated significance
<table>
<thead>
<tr>
<th>Engine</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>$t = -2.16$</td>
<td>$p &lt; .05$</td>
</tr>
</tbody>
</table>

Table 3. Results for Sentiment Comparison (Human vs. Machine)

Crucially, there was no significant difference between the sentiment scores of human translations and both Google and Google Neural Network. This means that the sentiment scores from Google and Google Neural Network does not differ significantly from human translation. This proves that these two engines’ performance was in line with human performance, and consequently in these cases, sentiment can be considered as successfully preserved.

6. Conclusion

In this paper, we provide evidence that sentiment can be preserved after translation of an English gold standard corpus into German by machine engines, namely Google and Google Neural Network when they are integrated in GeoFluent. With this prerequisite fulfilled, we can either use the Derived approach to convert English data to another language and subsequently train a sentiment classifier on that data. Alternatively, we can use the Direct Translation approach to transfer multilingual data to English and use our already built English sentiment classifier. As these approaches do not need a human translator, time and costs can be greatly reduced, without an apparent, major loss in quality for the purposes of sentiment analysis. This is a crucial step for building an affordable multilingual sentiment platform in the domain of finance, to overcome the language barriers and help SME to analyze multilingual social content.

We have many directions for further research in the future that go from the integration of more language-specific processing rules in GeoFluent to enhancing the performance of machine translation, to benchmarking financial sentiment classifiers trained with Native and Derived approaches.
ACKNOWLEDGMENTS

This work is funded by the SSIX Horizon 2020 project (Grant agreement No 645425). 3

References


GeoFluent (Lionbridge Inc.) http://www.lionbridge.com/GeoFluent/


Source code for calculating the BLEU score:
http://www.nltk.org/_modules/nltk/translate/bleu_score.html

---

3 The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.
A New Methodology to Maximize the Strength of SMT and NMT

MT Summit XVI

Yu Gong
August 14th, 2017

SMT vs. NMT


2. [http://opennmt.net/](http://opennmt.net/)

Figure-1

Figure-2
Which one is better?

- More and more attention to Neural MT
- Improved translation quality over SMT
- A milestone in machine translation

Is it true?

The Data

- Selection of “real world” customer data collected over a three month period
- Catalogue of technical tools
- German → English
- ~ 5,000 Segments

Automatic Evaluation Results

<table>
<thead>
<tr>
<th></th>
<th>NMT</th>
<th>Moses</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>23.68</td>
<td>47.98</td>
</tr>
<tr>
<td>METEOR</td>
<td>28.46</td>
<td>38.26</td>
</tr>
</tbody>
</table>

Table 1: BLEU and METEOR scores.

Anne Beyer, Vivien Macketanz, Aljoscha Burchardt and Philip Williams, Can out-of-the-box NMT Beat a Domain-trained Moses on Technical Data? EAMT 2017
Some Findings

- Professional translators prefer translations of NMT systems over translations of the SMT systems*
- NMT systems are better at handling word ordering and morphology, syntax and agreements (including long distance agreements) than the SMT systems*
- SMT systems are better at handling terminologies than the SMT systems
- Human comparative evaluation is crucial when comparing MT systems from fundamentally different approaches*
What Can We Do?

Open Machine Translation Toolset (OMT²)

- Streamline the process of creating workable MT models
- Help users choose the best model by evaluating MT output
- Integrate machine translation into enterprise localization process
- Enable users to try the latest machine translation technology with least effort
**Architecture**

**OMT² Features**

- Parse TMX: Extract corpus from Translation Memory eXchange (TMX).
- Clean up corpus: Remove garbage tags, those sentences length of which are not suitable for training a MT model.
- Tokenization & Segmentation: Call third party tools to tokenize or segment corpus.
- Split corpus: Split the original corpus by randomly selecting sentences for different purposes: training, validating and testing.
- Train: Train an MT model by calling OpenNMT or Moses scripts.
- Score: Use BLEU to give users a sense of how good the model is.
- Select the best model: automatically choose the model with highest BLEU score.
- Translation: Enable users to translate content by RESTful API.
OMT² RESTful API

Request

Response
{ "translation": {  "SMT": "你好",  "NMT": "你好" }}

Sample Output in CAT Tools

1. Po the Panda is the laziest animals in all of the Valley of Peace, but unwittingly becomes the chosen one when enemies threaten their way of life.

   SMT: 宝熊是和平谷中最懒惰的动物, 但是当敌人威胁生活方式时，不知不觉地成为选择的动物。

   NMT: 熊猫是所有和平谷中最懒的动物, 但是当敌人威胁他们的生活方式时, 它不知不觉地变成了一个被选中的人。
Demo

Q&A
Thank You

gongy@vmware.com
Rule-based MT and UTX Glossary Management – Honda’s Case
Dealing with Thousands of Technical Terms

MT Summit 2017 (Nagoya, Japan)

Saemi Hirayama
CAT tool leader, Honda R&D Americas, Inc.
Yuji Yamamoto
Founder/representative, CosmosHouse

Contents

1. Speakers
2. Honda MT overview
3. Issue 1: MT migration
4. Issue 2: term inconsistency
5. Terminology management continues
Speakers

Saemi Hirayama
An in-house translator/CAT tool leader at the Ohio Center of Honda R&D Americas, Inc.

Yuji Yamamoto
Founder/representative, CosmosHouse
<http://cosmoshouse.com>
UTX team leader at AAMT (Asia-Pacific Association for Machine Translation)

Honda R&D Americas Inc.
(hereafter referred to as Honda R&D)

Facilities in North America
Multiple facilities strategically located close to production facilities

<table>
<thead>
<tr>
<th>West to East</th>
</tr>
</thead>
<tbody>
<tr>
<td>HRA Silicon Valley Lab (HSVL)</td>
</tr>
<tr>
<td>HRA Los Angeles Center</td>
</tr>
<tr>
<td>HRA Advanced Design Studio (Downtown Los Angeles)</td>
</tr>
<tr>
<td>HRA Denver</td>
</tr>
<tr>
<td>HRA Cincinnati</td>
</tr>
<tr>
<td>HRA Ohio Center</td>
</tr>
<tr>
<td>HRA OSU (MIX)</td>
</tr>
<tr>
<td>HRA Detroit</td>
</tr>
<tr>
<td>HRA Florida</td>
</tr>
<tr>
<td>HRA South Carolina</td>
</tr>
<tr>
<td>HRA Canada (Markham, Ontario)</td>
</tr>
<tr>
<td>HRA Burlington, N.C.</td>
</tr>
<tr>
<td>HRA North Carolina</td>
</tr>
<tr>
<td>HRA Halifax, Nova Scotia, Canada</td>
</tr>
</tbody>
</table>
Honda R&D Americas Inc.
Creating New Value in the U.S.

- Product Research & Development
- Product Styling Design
- Environmental Technology Development
- Safety Technology Research and Development

**Honda R&D’s needs**

1. JA to EN, EN to JA
2. Technical documents written by engineers
3. Used for translation needs by global associates in daily business operations
4. Term-level accuracy and consistency are important
5. Speed is crucial
Honda R&D MT overview

Over a decade ago, Honda R&D adopted an RBMT (Rule-based Machine Translation) system

The current MT is a RBMT subsystem

Engineers use it to translate documents and emails

In-house translators also use it to process translation requests from engineers

Honda R&D MT overview

• Honda Jargon dictionaries categorized and added to the MT

• A feedback function added to the web-based MT for mistranslations/ unregistered terms to keep the dictionaries up-to-date
Honda R&D MT achievement

- In-house translations reduced and outsourcing cost cut by half
- Significant translation speed increase
- Better communication with accurate technical terms

Why was RBMT chosen at Honda R&D?

1. 80,000 terms, including many Honda-only terms
2. Many incomplete fragmental phrases/very few fixed phrases
3. File formats: complicated slides
4. No two documents are alike
Why is neural/statistical MT not used at Honda R&D?

1. Term-level accuracy and consistency are poor in NMT
2. Human-translated corpus is too small
   Because the majority of translations are lightly post-edited machine translations
3. Protection of intellectual property and secrecy
4. Higher cost
5. Most documents do not repeat

Issue 1: MT migration (RBMT to RBMT)

80,000 Honda terms in Fujitsu ATLAS were needed to be imported into a new MT system, Toshiba’s The Honyaku.

<table>
<thead>
<tr>
<th>側圧</th>
<th>side pressure</th>
</tr>
</thead>
<tbody>
<tr>
<td>傘裏</td>
<td>back of umbrella</td>
</tr>
<tr>
<td>傾斜角度</td>
<td>inclination angle</td>
</tr>
</tbody>
</table>

No Honda terms

RBMT (ATLAS) customized dictionaries

80,000 terms

RBMT (The Honyaku) customized dictionaries
Solution 1: MT migration (RBMT to RBMT)

Solution: Conversion through UTX format. The customized dictionaries were transferred to the new MT.

<table>
<thead>
<tr>
<th>RBMT (ATLAS) customized dictionaries</th>
<th>UTX</th>
<th>RBMT (The Honyaku) customized dictionaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>側压</td>
<td>side pressure</td>
<td>側压</td>
</tr>
<tr>
<td>傘裏</td>
<td>back of umbrella</td>
<td>傘裏</td>
</tr>
<tr>
<td>傾斜角度</td>
<td>inclination angle</td>
<td>傾斜角度</td>
</tr>
</tbody>
</table>

80,000 terms

UTX – glossary standard for sharing and reuse

UTX (Universal Terminology eXchange)
- Developed by AAMT, initially as RBMT dictionary data exchange format
- Later restructured as a structured glossary format
- Used by companies and organizations such as Japan Patent Office

http://www.aamt.info/english/utx/
Issue 2: term inconsistency

Identical Honda jargon was being translated inconsistently at various company sites around the world.

Correlation meeting?
Coordination meeting?
...................................?

“整合会”

“correlation meeting”
“arrangement meeting”
“collaboration meeting”
“coordination meeting”
“adjustment meeting”

Honda Terms were being translated inconsistently in 6 Regions Worldwide
Solution 2: term inconsistency

• Term statuses (approved, non-standard, forbidden etc.) were added.

• 1:n, n:1, n:n source/target term pair relationships are clearly defined.

• J to E glossary now also works as E to J.

<table>
<thead>
<tr>
<th>#term:ja</th>
<th>term:en</th>
<th>term status:ja</th>
<th>term status:en</th>
</tr>
</thead>
<tbody>
<tr>
<td>整合会</td>
<td>correlation</td>
<td>approved</td>
<td>approved</td>
</tr>
<tr>
<td></td>
<td>meeting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>整合会</td>
<td>coordination</td>
<td></td>
<td>forbidden</td>
</tr>
<tr>
<td></td>
<td>meeting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>整合会</td>
<td>collaboration</td>
<td></td>
<td>non-standard</td>
</tr>
<tr>
<td></td>
<td>meeting</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Terminology management at Honda R&D

• 2013: UTX was introduced, transferring 80,000 terms from the old MT to a new one.

• 2014: a terminology committee was established to review existing/new terms to update the MT dictionaries monthly.
Reported useful by 98% of users

196 respondents: local US staff 80%, Japanese staff 20%

Terminology management continues

1. Review terms and term statuses
2. Add new terms
3. Delete unnecessary/obsolete terms
4. Categorize terms

…to improve translation accuracy and efficiency
For fellow MT user companies

• Glossaries control your company vocabulary
  Quality of human/machine translation can be improved with terminology management
• Proper terminology management pays off!

Solution 2: term inconsistency

• Term statuses (approved, non-standard, forbidden etc.) were added.
• 1:n, n:1, n:n source/target term pair relationships are clearly defined.
• J to E glossary now also works as E to J.

<table>
<thead>
<tr>
<th>#term:ja</th>
<th>term:en</th>
<th>term status:ja</th>
<th>term status:en</th>
</tr>
</thead>
<tbody>
<tr>
<td>整合会</td>
<td>correlation</td>
<td>approved</td>
<td>approved</td>
</tr>
<tr>
<td>整合会</td>
<td>coordination</td>
<td>forbidden</td>
<td>non-standard</td>
</tr>
</tbody>
</table>

Proceedings of MT Summit XVI, Vol.2: Users and Translators Track  
Nagoya, Sep. 18-22, 2017 | p. 77
Future actions at Honda R&D

1. Further tuning of UTX glossary
2. UTX for terminology check
   • Can be used for post editing neural/statistical MT if necessary
   • Terminology tool training for translators

Take away

1. Glossaries are necessary for both humans and MT
2. UTX glossary management has been effective at Honda R&D
3. Neural MT may not be the only future – users are satisfied with RBMT
A detailed investigation of Bias Errors in Post-editing of MT output

Silvio Picinini
Localization, eBay Inc., San Jose, CA, USA
spicinini@ebay.com

Nicola Ueffing
Machine Translation Science Lab, eBay Inc., Kasernenstr. 25, Aachen, Germany
nueffing@ebay.com

Abstract

The use of post-editing of machine translation output is increasing throughout the language technology community. In this work, we investigate whether the MT system influences the human translator, thereby introducing "bias" and potentially leading to errors in the post-editing. We analyze how often a translator accepts an incorrect suggestion from the MT system and determine different types of bias errors. We carry out quantitative analysis on translations of eCommerce data from English into Portuguese, consisting of 713 segments with about 15k words. We observed a higher-than-expected number of bias errors, about 18 bias errors per 1,000 words. Among the most frequent types of bias error we observed ambiguous modifiers, terminology errors, polysemy, and omissions. The goal of this work is to provide quantitative data about bias errors in post-editing that help indicate the existence of bias. We explore some ideas on how to automate the finding of these error patterns and facilitate the quality assurance of post-editing.

1. Introduction

The use of machine translation (MT) for facilitating the work of translators is increasing throughout the language technology community. The human translator receives an automatically generated translation from the system, and then corrects the errors made by the system. This is called post-editing. As post-editing will gain even more importance, we believe that the quality of this work needs to be evaluated. Translations suggested by MT systems contain errors, and - for several reasons, such as time pressure - the posteditor might leave these MT errors uncorrected. We are calling this effect “bias”, as in the posteditor being “biased” by the MT suggestion, and accepting translation errors.

In our work, we investigated whether the MT system influences the human translator, thereby introducing bias and potentially leading to errors in the post-editing. We analyzed how often a translator accepts an incorrect suggestion from the MT system. Furthermore, we explored the types of bias errors and performed a quantitative analysis.

Our analysis was carried out on translations of eCommerce data from English into Portuguese, consisting of 713 segments with about 15k words. In addition to the MT output and the post-editing, we carefully curated a golden post-editing reference. Using this golden reference, we calculated edit distances and related scores, and then classified and quantified the types of errors that emerged. We observed a higher-than-expected number of bias errors, about 18 bias errors per 1,000 words. Among the most frequent types of bias error we observed ambiguous modifiers, terminology errors, polysemy, and omissions.

The goal of this work is to provide quantitative data about bias errors in post-editing that helps indicate its existence. Additionally, we will provide data about certain types of error patterns that lead to bias. We explore some ideas on how to systematically find these error patterns.
and facilitate the quality assurance of post-editing. Educating post-editors about bias and about these patterns can help improve the quality of the post-editing work, and therefore the final translation quality delivered to the user.

An early analysis of post-editing of machine translation output is presented in (Krings, 2001). This publication discusses the post-editing process and the quality of machine translations and post-editing, but does not have a quantitative analysis of errors. More recently, (Blain et al., 2011) presents a qualitative analysis of post-editing, focusing on reducing the post-editing effort. In addition to this analysis, the authors present methods for learning corrections from post-editings and improving the MT systems which generated the translations.

2. Analysis of Bias Errors

2.1. Data

We worked on translations from English into Portuguese in the eCommerce domain. The text are descriptions of items which are for sale on the eBay site. The English descriptions, consisting of 713 segments with 15k words in total, were automatically translated using the Microsoft statistical machine translation system, and were post-edited by a human translator, whom we will call post-editor 1 going forward. These post-editings were carefully reviewed by another language expert, whom we will call post-editor 2, who created perfect translations to be used as golden references.

2.2. Methodology

We performed a detailed manual analysis of the post-editings from post-editor 1, comparing them against the golden reference from post-editor 2, in order to detect bias errors. For each error corrected by post-editor 2, we analyzed source, machine translation, and post-editing for potential bias. We classified the errors into certain groups which will be described in section 3.

We used edit distance (Word Error Rate – WER) in two significant ways. First, the distance calculated between the machine translation and the post-editing. This is an indication of where post-editing happened and how much. Based on those data, we developed a process (described in a section below) to identify instances of lack of post-editing:

- If the post-editor does not post-edit a segment (for example, by skipping it), the edit distance is zero. This could look like all MT errors were accepted and there was bias, but in reality the posteditor just missed the entire segment. We wanted to find and exclude these instances from the bias analysis.
- If the post-editor rushes through the task and make just one change in a segment, and there were others to make, this will result in a low edit distance. This would look like bias when it is not bias, it is just lack of proper post-editing. We also wanted to find these instances and exclude them.

Second, we used the edit distance between the golden reference and the post-editing. The primary use of it was to triage the segments to be analyzed. If the edit distance was zero, this meant that the golden reference agreed in full with the post-editing, so this segment should not be part of the analysis.

The edit distance between post-editing and golden reference can indicate:

- If the edit distance is low, this is an indication that the post-editing was generally good and not many changes were needed.
- If the edit distance is significant:
  - There could be a lack of knowledge – the post-editing made changes and they were wrong, so the golden reference corrected this. This
could appear as high PE-MT distance and also high Golden-PE distance.

- There could be bias – the post-editing accepted the MT and the golden reference changed it. This could appear as lower PE-MT distance and higher Golden-PE distance.

We looked into the numbers for the edit distance through the WER scores, see Table 1. The results are consistent with our expectations: The PE-Golden is higher with bias than without it, which means that there were more corrections for bias segments, as expected. The PE-MT is slightly lower with bias compared to without it, which means that there were fewer changes by post-editors in segments with bias, and therefore they left more errors in them.

<table>
<thead>
<tr>
<th>avg. WER</th>
<th>All</th>
<th>without bias</th>
<th>with bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE vs. golden</td>
<td>0.12</td>
<td>0.09</td>
<td>0.20</td>
</tr>
<tr>
<td>PE vs. MT</td>
<td>0.25</td>
<td>0.26</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Table 1. Average WER of post-editing (PE) vs. golden reference and vs. MT output

**Finding and excluding content with lack of post-editing**

The bias that we are trying to identify happens when the post-editor looks at the machine translation and makes a conscious decision to accept it, and the machine translation is wrong. However, it could happen that the post-editor would skip working on a segment, or could make one change at the beginning of a segment and leave the rest untouched. These would not be examples of bias, they would be examples of lack of complete post-editing. In order to try to identify this phenomenon, and exclude it from our analysis of bias, we went through the steps described below.

1. Generated WER scores for each segment, between the post-edited version and the initial MT version (PE-MT).
2. With numbers for each segment, we plotted these numbers on a chart (Figure 1):

![Figure 1. Segment-level WER Post-editing vs. MT](image-url)
This chart shows regions of data where the volume of post-editing seems lower than the rest of the chart. Further investigation showed that the post-editor indeed failed to do a complete post-editing on segments in this region.

3. We looked for a different chart display that would make this phenomenon more visible than plotting the scores. Therefore, we calculated the average of the WER for the past 30 segments, and plotted this rolling average of distances (shown in Figure 2). The orange line is the average for the file.

![Rolling average WER Post-editing vs. MT on 30 segments](image)

Figure 2. Rolling average WER Post-editing vs. MT on 30 segments

This type of chart shows the amount of post-editing effort progressing through the file. If the post-editor, for example, rushes the work towards the end of the file because of a deadline, this will be reflected in a lower WER/edit distance in a series of segments in that region of the file. This lowering will appear in the chart, as the rolling average will go down for that region. This visualization showed two regions of interest (where the chart shows the lowest values), one region around segment number 221 and the other region around segment number 441. After investigating these regions, we confirmed that the second one had segments lacking post-editing.

4. We looked for one more type of chart and we plotted the “Rolling % of zero-WER in 30 segments” and “Rolling % of low WER (<4%) in 30 segments”. In this chart (shown in Figure 3) we tracked the % of zeros in the past 30 segments. A concentration of segments with no post-editing would start to increase the percentage as we move through them, so regions in this chart with peaks are our regions of interest. We did the same for “% of lows” (shown in Figure 4), tracking not only zero changes but also low % of changes up to 4 %. These are segments that could have changed one character, for example.
Both charts were very effective in pointing out regions of interest (highest values on the chart, around segment # 441 as before. While at first it may seem counter-intuitive to look for the highest numbers when talking about low edit distance, it takes just a few seconds to realize that we are looking for “high concentrations” of low scores, and then the peaks on the charts make sense.

2.3. Findings

General observations

“Is there a significant volume of bias?”, that was the question that we wanted to explore for this particular case. While a “Yes” answer can’t be easily generalized to other cases, we hope that there is value in concluding that (1) bias happens and (2) this is an issue that needs attention when thinking about improving the post-editing quality.
Types of errors/causes found

Our analysis did not start with a defined set of standard error types. Instead, as error patterns emerged, they became a type. We are used to error typologies, but the classification used in this work is trying to look at causes of MT errors. Some of these descriptions below will look more clearly like a cause, such as “Modifiers to Multiple Words” or “Multiword expressions” and others may look like a traditional error type, but there is still a cause behind it. Whatever the causes are, we should just keep in mind that these causes created an MT error, and then bias occurred when that MT error was not changed.

- Multiple Modifiers or Words (MMoW) – this pattern describes situations where a modifier may or may not apply to several words around it. This ambiguity is difficult for the MT to resolve. This more frequently applies to nouns and adjectives, but we opted for a more general name because there are some examples of those, and the same principle applies. Examples of this situation are shown in Figure 5 and 6:

![Figure 5. Example 1 of Multiple Modifiers or Words](image)

Veteran musicians, DJs, and public speakers are taken aback by the sensitivity and reliability of the Samson QMIC.

We as humans intuitively know that this is talking about veteran musicians but also veteran DJs and veteran public speakers. However, the MT engine does not know that, and will produce a translation that says, “DJs, public speakers and veteran musicians are taken…”, and the cause of the error is a modifier adjective that applies to multiple nouns.

Another example:

The hypercardioid capsule is built to match frequency and sensitivity of lectern, choir, and boundary mics.

![Figure 6. Example 2 of Multiple Modifiers or Words](image)

In this situation, we have three modifiers applied to one word. We as humans use the context to understand that mics (microphones) probably have frequency and sensitivity and therefore lectern, choir and boundary are three different types of microphones. So this sentence actually means “…lectern mics, choir mics and boundary mics.” The MT does not know that and will produce a translation that sounds like “… sensitivity of boundary mics, and choir and lectern.”, and the cause of the error is multiple modifier (you can see them as nouns or adjectives) that apply to one noun.

This pattern appeared a significant number of times and tends to be difficult for MT. We decided to explore further this pattern in two ways, explained later in this paper:
Can we find this pattern more systematically?

Does this issue occur also for Neural MT?

- Multiword expressions (MWE) – these are issues where a sequence of words has a completely different meaning than the individual words. Examples of this pattern are idioms and phrasal verbs. It is a difficult construction for the MT to handle because of the change in meaning, so it is a cause of errors made by MT. Examples include “makes an impression”, “cut short”, “built in”, “turn over to”.

- Polysemy - Polysemous words are words with multiple meanings and therefore multiple translations. In our case, we look at all issues related to polysemous words that have two competing meanings that are popular in the corpora and confuse the MT engine. This is a cause of errors for MT. Consider, for example, “a choice of restaurants to eat”. The more common meaning of “choice” is probably “to make a choice” but in this example, you are not actually making a choice, and instead the meaning is “variety of options”. If this meaning is less common in the corpora, the MT may make an error. In “performance-conscious photographer”, “conscious” means “photographer concerned with the performance”, but it was translated literally as “did not lose conscience”. So the translation ended up sounding like “performance-did-not-pass-out photographer”.

Other examples include:

- In “Enter a new world of creativity”, “enter” was translated as “insert” as in “enter a password” instead of “walk into” a new world.
- “fleece-lined compartments” had “lined” translated as “aligned” instead of “covered with fleece”
- In “Publishes…materials of benefit to the bar”, “bar” refers to lawyers and was translated as the place to go for drinks.
- In “Washer…including cycles for active wear”, “wear” refers to clothing and was translated with the meaning as in “wear and tear”.

- Mistranslation - In general, “Mistranslation” represents causes that made the MT engine produce a mistranslation. However, every error can be considered a mistranslation. In this work, we classified all possible issues into specific categories. The issues left to be classified as mistranslation are the ones where the translation is wrong but the cause can’t be easily identified. The example of “parent and child” translated as “father and son” should illustrate this category well. We don’t know exactly why the translation is wrong, we only know that it is. This goes into a “Mistranslation” category.

Other Mistranslation examples include:

- “allow concentration to be focused elsewhere” had “elsewhere” translated literally as a location, when “elsewhere” here means focused on “something else”
- “reduces eye fatigue and neck pain” had “neck pain” translated as “throat pain”
- “overcooking” translated as “burned meals”

- Do Not Translate terms – brands and other terms that should not be translated are a cause of errors for MT when the engine has to decide if the term is a brand or a common word. Generic examples would be brands like Gap, Guess or Coach. In our case, examples include the brand Philosophy and a product name called JBL Venue Stadium.

- Terminology – this cause of errors appears when the MT does not know the proper terminology for a certain subject matter. Examples include “focal length” for cameras, “devices” for heraldry, “refrigerator” and “green gas”.

Part of Speech (POS) – we were interested in this specific cause of error, when the MT would translate a word using the wrong POS for it. Some examples we found showed significant ambiguity that would cause MT errors, some of them difficult even for humans to resolve. Examples include:

- “Nuts & Bolts component utilities include…” where “component” is an adjective meaning “utilities that compose the Nuts and Bolts…”. The translation treated it as a noun.
- “The dual apertures of the Vivitar MC Macro Focusing Zoom allow for more flexibility in varying light”, where varying is an adjective meaning “light that can vary”. The translation meant “… more flexibility in the ability to vary light”, treating “varying” as a verb.
- “One example was gilt -- a process presumably done after striking…”, where gilt is a noun (the action named “gilt”) and it as translated as the adjective “gilt”. In sentences where the structure is “<subject> was xxxx”, the xxxx is usually an adjective, but in our example, it was not.

Omission of the initial article – it is a common style in English to omit an article at the beginning of a sentence. Examples with and without article include:

- "AutoCAD LT 2D CAD design software simplifies tasks" vs. “The AutoCAD LT 2D CAD design software simplifies tasks”
- "Zeus IOPS eliminates the wait time" vs. “The Zeus IOPS eliminates the wait time”
- "Familiar six-button configuration provide direct access" vs. “The / A Familiar six-button configuration provide direct access"
- "CenterFlex technology helps enable […]" vs. “The CenterFlex technology helps enable […]"

While readers of English are used to this construction, the MT notices that pattern and consistently produces translations that miss the initial article in several languages. The impact of that is very different from English because readers of those languages are used to the article being present virtually all the time. This is a systematic cause of MT inadequacy. The bias in post-editing consists of not adding the initial article on the target language.

Untranslated words – we found instances of words that should have been translated and were not. Examples include: “POI” (acronym for Points of Interest), "non-resonant", "adaptogen", "dot inlay”. These issues may be linked to these words being out of vocabulary.

Omission – we tracked omissions made by MT and not corrected by post-editing. Examples include: card in “card printers”, sharp.

Addition – same as omission for words added by MT and not removed. Example: added “obtain”.

Prepositions – significant changes of meaning can be caused by a preposition. A different preposition or its omission in the translation may have an impact. Example:

- In “save consumer’s money by reducing the operating costs”, the preposition “by” is what defines the meaning as “the reduction of operating costs is what will save consumers money”. The translation sounded as “save consumers money, reducing the operating costs” and actually reversed the meaning as if “save consumers money” would “reduce the operating costs”.
• “The sole in the Callaway Wedge” was translated as “The sole in wedge shape of the Callaway” changing the meaning just with one preposition.
• Gender agreement – we tracked when the MT made the wrong agreement and was not corrected
• Number agreement – same as gender for singular/plural
• Word order – MT created wrong word orders and they were not corrected.
• Grammar – MT created wrong word orders and they were not corrected.
• Verb tense – MT used the wrong tense for a verb and this was not corrected. Examples include:
  o In “Getting a camera with a greater number of megapixels means cropping and enlarging won’t adversely affect picture quality”, the gerund “getting” actually does not mean that you are actually already getting the camera at the moment. This would be the meaning for the gerund. Instead, it means the hypothetical action of the infinitive, something like “To get a camera… means cropping… won’t affect…”. This infinitive is the tense that is required to appear in the translation. The MT will translate as a gerund and the post-editing needs to change to an infinitive. If this does not happen, there is a bias.
• English has almost no difference between the subjunctive mode and the indicative. The construction “If I were invisible” instead of “If I was invisible” may be the most visible instance of differences in the mode. Yet, there is a popular song that uses the “was” form. This similarity will cause the MT to make errors translating subjunctives as indicatives. If this is not corrected by post-editing, there is a bias. Examples include:
  o “so that they can be lifted” in “the rockets are fitted with magnets so that they can be lifted and loaded with cranes”
  o “(would) freeze herself” and “would cause” in “It seems as if Hilda, while trying to scare up a dancing partner, accidentally freezes herself and causes objects to fly throughout the house”
  o “to be found” in “Usually the puzzles require items to be found and then executed”. It means “the puzzle requires that items be found”, and not “requires items that will be found”.
• Spelling (including language rules) – Brazilian Portuguese had a spelling reform. Therefore, there are new language rules in place. The corpora used to train the MT engine contains content created before the reform. Therefore, there are spelling errors training the MT, and they will appear in the translation. If they are not corrected, there will be bias.

Spelling reforms and corpora for MT will pose a certain challenge for MT systems. Many languages use corpora from different flavors of the language, such as European Portuguese and Spanish versus those used for Brazil and Latin America. The spelling of Portuguese varies a little bit between Brazil and Portugal, so the MT ends up making a few Portugal suggestions for Brazil and Spain suggestions for Mexico or Colombia. The advantages of having more corpora by doing this outweigh the downsides of it. If the corpora will not be fixed, the role of post-editing will include correcting these “cross-border” spelling issues.

**Errors per 1k words**

The main number that we obtained was the number of bias errors per 1k words. Although there is no official standard for the industry, we typically consider translations as high quality if the
number of errors per 1,000 words does not exceed 2, based on our own personal experience. The total number that we found in this work is given in Table 2.

<table>
<thead>
<tr>
<th>Total Number of words</th>
<th>14,986</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Number of bias errors</td>
<td>270</td>
</tr>
<tr>
<td>Errors per 1k words</td>
<td>18.02</td>
</tr>
</tbody>
</table>

Table 2. Number of bias error vs. number of words in post-editing

This number is about nine times a reference for quality, indicating that the total number of bias errors is significant. We should keep in mind that these are only bias errors. In addition to these, there are regular non-bias errors, where the post-editor makes changes and they are still not correct. The entire picture of quality is comprised of bias + non-bias errors.

### Numbers for each type of error

<table>
<thead>
<tr>
<th>Polysemy</th>
<th>Mis-translation</th>
<th>Multiple Modifiers or Words</th>
<th>Multiword Expressions</th>
<th>Omission initial article</th>
<th>Do Not Translate terms</th>
<th>Untranslated</th>
<th>Omission</th>
<th>Addition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of errors</td>
<td>54</td>
<td>56</td>
<td>22</td>
<td>14</td>
<td>10</td>
<td>9</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>Errors per 1k words</td>
<td>3.60</td>
<td>3.74</td>
<td>1.47</td>
<td>0.93</td>
<td>0.67</td>
<td>0.60</td>
<td>1.00</td>
<td>1.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Terminology</th>
<th>Gender agreement</th>
<th>Number agreement</th>
<th>Prepositions</th>
<th>Word order</th>
<th>Spellings (incl Lang Rules)</th>
<th>Grammar Verb tense</th>
<th>Paraphrasing</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of errors</td>
<td>14</td>
<td>7</td>
<td>12</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Errors per 1k words</td>
<td>0.93</td>
<td>0.47</td>
<td>0.80</td>
<td>0.53</td>
<td>0.53</td>
<td>0.27</td>
<td>0.60</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Table 3. Numbers for each type of error

Table 3 lists the frequency of each of the error types which we defined during our analysis. The breakdown per type shows that several types of causes of errors (described previously) are deserving of further attention:

- Polysems
- Multiple modifiers or words
- Multiword expressions
- Terminology
- Omissions
- Untranslated words
Other causes of mistranslation

**Detailed Analysis for Errors caused by Multiple Modifiers or Words (MMoW)**

We analyzed the error caused by Multiple Modifiers or Words in more detail. We found 22 instances of this error, indicating about 1.5 errors per 1k words. This is a significant number for just one type of error.

Next, we were interested in the question whether this error occurred with the same frequency for different types of MT systems. Therefore, we compared the output from 2 different types of MT systems for these 22 segments to find out whether these errors stem from inherent complexity of the source segment. We used this issue to have a sense of the impact that Neural Machine Translation may have on the quality. The hypothesis is that if NMT produces an output that contains less causes of errors, there will be less errors and therefore less bias. We wanted to see if this happened.

We looked into the 22 issues identified originally on Microsoft Statistical MT and created a NMT output from Microsoft Neural MT for them. We then evaluated to see how many of the original 22 errors were still present in the NMT output. We found that 10 out of 22 times, the NMT system corrected the error, meaning that it improved over the SMT system in 45% of the cases. However, in the remaining 10 segments, we still observed the same type of error in the NMT output. These results indicate that NMT tends to produce less errors than SMT for the post-editing corrections. This leads to better post-editing quality. However, 55% of the errors in the SMT output were still present in the NMT, indicating that the issues that a significant portion of the issues that are difficult for SMT remain an issue for NMT. This seems to indicate that the work on patterns that we started here would be a worth pursuit in improving NMT, and in evaluating it.

Once we identify that Multiple Modifiers or Words was an issue worth our attention, we thought of how we could find these expressions in a more semi-automated way. The process that we used can be described in these general steps:

1. Run a POS tagging of the source content
2. List tokens and tags and simplify the POS tags to a minimum; see Table 4 for examples. Create patterns indicating errors and find these patterns in the tagged content

We applied a simple formula to find a pattern: adjective-noun-noun and found the example “enhanced telephony capability” above. We also looked for another pattern: adjective or noun-noun-“and”-noun. We found examples such as “cook time and temperature” with this pattern. Analyzing the data, we found that:

- Using only these two narrow formulas we already found 7 out of 22 issues (32%).

This indicates that a few formulas could find the majority of the patterns.

<table>
<thead>
<tr>
<th>Token</th>
<th>Tag</th>
<th>Simplified Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>DT,B-NP-plural</td>
<td>DT</td>
</tr>
<tr>
<td>enhanced</td>
<td>JJ,enhance/VBD,enhance/VBN,I-NP-plural</td>
<td>JJ</td>
</tr>
<tr>
<td>telephony</td>
<td>NN,U,I-NP-plural</td>
<td>NN</td>
</tr>
<tr>
<td>capability</td>
<td>NNS,E-NP-plural</td>
<td>NN</td>
</tr>
</tbody>
</table>

Table 4. Examples of POS tags and simplified tags for English tokens
We manually analyzed the 22 MMoW issues to find out how many were suitable to be found with formulas/patterns. Out of 22, 20 of them could be found. This indicates that most of the MMoW issues are findable with patterns, and that there is potential to semi-automate the harvesting of these terms from a content tagged with POS.

3. Conclusions

1. We found significant bias in the post-editing of MT. This cannot be generalized to all cases, but it shows that the bias exists and is an issue to be considered as part of improving post-editing.
2. We found patterns that cause MT errors and can cause significant bias. These patterns should be considered for improvement of post-editing and for measurement of post-editing quality.
3. We found that it is possible to apply some automation in detecting the error patterns that cause errors on MT.
4. We found that Neural MT is likely to reduce the errors from bias by eliminating the original MT error. However, a significant percentage of the issues that cause errors on MT are not resolved by Neural MT and remain of interest for improving and measuring the quality of Neural MT.

4. Future Work

1. The semi-automated finding of patterns should be explored further. Once a representative number of instances of patterns is obtained, different metrics can be calculated. For example, we could find that there are 100 instances of MMoW in the content. If, upon reviewing them, we find, for example, 43 errors, this indicates that this type of error is produced by MT 43% of the time.
2. We think that there is potential in measuring the quality of the MT output based on difficult issues instead of a random sample. If a system 1 performs better than another system 2 on polysemous words, multiple modifiers or words, and multiword expression, it is likely that this system 1 will perform better on any translation than system 2. The same reasoning of measuring difficult words can be applied to measuring post-editing quality. We would like to create a measurement method that is not based on random sampling nor error typology, that targets difficult words, that is not subjective (make simple binary decisions), that is fast, cost-effective and suitable for crowdsourcing (with bilingual people and not professional linguists). We are working on this topic.

References


Terminology post-editing of neural MT by UTX glossary data

MT Summit 2017

YAMAMOTO Yuji
UTX team leader, AAMT
http://www.aamt.info/english/utx/

Presenter: Yamamoto Yuji

- CosmosHouse Founder/Representative
- Language/translation consultant
- AAMT UTX team leader
- ISO/TC37 (terminology) committee member
- Contact at http://cosmoshouse.com/mail.htm
Agenda

1. Background – terminology and NMT
2. UTX – a structured glossary format
3. Terminology post-editing
4. Conclusion

Background – terminology and NMT
Terminology is an essential part of systematic translation

Commercial translation requirements

- Glossaries are established by client companies. e.g. Microsoft glossaries
- Use of a company vocabulary is not optional. You are required to use certain terms for translation.
NMT problems

1. NMT mistranslates low-frequency words
2. NMT cannot reflect an existing glossary
3. NMT lacks terminological consistency

Problem 1: NMT mistranslates less-frequent terms

- such as proper names and technical terms
- e.g. auxiliary verb → *補助動詞 (助動詞 is correct)
- response rate → *奏功率 (回答率 is correct)
- Missing or repeated translation
Problem 2: NMT cannot reflect an existing glossary

- e.g. “liaison” in ISO context
- A glossary is not an issue for general MT users
- A glossary is essential in a systematic translation
- Many companies are not managing glossaries in an organized manner
- Translation problems are hidden in such an environment

Problem 3: NMT lacks terminological consistency

- e.g. International Standard→国際規格、国際標準
  - resource→資源、リソース
- Terminology consistency is not an issue for general MT users
- But terminology consistency is important in systematic translation
Prevalence of RbMT in Japan

- Strong demands for translation.
- EN-JA bilingual market.
- Early MT commercialization since 1990s.
- Many commercial RbMT packages are sold.

Japanese is an influential language, but its market is bilingual

![Japanese language market share](https://journal.lib.uoguelph.ca/index.php/perj/article/view/826/1358#WY_eh1GrSHs)

125 million speakers

Top languages in global information production
Sergey Lobachev
Casual Reference Librarian
London Public Library
Bilingual or multilingual scenario?

- Japan – Japanese and English
- Europe, Americas, Africa, etc. - multilingual

Terminology management must be simplified

- Or it will never be implemented.
- Multilingual complexity is not necessary for a bilingual environment.
- A simple Excel sheet is too simple.
UTX – a structured glossary format

What is UTX (Universal Terminological eXchange)?

Simple but structured glossary data format

for terminology tools and MT
UTX is free to use

UTX specification is free
- Many UTX glossaries are free
- UTX Converter is free
  - (open source)

UTX facilitates sharing and reusing of glossaries

Translation Client

Language Service Provider

Translator A

Translator B

UTX glossaries
Conversion to/from UTX

Excel

MT dictionaries

Termbases

UTX Converter

UTX glossary sample
Manage essential glossary data in a standardized format

Information about the glossary (creation date, license, etc.)

<table>
<thead>
<tr>
<th>#src</th>
<th>tgt</th>
<th>src:pos</th>
<th>term status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia-Pacific Association for Machine Translation</td>
<td>亚洲太平洋机器翻译协会</td>
<td>properNoun</td>
<td>approved</td>
</tr>
<tr>
<td>dictionary administrator</td>
<td>字典管理员</td>
<td>noun</td>
<td>approved</td>
</tr>
<tr>
<td>contributor</td>
<td>用语提交者</td>
<td>noun</td>
<td>provisional</td>
</tr>
<tr>
<td>domain</td>
<td>领域</td>
<td>noun</td>
<td></td>
</tr>
<tr>
<td>glossary</td>
<td>词汇表</td>
<td>noun</td>
<td></td>
</tr>
<tr>
<td>bidirectional</td>
<td>双向</td>
<td>adjective</td>
<td>approved</td>
</tr>
<tr>
<td>merge</td>
<td>合并</td>
<td>verb</td>
<td>approved</td>
</tr>
</tbody>
</table>

Source term (American English)  Target term (Chinese)  Part of speech  Term status

Term status provides reliability
JPO (Japan Patent Office) UTX glossary

- Created by JPO, converted by AAMT
- Available for free
- Japanese to English
- 130 thousands entries
- Only rare technical terms
  - User-defined terms not included in technical dictionaries shipped with the package

JPO glossary covers all IPC sections

IPC (International Patent Classification)
- A human necessities
- B performing operations; transporting
- C chemistry; metallurgy
- D textiles; paper
- E fixed constructions
- F mechanical engineering; lighting; heating; weapons; blasting
- G physics
- H electricity
## Examples of entries

<table>
<thead>
<tr>
<th>Domain (semantic feature)</th>
<th>Entries</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plants (common names, species, scientific names, etc.)</td>
<td>5498</td>
<td>白いぼキュウリ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>メラレウカ・アルテルフォリア</td>
</tr>
<tr>
<td></td>
<td></td>
<td>いらくさ科植物</td>
</tr>
<tr>
<td>Animals (common names, scientific names, etc.)</td>
<td>3025</td>
<td>ヤブカ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>モンシロチョウ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ユーグレナ</td>
</tr>
<tr>
<td>People (personal names, titles, etc.)</td>
<td>1316</td>
<td>昌聰</td>
</tr>
<tr>
<td></td>
<td></td>
<td>調香士</td>
</tr>
<tr>
<td></td>
<td></td>
<td>登壇者</td>
</tr>
<tr>
<td>Companies and organizations</td>
<td>7340</td>
<td>日本釀造協会</td>
</tr>
<tr>
<td></td>
<td></td>
<td>獵友会</td>
</tr>
<tr>
<td></td>
<td></td>
<td>インド技術研究所</td>
</tr>
<tr>
<td>Others</td>
<td>46975</td>
<td>Chemistry, medicine, machine, engineering, and other technical terms.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>オキシジフタル酸ニ無水物</td>
</tr>
</tbody>
</table>

## Patent documents characteristics

1. Extremely long sentences
2. Ambiguous sentence structure
3. Peculiar writing style
4. Many technical terms (obfuscation)
Terminology post-editing

What is “terminology post-editing”?

- post-editing method focused on terminology checking
- requires structured glossary data that has **strong correlation** with the source documents
Terminology post-editing: merits and limitations

- **Merits**
  - Fully- or partially-automated check
  - Check with no lingual knowledge

- **Limitations**
  - Accuracy is insufficient (requires other criteria for a full quality assessment)

Quick check or post-editing

Terminology check in SDL Trados

![Terminology check in SDL Trados](image)
Patent NMT translation checked by UTX glossary data (SDL Trados)

Patent NMT translation checked by UTX glossary data (Memsource)
Patent NMT translation checked by UTX glossary data (ApSIC Xbench)

Result: potential term errors

391 segments (sentences). More error detection is not necessarily better.

<table>
<thead>
<tr>
<th>Detected potential term errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDL Trados</td>
</tr>
<tr>
<td>Memsource</td>
</tr>
<tr>
<td>ApSIC Xbench</td>
</tr>
</tbody>
</table>

Examples of incorrectly translated terms:

- 請求項/claim
- デシテックス/decitex
- 経編/warp weaving
- センサシステム/sensor system
Patent NMT translation checked by UTX glossary data (ApSIC Xbench)

Result: potential term errors

391 segments (sentences). More error detection is not necessarily better.

<table>
<thead>
<tr>
<th>Detected potential term errors</th>
<th>SDL Trados</th>
<th>Memsource</th>
<th>ApSIC Xbench</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1237</td>
<td>372</td>
<td>603</td>
</tr>
</tbody>
</table>

- Examples of incorrectly translated terms:
  - 請求項/claim
  - デシテックス/decitex
  - 経編/warp weaving
  - センサシステム/sensor system
Conclusion

1. NMT has many terminological flaws.
2. Glossary data and terminological check can find potential term errors.
3. To do so, you need a simple but structured glossary data format (such as UTX).
4. The UTX format was proved to be effective in finding potential errors.

More info

- Search for “UTX glossary”
- Contact at [http://cosmoshouse.com/mail.htm](http://cosmoshouse.com/mail.htm)
- We welcome your feedback!
Harvesting Polysemous Terms from e-commerce Data to Enhance QA

Silvio Picinini
Localization, eBay Inc., San Jose, CA, USA
spicinini@ebay.com

Abstract

Polysemous words can be difficult to translate and can affect the quality of Machine Translation (MT) output. Once the MT quality is affected, it has a direct impact on post-editing and on human-assisted machine translation. The presence of these terms increases the risk of errors. We think that these important words can be used to improve and to measure quality of translations. We present three methods for finding these words from e-commerce data, based on Named Entity Recognition, Part of Speech and Search Queries.

1. Introduction

Polysemous are words or sets of words with multiple meanings. For this work, we consider a broader definition that reflects what polysemy causes in Machine Translation (MT). A polysemous word here is “a word that can have different translations”. This means that the MT engine can be confused about which translation is the correct one; this in turn affects the quality of the machine translation. Instead of “polysemous”, we could call these “polytranslation” terms, and just invent this word.

For example, a word that can be assigned multiple POS (parts-of-speech) tags may have a similar meaning, but it will be translated differently if it is a noun, a verb, or an adjective. Example: Print a report (verb), print magazine (adjective), this fabric has a nice print (noun).

A brand that is also a common word (e.g. Gap, Guess, Coach) will be left untranslated when referring to a brand and will be translated when used as a common word.

Also, a word like “mixer” may have a generic meaning of “a device that mixes”, but in the real world, it can refer to very different products, such as a kitchen mixer or a sound mixer for music. These are two very different devices with very different translations. Also, it can be a party (singles mixer), a very different meaning.

This work presents three new processes that leverage eBay e-commerce data to harvest polysemous words, so that these can be used for different applications, such as the ones described in this paper.

Before going into the methods, we make two general points about data below.

2. Leveraging semantic value and relevance added by the public

We think that it is important to make a point about the importance of capturing semantic meaning from user behavior. It is a massive and no-cost source of information, therefore, we should be interested in using it. One of the methods described uses information from buyer behavior on eBay. By entering a query and then going into a certain category, the buyer is associating meaning to the query, which is comprised of one or two words. The same happens when a seller describes an item for sale and chooses a category for it, giving the words of the
item additional context and meaning. It is also important to capture relevance from user behavior. If we capture how frequent certain meanings are, we are capturing how relevant they are. This “quantification” is something that traditional dictionaries cannot do, and is a significant difference compared to dictionaries.

All of this can be seen as a “public semantic annotation” work that is being done without cost for the companies. While companies collect vast amounts of data, we are all constantly looking for ways to enhance the meaning of this data, and the examples in this work are in line with that overall effort.

3. Harvesting relevant words in context

Another important point about polysemous words and synonyms is these words are relevant because of their meaning relationship; polysemous words have a single form with multiple meanings and synonyms have multiple forms with a single meaning. There are also other types of relationships that can be of interest, such as hyponyms, hypernyms, and meronyms (a word that is a constituent part or a member of another word).

These words are useful in many ways, such as improving and measuring MT, but also improve search queries and possibly classification. The challenge is to find “applied” examples of these words in a specific context. While a dictionary or WordNet can tell us that words are synonyms, it will not tell us that “camcorder” is a synonym for “video camera” or that “flash drive” is the same as a “pendrive”. Finding these words applied to a context, an industry, or a subject matter should be more useful than generic words.

4. Applications of polysemous words

There are some possible applications for polysemous words in machine translation:

- Select data containing these words and create training and testing data for them

Since these words are more likely to be mistranslated, they are more likely to require more in-context training data to help the engine disambiguate different situations. So one application of polysemous words is to find examples of content with these words and have it translated/post-edited. This will allow the creation of training data for the engine to learn how to better handle these words, and the creation of testing data to evaluate the translation.

- Evaluate the quality of the MT of these words

The evaluation of the MT output quality is usually performed on the entire content (using automated metrics) or on a sample (if human evaluation is used). In both cases, there is usually no “selection” of certain segments matching some certain criteria which should be measured, the segments are randomly chosen. However, polysemous words could be used to provide an insight on the quality of the machine translation, using selected “more difficult” words. eBay has started collecting some data around this.

- Evaluate the quality of the post-editing of these words

Training and testing data may be created through post-editing. The quality of that post-editing work needs to be evaluated. Polysemous words are more likely to be mistranslated and
be wrong in the MT output. The post-editing process is supposed to correct those errors. If the error is corrected by the post-editor, this is an indication that the post-editing is of good quality. If the error is not corrected, this is an indication that the post-editing may not be of good quality, and may need further work before being used as training or testing data. The evaluation of post-editing is usually done by an evaluator on a random sample.

Looking at how polysemous words were post-edited is a way to assess the quality of the post-editing work and is also an indication of the final quality of the content that is going to become training or testing data.

5. Three processes to harvest polysemous words

5.1. From eBay search queries

This process is based on associating different categories to the same query. The premise is that if a word is associated with two very different categories, they are likely to have very different meanings, and there is a good chance that the word is polysemous.

Customers enter search queries on eBay. After seeing the results of their queries, they take an action that leads to a certain category. This is an indication of the meaning of the word that was entered as a query. Let’s consider an example with the word “mixer”. A query for that word does not clarify if the customer is looking for a sound mixer or a kitchen mixer. However, after the query display results, the customer takes action to look into one of these different devices. Once the customer acts, there is now a category that can be attached to the word in the query.

eBay creates a column called Leaf Category Histogram. It looks like Figure 1:

![Leaf Category Histogram](image)

This column contains the identification of the most frequent categories (in black above) accessed by the customer after entering a query. It also contains the % of instances where the customer went to a certain category (in red above). This number is an indication of the “intensity of the polysemy”. If a word like “mixer” goes 60% of the time to a music category (for sound mixer) and 40% to a kitchen category, this is an indication that there is significant interest for both meanings. If another word has a 99.8% frequency and the second category is, for example, less than 0.1%, this is an indication that one meaning is nearly universal and the other is extremely rare. This can inform our harvesting of polysemous words.

Starting from that data, we find the higher level eBay categories associated with the word. We are interested in finding big differences in categories, which would be more likely to have different meanings. Once we manipulate the data, we arrive at information that looks like Table 1:

```
<table>
<thead>
<tr>
<th>Word</th>
<th>Category 1</th>
<th>Frequency 1</th>
<th>Category 2</th>
<th>Frequency 2</th>
<th>Is Cat 1 diff from 2?</th>
<th>Is Freq 2 &gt; 2%?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixer</td>
<td>Music</td>
<td>60%</td>
<td>Kitchen</td>
<td>40%</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
```

Table 1. Data after manipulation
The last two columns are formulas. With this data we can filter the last two columns and the result will be a list of polysemous candidates. Table 2 show some examples we found in our initial results:

<table>
<thead>
<tr>
<th>Word</th>
<th>Category 1</th>
<th>Category 2</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vans</td>
<td>Clothing, Shoes &amp; Accessories</td>
<td>eBay Motors</td>
<td>brand vs. type of car</td>
</tr>
<tr>
<td>notebook</td>
<td>Computers/Tablets &amp; Networking</td>
<td>Books</td>
<td>computer vs. writing</td>
</tr>
<tr>
<td>Fossil</td>
<td>Jewelry &amp; Watches</td>
<td>Collectibles</td>
<td>brand vs. actual fossil</td>
</tr>
<tr>
<td>mixer</td>
<td>Musical Instruments &amp; Gear</td>
<td>Home &amp; Garden</td>
<td>sound table vs. dough mixer</td>
</tr>
<tr>
<td>roadrunner</td>
<td>eBay Motors</td>
<td>Toys &amp; Hobbies</td>
<td>car vs. character</td>
</tr>
<tr>
<td>Pebble</td>
<td>Cell Phones &amp; Accessories</td>
<td>Pet Supplies</td>
<td>brand of watches</td>
</tr>
<tr>
<td>torch</td>
<td>Sporting Goods</td>
<td>Business &amp; Industrial</td>
<td>flashlight vs. hot flame</td>
</tr>
</tbody>
</table>

Table 2. Results from queries

A quick human triage to validate which of these candidates are good produces our final list.

The initial results indicate that this process is efficient. A list of about 1900 queries yielded about 40 candidates. A human triage that took about an hour yielded 19 final terms, about 1% of the initial data.

5.2. From NER data

For Named Entity Recognition, we tag individual tokens, mapping them to different tags according to their meaning. The premise for finding polysemous words in this process is that the same word can be tagged with different tags, and if these tags indicate a significantly different meaning, there is a good chance that the word is polysemous.

This process leans on the concept of polysemous words being defined by “how words are translated”. The most benefit from this process comes from differentiating words that are not translated from words that are translated. The MT engine may be confused and translate brand names, or do not translate common words because they are commonly brand names. The word “charger” can refer to the car Dodge Charger. This is a product name and won’t be translated. But it can also refer to a charger for a cell phone. This is a common word and will be translated. Therefore, it is possible that there is “a charger in a Charger”, and the MT has to deal with this ambiguity.

We start with a list of tokens and tags for a certain category. Once we sort it by token, we will see that some tokens are tagged with different tags. Some NER tags indicate that the token should not be translated: Brand and Product Name. Other tags indicate that the meaning tends to be a common word: Type, Color. We organize the data with additional columns: Do Not Translate indicates when a token is tagged with Brand or Product Name. Translatable indicates when the token is tagged with a category that is usually a common word, and therefore translatable. Once the data is organized in this way, a few manipulations with sorting, filtering and formulas will produce the list of candidates that we are looking for.
Table 3 shows what the data looks like:

<table>
<thead>
<tr>
<th>Word</th>
<th>Token</th>
<th>Do Not Translate?</th>
<th>Translatable?</th>
<th>Contains Translatable and DNT?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Charger</td>
<td>t</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charger</td>
<td>p</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 3. Data after manipulation

Table 4 shows some of our initial results:

<table>
<thead>
<tr>
<th>Token</th>
<th>Contains DNT tag?</th>
<th>Contains translatable tag?</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>b</td>
<td>c</td>
<td>Black and Decker brand vs. color</td>
</tr>
<tr>
<td>Case</td>
<td>b</td>
<td>n</td>
<td>Case Logic brand</td>
</tr>
<tr>
<td>Charger</td>
<td>md</td>
<td>ta</td>
<td>Dodge Charger car vs. device</td>
</tr>
<tr>
<td>RAM</td>
<td>md</td>
<td>ta</td>
<td>Dodge RAM pickup vs. RAM memory</td>
</tr>
<tr>
<td>Range</td>
<td>md</td>
<td>f</td>
<td>Range Rover brand vs. common word</td>
</tr>
<tr>
<td>Seat</td>
<td>ma</td>
<td>n</td>
<td>Car maker in Spain</td>
</tr>
</tbody>
</table>

Table 4. Results from NER

5.3. From Part of Speech data (POS)

This process is based on identifying when a word is used with different parts of speech in a certain content. It is very common for MT engines to make errors because of a word that is written in the same way, but can be a verb, a noun, or an adjective for example. While the English language does not have any difference for the usage of that word, other languages will have lots of variations for the different POS. Adjectives will have gender in Romance languages, and verbs will have a variety of forms. This brings again the concept that “translations will be different for the same word”, and this may confuse the MT engine and affect the MT quality.

The premise for this process is that if a word is associated with two different POS types, there is a good chance that the word is polysemous (will have different translations).

We run a POS tagger on the content, and the result looks like this:

```
```
With some manipulation, we create a list with two columns: word and POS tag. We sort that list by word, and secondarily by tag, and then we are on our way to identify words that have more than one POS tag, as shown in Table 5:

<table>
<thead>
<tr>
<th>Word</th>
<th>Tag</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessory</td>
<td>J</td>
<td></td>
</tr>
<tr>
<td>Accessory</td>
<td>J</td>
<td></td>
</tr>
<tr>
<td>Accessory</td>
<td>N</td>
<td>accessory tagged as N (noun) and J (adjective)</td>
</tr>
</tbody>
</table>

Table 5. Data after POS tagging and manipulation

Different POS taggers will have different tags, but this process only requires:
- Creating a vertical list of words and tags (usually simple introduction of CR characters)
- Identifying a different part of the tag for nouns, adjectives and verbs (sometimes the first letter of the tag will be enough, as above)

We can also subtotal the list by words and tag, and we will then have information about the frequency that each word and tag occurs. This number indicates the candidates with better potential. One word may have a 60%/40% ratio between noun and verb, while another word may have a 99%/1% ratio. If the same proportion appears in the training data, the first situation will more likely confuse the MT engine than the second situation.

Table 6 below show some of our initial results:

<table>
<thead>
<tr>
<th>Word</th>
<th>Tag</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessory</td>
<td>N</td>
<td>accessory tagged as N and J</td>
</tr>
<tr>
<td>Acted</td>
<td>V</td>
<td>acted tagged as V and J</td>
</tr>
<tr>
<td>Adapted</td>
<td>V</td>
<td>adapted tagged as V and J</td>
</tr>
<tr>
<td>Added</td>
<td>V</td>
<td>added tagged as V and J</td>
</tr>
<tr>
<td>Adhesive</td>
<td>N</td>
<td>adhesive tagged as N and J</td>
</tr>
<tr>
<td>Adjusted</td>
<td>V</td>
<td>adjusted tagged as V and J</td>
</tr>
<tr>
<td>Adore</td>
<td>V</td>
<td>adore tagged as V and N</td>
</tr>
<tr>
<td>Affected</td>
<td>V</td>
<td>affected tagged as V and J</td>
</tr>
</tbody>
</table>

Table 6. Results from POS

6. **Quantification effect enhances relevance**

The processes presented have a “quantification” effect on the meaning. A term could be polysemous and one of the meaning could be very rare. In practical terms, this would not be a significant polysemy case, because there is no volume for that meaning. The eBay data helps indicating how often a term has one meaning versus another, by connecting the meaning to a frequency number.

In queries, the frequency is defined by the category that follows the term. In NER, the frequency indicates how often each meaning appears in one category, but we can also look across categories. In POS, this effect also appears. In absolute terms, a certain word can be
tagged with several parts of speech. However, one of these POS may be very rare in the context being analyzed, so this variation would not appear in the results.

These are positive effects, because they introduce the frequency/relevance into the analysis and results, as opposed to an analysis based just on the absolute existence of multiple meanings or POS in a dictionary.

7. Conclusion

The processes described here are finding words with limited human effort, indicating that they are efficient. These words are valuable for eBay because they take into account the eBay context. For example, Fossil is a noun and a brand, but a dictionary would not contain the brand. So these processes are finding words in a way that could be difficult to find with other resources. There is also value in the “quantification” of how frequent these words are.

The methods for harvesting polysemous words presented here are only possible due to the wealth of linguistic data that eBay has. We hope that other companies that have data will find these ideas useful, and those who do not have data will feel inspired to create data and use it.
Translation Dictation vs. Post-editing with Cloud-based Voice Recognition: A Pilot Experiment

Julián Zapata, julianz@intr.co
University of Ottawa & InTr Technologies, Ottawa-Gatineau, Canada
Sheila Castilho, sheila.castilho@adaptcentre.ie
ADAPT Centre/School of Computing, Dublin City University, Dublin, Ireland
Joss Moorkens, joss.moorkens@adaptcentre.ie
ADAPT Centre/School of Applied Language & Intercultural Studies, Dublin City University, Dublin, Ireland

Abstract
In this paper, we report on a pilot mixed-methods experiment investigating the effects on productivity and on the translator experience of integrating machine translation (MT) post-editing (PE) with voice recognition (VR) and translation dictation (TD). The experiment was performed with a sample of native Spanish participants. In the quantitative phase of the experiment, they performed four tasks under four different conditions, namely (1) conventional TD; (2) PE in dictation mode; (3) TD with VR; and (4) PE with VR (PEVR). In the follow-on qualitative phase, the participants filled out an online survey, providing details of their perceptions of the task and of PEVR in general. Our results suggest that PEVR may be a usable way to add MT to a translation workflow, with some caveats. When asked about their experience with the tasks, our participants preferred translation without the ‘constraint’ of MT, though the quantitative results show that PE tasks were generally more efficient. This paper provides a brief overview of past work exploring VR for from-scratch translation and PE purposes, describes our pilot experiment in detail, presents an overview and analysis of the data collected, and outlines avenues for future work.

1. Introduction
Machine translation (MT) post-editing (PE) and voice recognition (VR) technology are gaining ground in both translation technology research and the translation industry. Over 50% of international Language Service Providers now offer a PE service using dedicated MT engines integrated into translators’ computer-aided translation environments (Lommel and DePalma, 2016). In a recent survey of 586 translators in the UK, 15% responded that they use VR technology in their work (Chartered Institute of Linguists et al., 2017). These disparate technologies tend not to be deployed in tandem, although both offer translators the potential to increase productivity and reduce the technical effort usually required to translate from scratch when using conventional word-processing hardware and software.

We carried out a pilot experiment to investigate the effects on productivity and on the translator experience (TX) (Zapata, 2016a) of integrating PE with VR and translation dictation (TD) using a sequential mixed-methods design. In the quantitative phase, four
translators performed four translation tasks under four different conditions: (1) conventional TD (i.e., sight-translating using a digital dictaphone), (2) PE in dictation mode (PED) (i.e., dictating approved or amended segments into the same dictaphone), (3) TD with VR (TDVR) (using a cloud-based VR system on a tablet), and (4) PE with VR (PEVR) (using the same VR system as in task 3). The quantitative experiments consisted of three phases during which task times were measured and some input data were collected. Phase I consisted of dictating and post-editing with dictaphone or the VR system; phase II consisted of manually transcribing the recordings from tasks 1 and 2 on the researcher’s laptop; and phase III consisted of revising/editing all four translations. As has been noted in a great deal of research about PE, productivity increases alone do not make a tool desirable for translators (see Teixeira, 2014; Moorkens and O’Brien, 2017). Translator attitudes and usability, the TX, are important factors in the adoption of any technology. For this reason, we have appended a follow-on qualitative phase, wherein the participants filled out an online survey, providing details of their perceptions of the task and of PEVR in general.

In this paper, we present our pilot experiment in detail. The paper is structured as follows: First, we provide a brief overview of past work exploring VR for from-scratch translation and PE purposes. Then, we describe the experimental setup, and present an overview and analysis of the quantitative and qualitative results. In the conclusion, we describe avenues for future work.

2. Related Work

2.1. TD and VR

The idea of using human voice to interact with computers and process texts is as old as the idea of computers themselves. For decades, and in recent years more than ever before, voice input has been widely used in a vast array of domains and applications, from virtual assistants on mobile phones to automated telephone customer services; from professional translation to legal and clinical documentation.

Simply put, VR (also known as voice/speech-to-text or automatic speech recognition) technology recognizes human-voice signals and converts them into digital data. The earliest experiments in VR suggested that voice input was expected to replace other input modes such as the keyboard and the mouse in full natural language communication tasks. However, it was soon discovered that speech often performed better in combination with other input modes such as the keyboard itself, as well as touch, stylus and gesture input on multimodal interfaces (Bolt, 1980; Pausch and Leatherby, 1991; Oviatt, 2012).

In translation, there has been a long interest in speaking translations instead of typing them. First, in the 1960s and 1970s professional translators often collaborated with transcriptionists, and dictated their translations either directly to the transcriptionist or into a voice recorder (or dictaphone), before having them transcribed later (a technique often referred to as TD). In the 1990s and 2000s, researchers began to explore VR adaptation for TD purposes. Such developments focused mainly on reducing VR word error rates by combining VR and MT. Hybrid VR/MT systems are presented with the source text and use MT probabilistic models to improve recognition; translators simply dictate their translation from scratch without being presented with the MT output (Brousseau et al., 1995; Désilets et al., 2008; Dymetman et al., 1994; Reddy and Rose, 2010; Rodriguez et al., 2012; Vidal et al., 2006). More recently, further efforts have been made to evaluate the performance of translation students and professionals when using commercial VR systems for straight TD (Dragsted et al., 2009; Dragsted et al., 2011; Mees et al., 2013); to assess and analyze
professional translators’ needs and opinions about VR technology (Ciobanu, 2014 and 2016; Zapata, 2012), and to explore TD in mobile and multimodal environments (Zapata and Kirkedal, 2015; Zapata, 2016a,b).

2.2. PE and VR

In recent years, the potential of using VR for PE purposes has also been investigated (García-Martínez et al., 2014; Mesa-Lao, 2014; Torres-Hostench et al., 2017). García-Martínez and her collaborators (2014) tested a VR system integrated into a PE environment (both research-level cloud-based systems). They argue that voice input is more interesting than the keyboard alone in a PE environment, not only because some segments may need major changes and therefore could be dictated, but also because, if the post-editor is not a touch typist, the visual attention back and forth between source text, MT text and keyboard adds to the complexity of the PE task.

Mesa-Lao (2014) surveyed student translators, 80% of which (n=15) reported that they would welcome the integration of voice as one of the possible input modes for performing PE tasks. Thus, voice input offers a third dimension to the PE task, making it possible to combine different input modes or to alternate between them according to the difficulty of the task and to the changing conditions of human-computer interaction. Some experiments have also suggested specifically that for certain translators, text types and language combinations, the benefits of VR and PE integration may not be the same (e.g. in terms of efficiency, productivity and cognitive effort) (see Carl et al. 2016a and 2016b).

Tests with VR within a mobile PE app were reported, first by Moorkens et al. (2016), then by Torres-Hostench et al. (2017). Participants were impressed by VR quality and found it useful for long segments. However, they mostly preferred to use the keyboard due to limitations of the software for making minor edits to MT output.

In the following section, we describe our pilot experiment more in detail: our participants’ profile and our methodology.

3. Experimental Setup

3.1. Participants’ Profile

This experiment included a sample of native (Latin American) Spanish speakers. All four participants are either pursuing or have recently completed a doctoral degree in translation studies. Participants had in common at least a minimum level of acquaintance with the notions of MT, PE and VR. Our sample includes two men and two women between the ages of 26 and 43. Participants reported 3 to 12 years of translation experience, two have training in interpreting, and both of those are regular users of VR (and were therefore familiar with voice commands and other specificities related to dictating with VR). All participants reported to be occasional post-editors.

3.2. Methodology

For this study, we applied a sequential, explanatory mixed-methods design, using the follow-up explanations model, in which the qualitative data is intended to expand upon the quantitative results (Creswell and Plano Clark, 2007:72). We chose this methodology to answer the following two research questions:

1. Can PEVR be as or more productive than comparable approaches, with or without MT and VR?
2. Does the participants’ TX suggest that combining MT and VR is feasible for translation projects?

As mentioned in the introduction, four tasks were involved in the quantitative phase of this experiment, namely:

1) Conventional TD;
2) PED;
3) TDVR; and
4) PEVR.

A digital dictaphone was used for tasks 1 and 2. A commercial cloud-based speaker-independent VR system\(^1\) was used on an Android tablet for tasks 3 and 4. (See Zapata and Kirkedal (2015) for a description of the different approaches to VR technology with respect to users (i.e. speaker-dependent, speaker-adapted and speaker-independent systems)).

Source texts were 20-segment sections of newstest 2013 data used in WMT\(^2\) translation tasks. The test sets were analysed using the Wordsmith Wordlist\(^3\) tool to ensure that they were statistically similar, based on measurements for type/text ratio, average sentence length, and average word length. Table 1 shows the statistics of the test set.

<table>
<thead>
<tr>
<th>Text file</th>
<th>Type/token ratio (TTR)</th>
<th>Mean word length (in characters)</th>
<th>Word length std.dev.</th>
<th>Sentences</th>
<th>Mean (in words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Set 1</td>
<td>55.12</td>
<td>4.99</td>
<td>2.51</td>
<td>20</td>
<td>18.05</td>
</tr>
<tr>
<td>Test Set 2</td>
<td>55.73</td>
<td>4.80</td>
<td>2.63</td>
<td>20</td>
<td>19.65</td>
</tr>
<tr>
<td>Test Set 3</td>
<td>54.31</td>
<td>5.00</td>
<td>2.62</td>
<td>22</td>
<td>21.09</td>
</tr>
<tr>
<td>Test Set 4</td>
<td>54.20</td>
<td>5.18</td>
<td>2.69</td>
<td>20</td>
<td>17.25</td>
</tr>
</tbody>
</table>

Table 1. Test set statistics for source texts

A commercial-level MT system\(^4\) was used to translate the texts. All texts were printed out separately and presented to the participants in hard copy. Naturally, only in tasks 2 and 4 were participants presented with the segmented source and MT texts. The MT texts for tasks 1 and 3 were used only to calculate HTER scores (Snover et al., 2006); more details are provided in section 4.1.2.

Experiments were run individually (i.e. one participant at a time) over four days. A university study room was booked to perform the experiments.

Tasks were randomized as follows:

---

1\(^{1}\) Dragon Dictation, integrated in the Swype+Dragon app. See http://www.swype.com/.
2\(^{2}\) http://www.statmt.org/wmt13/
3\(^{3}\) http://lexically.net/wordsmith/
4\(^{4}\) Google Translate. See https://translate.google.com/.
Before performing any of the experimental tasks, participants were briefly instructed how to use the digital dictaphone (for tasks 1 and 2) and the VR system on the tablet (for tasks 3 and 4) (i.e., they were given the opportunity to dictate while testing a few voice commands such as punctuation marks, etc.).

The quantitative experiments consisted of three phases during which task times were measured and some input data were collected:

- Phase I - dictating and post-editing with dictaphone or the VR system on the tablet,
- Phase II - manually transcribing the recordings from tasks 1 and 2 (for TD and PED) on the researcher’s laptop; and
- Phase III - revising/editing all four translations on the researcher’s laptop.

It is important to highlight that during phase II, participants were instructed not to edit the translation, only transcribe what they heard. The documents in which dictations were performed on the tablet for tasks 3 and 4 in phase I were automatically saved into a cloud-based drive\(^5\) after dictation, and therefore immediately synchronized and available to be edited/revised on the researcher’s laptop in phase III.

In phase I, task times were measured using a stopwatch. In both phases II and III, Inputlog (Leijten and Van Waes, 2013) was used. Inputlog is a research-level program designed to log, analyse and visualize writing processes. The program provides data such as total time spent in the document, total time in active writing mode (i.e., of actual keystrokes), total time spent moving/clicking with the mouse, total number of characters typed, total switches between the keyboard and the mouse, etc. Beyond total task times alone, we were interested in collecting this kind of detailed input data, particularly for phase III. We are not reporting data other than task times here given the scope and limitations of this paper; we do consider, however, that input data analysis will be essential in larger-scale experiments.

Thereafter, in the qualitative phase, participants responded to a short online questionnaire, with socio-demographic questions, retrospective questions about the experiment, as well as questions providing insight on the TX with multimodal/mobile VR-enabled TD and PE applications (more details to be provided in section 4.4).

In the following section, some of the data collected is presented and analysed.

### 4. Results and Analysis

#### 4.1. Task Times Measures (Quantitative Phase)

In order to investigate the effects on productivity of integrating PE with VR and TD in the quantitative phase of this research, we have conducted analysis of the task times as follows:

1. Comparing tasks of the same nature with and without VR, that is, a) TD vs. TDVR (see 4.1), and b) PED vs. PEVR (see 4.2)

2. Comparing translation vs. PE within phases, that is: a) TD vs PED (4.3) and b) TDVR vs. PEVR (4.4).

We consider:

a) Translation and/or PE time (phase I + phase II), that is, the time participants needed to translate and/or post-edit, as well as the transcription time (for TD and PED);

b) Revision duration (phase III), that is, the total time participants needed to review/edit their translation/post-editing;

c) Total task time (phase I + phase II+ phase III), that is, the total time the participants needed to perform each task.

**TD versus TDVR**

When comparing both TD tasks (Table 3), i.e. the one performed with a dictaphone (TD) and the one performed with a VR program (TDVR), we can see that the total translation time is always shorter when participants use VR. A reminder to the reader that the total translation time in the dictaphone task includes the time participants need to transcribe their translations (phase II).

Regarding revision duration, however, tasks performed with VR seem to take longer to be completed. We speculate that this is because during the revision time, participants do not only review their translation but also must correct errors produced by the VR program.

<table>
<thead>
<tr>
<th>Participants</th>
<th>Task</th>
<th>Translation Time</th>
<th>Revision Time</th>
<th>Total Task Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Translation time</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transcription time</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ES1</td>
<td>TD</td>
<td>537</td>
<td>716</td>
<td>1253</td>
</tr>
<tr>
<td></td>
<td>TDVR</td>
<td>796</td>
<td>n/a</td>
<td>796</td>
</tr>
<tr>
<td>ES2</td>
<td>TD</td>
<td>688</td>
<td>1197</td>
<td>1885</td>
</tr>
<tr>
<td></td>
<td>TDVR</td>
<td>1330</td>
<td>n/a</td>
<td>1330</td>
</tr>
<tr>
<td>ES3</td>
<td>TD</td>
<td>846</td>
<td>1116</td>
<td>1962</td>
</tr>
<tr>
<td></td>
<td>TDVR</td>
<td>377</td>
<td>n/a</td>
<td>377</td>
</tr>
<tr>
<td>ES4</td>
<td>TD</td>
<td>700</td>
<td>1432</td>
<td>2132</td>
</tr>
<tr>
<td></td>
<td>TDVR</td>
<td>460</td>
<td>n/a</td>
<td>460</td>
</tr>
</tbody>
</table>

Table 3. TD vs TDVR (in seconds)

Overall, when considering all phases, total task time seems to be lower for TDVR, apart from participant ES2, who shows lower time when performing TD.
**PED versus PEVR**

Results for both PE tasks (PED and PEVR) were also compared (table 4). We notice that the PE time (total) is lower for all participants in the VR condition. As for revision, the time is higher in PEVR, which we assume is for the same reason described above: that participants also need to correct errors produced by the VR application. However, when considering all phases, participants were still faster post-editing with VR than with the dictaphone.

To compare how much PE was performed for each task, we have calculated the translation edit rate (HTER) (Snover et al. 2016). The HTER score is a measure that compares the raw MT output and the post-edited version, and goes from 0 to 1, where the higher number, the more modifications were made in the raw MT output. We can see in table 4 that most of the participants have an average score of 0.2 – which indicates that little post-editing was performed. However, participant ES3 displays more post-editing performed for the PED task (0.52).

<table>
<thead>
<tr>
<th>Participants</th>
<th>Task</th>
<th>PE Time</th>
<th>Revision Time</th>
<th>Total Task Time</th>
<th>HTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES1</td>
<td>PED</td>
<td>633</td>
<td>692</td>
<td>1325</td>
<td>238</td>
</tr>
<tr>
<td></td>
<td>PEVR</td>
<td>623</td>
<td>n/a</td>
<td>623</td>
<td>776</td>
</tr>
<tr>
<td>ES2</td>
<td>PED</td>
<td>822</td>
<td>604</td>
<td>1426</td>
<td>537</td>
</tr>
<tr>
<td></td>
<td>PEVR</td>
<td>910</td>
<td>n/a</td>
<td>910</td>
<td>606</td>
</tr>
<tr>
<td>ES3</td>
<td>PED</td>
<td>612</td>
<td>1366</td>
<td>1978</td>
<td>270</td>
</tr>
<tr>
<td></td>
<td>PEVR</td>
<td>344</td>
<td>n/a</td>
<td>344</td>
<td>475</td>
</tr>
<tr>
<td>ES4</td>
<td>PED</td>
<td>396</td>
<td>1725</td>
<td>2121</td>
<td>654</td>
</tr>
<tr>
<td></td>
<td>PEVR</td>
<td>1176</td>
<td>n/a</td>
<td>1176</td>
<td>1007</td>
</tr>
</tbody>
</table>

Table 4. PED vs PEVR (times are in seconds)

**TD versus PED**

As mentioned above, we also decided to consider the differences between translation and PE when both were performed in the same manner; that is TD and PED; and TDVR and PEVR.

Table 5 compares the results for TD and PED. When looking at the results for translation and PE translation time (total task time; last column), we notice that the results are mixed: while participants ES1 and ES2 were faster with TD, the other two participants (ES3 and ES4) were faster with PED. Interestingly, the transcription time is inversely higher, that is, participants ES1 and ES2 had higher transcription time for the TD tasks, whereas ES3 and ES4 had higher transcription time in PED. Now, when considering the total translation/PE time, we can see that the results are very close, the more visible differences lying for ES1 and ES2, where the former is faster with TD and the latter with PED.

In sum, when looking at the different time measures across phases, we notice no trend in the results. This indicates that, in general, there were not many differences between TD and PED.
### Table 5. TD vs PED (in seconds)

<table>
<thead>
<tr>
<th>Participants</th>
<th>Task</th>
<th>Translation/PE Time</th>
<th>Revision Time</th>
<th>Total Task Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Translation/PE time</td>
<td>Transcription time</td>
<td>Total</td>
</tr>
<tr>
<td>ES1</td>
<td>TD</td>
<td>537</td>
<td>716</td>
<td>1253</td>
</tr>
<tr>
<td></td>
<td>PED</td>
<td>633</td>
<td>692</td>
<td>1325</td>
</tr>
<tr>
<td>ES2</td>
<td>TD</td>
<td>688</td>
<td>1197</td>
<td>1885</td>
</tr>
<tr>
<td></td>
<td>PED</td>
<td>822</td>
<td>604</td>
<td>1426</td>
</tr>
<tr>
<td>ES3</td>
<td>TD</td>
<td>846</td>
<td>1116</td>
<td>1962</td>
</tr>
<tr>
<td></td>
<td>PED</td>
<td>612</td>
<td>1366</td>
<td>1978</td>
</tr>
<tr>
<td>ES4</td>
<td>TD</td>
<td>700</td>
<td>1432</td>
<td>2132</td>
</tr>
<tr>
<td></td>
<td>PED</td>
<td>396</td>
<td>1725</td>
<td>2121</td>
</tr>
</tbody>
</table>

Table 6 compares the results for TDVR and PEVR. We can see that total task times are lower for the first three participants when post-editing with VR than translating from scratch. Only participant ES4 was faster in the translation task. Interestingly, participant ES4 displayed close times for revision, whereas participant ES1 showed lower times to revise the translation. In sum, only participant ES4 showed higher times when post-editing than when translating from scratch, which suggests that PE with the help of VR could generally lead to higher productivity.

### Table 6. TDVR vs PEVR (in seconds)

<table>
<thead>
<tr>
<th>Participants</th>
<th>Task</th>
<th>Translation/PE Time</th>
<th>Revision Time</th>
<th>Total Task Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES1</td>
<td>TDVR</td>
<td>796</td>
<td>656</td>
<td>1452</td>
</tr>
<tr>
<td></td>
<td>PEVR</td>
<td>623</td>
<td>776</td>
<td>1399</td>
</tr>
<tr>
<td>ES2</td>
<td>TDVR</td>
<td>1330</td>
<td>1191</td>
<td>2521</td>
</tr>
<tr>
<td></td>
<td>PEVR</td>
<td>910</td>
<td>606</td>
<td>1516</td>
</tr>
<tr>
<td>ES3</td>
<td>TDVR</td>
<td>377</td>
<td>722</td>
<td>1099</td>
</tr>
<tr>
<td></td>
<td>PEVR</td>
<td>344</td>
<td>475</td>
<td>819</td>
</tr>
<tr>
<td>ES4</td>
<td>TDVR</td>
<td>460</td>
<td>1046</td>
<td>1506</td>
</tr>
<tr>
<td></td>
<td>PEVR</td>
<td>1176</td>
<td>1007</td>
<td>2183</td>
</tr>
</tbody>
</table>

#### 4.2. TX Analysis (Qualitative Phase)

In the follow-on, qualitative phase of this experiment, participants responded to an online questionnaire with sociodemographic questions (see Participant’s profile in section 3.1 above) and retrospective questions about the experiment, as well as questions providing insight on the TX with multimodal/mobile VR-enabled TD and PE applications. The notion of TX is inspired from the notion of user experience (UX) – extensively investigated in the field...
of human-computer interaction – and is defined as “a translator’s perceptions of and responses to the use or anticipated use of a product, system or service” (Zapata, 2016a).

In this section, we report on the results of our questionnaire.

**Subjectively Experienced Productivity**

The questionnaire included an item to ask participants to indicate which one of the four translation tasks they felt made them most productive, and which one made them least productive. Three participants believed that TDVR made them most productive when in fact they had performed the PEVR task faster. Two participants felt that they were slowest in the PED condition. This perception of slower pace when MT has been introduced, contradicting quantitative measurements that recorded increased speed, has been seen elsewhere by Plitt and Masselot (2010) and Gaspari et al. (2014). When compared to their actual productivity times, we note that apart from ES1 regarding TD (where he/she is least productive), the other participants perceive it differently from the actual numbers. Table 7 below shows the perceived productivity against the actual productivity, where $l/L =$ least, $m/M =$ most, lower-case letters are for the perceived productivity and capital letters for the actual productivity.

<table>
<thead>
<tr>
<th>Participant</th>
<th>TD</th>
<th>PED</th>
<th>TDVR</th>
<th>PEVR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES1</td>
<td>$l/L$</td>
<td></td>
<td>m</td>
<td>M</td>
</tr>
<tr>
<td>ES2</td>
<td>l</td>
<td></td>
<td>$m/L$</td>
<td>M</td>
</tr>
<tr>
<td>ES3</td>
<td>l</td>
<td></td>
<td>$m/L$</td>
<td>M</td>
</tr>
<tr>
<td>ES4</td>
<td>m</td>
<td></td>
<td>l</td>
<td>$l/M$</td>
</tr>
</tbody>
</table>

Table 7. Subjectively experienced productivity against actual productivity

**Subjectively Perceived Quality**

The questionnaire also included an item to ask participants to indicate which one of the four translation tasks they felt would result in the best quality, and which one would result in the worst quality (that is, quality of the final target text). Table 8 shows that two of the four participants were confident enough in the PEVR process, that they expected the output texts from that process to be of high quality.

<table>
<thead>
<tr>
<th>Participant</th>
<th>TD</th>
<th>PED</th>
<th>TDVR</th>
<th>PEVR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES1</td>
<td>worst</td>
<td></td>
<td></td>
<td>best</td>
</tr>
<tr>
<td>ES2</td>
<td>worst</td>
<td></td>
<td></td>
<td>best</td>
</tr>
<tr>
<td>ES3</td>
<td>worst</td>
<td></td>
<td></td>
<td>best</td>
</tr>
<tr>
<td>ES4</td>
<td>best</td>
<td></td>
<td></td>
<td>worst</td>
</tr>
</tbody>
</table>

Table 8. Subjectively perceived quality

**Challenges for VR-enabled TD and PE**

A further question asked participants to elaborate on what they thought are the challenges of VR, on the one hand, and of MT, on the other hand, to provide translators with a useful VR-enabled TD and PE tool.
Participants found VR to be reasonably accurate, but with room for improvement, particularly regarding "proper names and figures". Participants preferred translation without the 'constraint' of MT as they considered the suggestions artificial. Participant ES2 wrote that "the Spanish translation sounded more like a transliteration of a technical text in English, and this is not translation as far as I understand". The added cognitive load when MT is added to source and target texts may be initially off-putting for translators, and may add to the perception of decreased speed when MT is introduced to the workflow. They recognized that VR and MT could aid productivity, but would prefer to add MT electively. Participant ES1 wrote that "a translator or post-editor should have the option to translate from scratch by default, and request the help from the machine only when needed". Participant ES2 agreed: “For quality purposes, I prefer the [VR] translation from scratch or post-editing from [translation memories] where you have more leeway.” In the opinion of participant ES4, “MT makes work faster but not necessarily better. It somehow guides the work towards the paradigmatic level. I think the overall cohesion of the document is affected.”

Advantages and Disadvantages of Mobile versus PC-based TD and PE

Finally, participants were asked to elaborate on the perceived advantages and disadvantages of using a mobile TD and PE tool (i.e., on a mobile device such as a smartphone or a tablet) versus a laptop- or PC-based tool. Several mentioned the flexibility of a mobile device, and participant ES2 suggested that "it may help translators to develop interpreting strategies; such as segmentation, quick thinking, anticipation, short-term memory, etc.” Two participants mentioned the difficulties of working in a noisy environment and of speaking translations in a public place. Participant ES3 felt that, although PEVR felt fast to him/her, it was difficult to edit retrospectively. He/she added that if there was “a way to make it more seamless between the keyboard and the mic, a balance so to say, then that'd be amazing.”

5. Conclusion and Future Work

We have reported a pilot experiment on the use of a cloud-based voice recognition (VR) application for translation dictation (TD) and post-editing (PE), using both quantitative and qualitative methods.

In answer to our first research question, based on this small-scale pilot experiment, PE with VR can be as or more productive than comparable approaches, with or without machine translation (MT) and VR. When looking at quantitative data alone, our results showed that, in general, PE with the aid of a VR system was the most efficient method, being the fastest for three of the participants. Interestingly, PE in dictation mode (PED) was the slowest for two participants, followed by TD and TD with VR (TDVR). In the quantitative data, however, we observe that most participants perceived productivity to be higher in the TDVR condition, and expressed a preference to translate/dictate from scratch and have PE added as an option.

One of the issues we identified in our experiment is high revision/editing times in the VR tasks; transcriptions by the VR system were far from flawless, leading to higher revision/editing times. VR applications may produce errors due to translators’ lack of familiarity with TD and insufficient training in how to speak to a VR system, especially for properly adding punctuation using the appropriate commands. Trainers and researchers in translation have explicitly affirmed that training in sight translation, TD, and VR will be essential to succeed with (mobile) voice-enabled tools and devices (Mees et al. 2013; Zapata...
and Quirion, 2016). We noted also that some foreign-language words (e.g. Russian names) in the source texts caused a few misrecognitions in Spanish VR. Moreover, we noticed that some participants would often wait until the software had transcribed a sentence or chunk of a sentence onto the word processor page to continue speaking, which tends to confuse the system (as opposed to when the dictation is continuous). Lastly, if the user pauses for several seconds, the VR system “stops listening” and disconnects, which also causes both the system and the user to lose the flow of the dictation.

Another point to highlight is that the participants’ typing skills may considerably affect translation times. If our time task measures excluded the transcription time in TD and PED, the whole productivity picture would change. Considering this and the issues described in the previous paragraph, the ideal scenario would be one in which translators do not need to transcribe their dictation, either in TD or PE. Instead, they would have a VR system with human-like transcription capabilities, keeping dictation, transcription, and editing/revision times (as well as recognition errors) to a minimum.

In answer to our second research question, participants’ TX suggests that combining MT and VR is indeed feasible for translation projects, with some caveats. When asked about their experience with the tasks, our participants seem to have preferred translation without the ‘constraint’ of MT as they considered the suggestions artificial, though the quantitative results show that the PE task was more efficient than that of translation from scratch. The results of this small-scale experiment suggest that PE with VR (PEVR) may be a usable way to add MT to a translation workflow, and is worth testing at a larger scale.

For future work, we intend to carry out experiments with more participants and language pairs. Further experimentation will include input logging, as well as eye-tracking technologies to collect empirical data on cognitive effort when using VR for TD and PE. We also seek to evaluate the impact of training translators in TD and VR over a period of time before performing TDVR and PEVR tasks. Also, we will include objective measures of quality (with the participation of expert evaluators) to compare it with the participants’ perceived quality of the target texts. Another avenue for future work is to investigate a collaborative scenario in which translators/post-editors collaborate with transcriptionists and/or revisers who would take part in the different phases of the experiment. This list of ideas for future work is of course non-exhaustive; the possibilities seem endless.

The unprecedented robustness of VR technology and its availability on mobile devices via the cloud opens a world of possibilities for human-aided MT and human translation environments. By keeping human translators at the core of research, with strong consideration of their perceptions and preferences for new technologies and applications, we can advance towards finding the right balance in translator-computer interaction (O’Brien, 2012), towards establishing what it is that the machine can do better than humans, and what it is that humans can do better than the machine.

Acknowledgement

We would like to thank our anonymous participants for their time and involvement in this pilot experiment. This work was supported by the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded under the European Regional Development Fund.

References


TOIN

WILL NEURAL MT BE A BREAKTHROUGH IN ENGLISH-TO-JAPANESE TECHNICAL TRANSLATION?

Tsunao Mikasa and Nobuko Kasahara, TOIN Corporation
September 2017

AGENDA

- **Question**: Is MT really usable for English-to-Japanese translation?
- **Pilot project** we carried out for assessing quality and productivity
  - Overview
  - MT engines examined
  - Methodology and assumptions
  - Results
- **Conclusion**
QUESTION

Is MT usable for English-to-Japanese (E2J) translation services where the required quality is at the same level as Human Translation (HT)?

- Until recently, the answer was NO; to obtain certain productivity gains in post-editing, quality of final translation needed to be compromised.
- In other words, only “Light PE” was worth considering, and “real” translation was achievable only by human translators with no help of “machine translators.”

QUESTION (CONT'D.)

Is MT usable for English-to-Japanese (E2J) translation services where the required quality is at the same level as Human Translation (HT)?

- We claim that the answer will be YES if using the latest MT technologies, in particular neural engines (under some reasonable assumptions about content types).
- In other words, MT will enable most E2J translators to achieve the same quality without compromise at higher productivity (except for some special content types, such as marketing materials).
PILOT PROJECT

To examine our claim, we carried out a simple pilot project for accessing quality and productivity in Human Translation (HT) and Post-Editing (PE)

Key Assumptions:

- We focused on Technical documents, as this sector accounts for the largest portion of many language service providers in Japan
- PE quality was required to be the same level as HT, since our interest was in examining whether HT quality can be achieved by PE without any compromise in quality (not “Light PE”)

MT ENGINES EXAMINED

We examined two engines which are recognized as ones of the best Neural and Statistical English-to-Japanese MT engines:

- Google NMT—Neural
- NICT みんなの自動翻訳@TexTra®—Statistical

(Note: NICT has recently also released its Neural engine)
METHODOLOGY AND ASSUMPTIONS

Content translated: A typical technical document, User Manual of a major PLM software product
- Not too technical, easy-to-understand for the average user (and for translators!)

Volume:
- 5k words for PE/HT productivity evaluation
- Additional 10k words for MT quality evaluation

METHODOLOGY AND ASSUMPTIONS (CONTD.)

Sample segments:

<table>
<thead>
<tr>
<th>Segm Source Segment</th>
<th>MT Target Segment</th>
<th>MT Engine</th>
<th>PE-er</th>
<th>Post-edited Target</th>
<th>Translated Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>245: The action is only available when creating or editing a change task.</td>
<td>この操作は、変更タスクを作成または編集するときに使用できるようになります。</td>
<td>NICT</td>
<td>KH</td>
<td>この操作は、変更タスクを作成または編集するときにのみ使用できます。</td>
<td></td>
</tr>
<tr>
<td>246: The action is only available when you access the Resulting Objects table from the change task information page.</td>
<td>操作は、変更タスクの情報ページの「結果オブジェクト」(Resulting Objects)テーブルにアクセスした場合にのみ使用できます。</td>
<td>NICT</td>
<td>KH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>247: Open a new window to edit the change task.</td>
<td>新しいウィンドウを開き、変更タスクを編集します。</td>
<td>NICT</td>
<td>KH</td>
<td>新しいウィンドウを開き、変更タスクを編集します。</td>
<td></td>
</tr>
<tr>
<td>248: Set effectiveness on an object.</td>
<td>オブジェクトの有効性を設定します。</td>
<td>NICT</td>
<td>KH</td>
<td>オブジェクトの有効性を設定します。</td>
<td></td>
</tr>
<tr>
<td>249: View effectiveness on an object.</td>
<td>オブジェクトの有効性を表示します。</td>
<td>NICT</td>
<td>KH</td>
<td>オブジェクトの有効性を表示します。</td>
<td></td>
</tr>
</tbody>
</table>
**METHODOLOGY AND ASSUMPTIONS (CONT'D.)**

Resources—Linguists (Translators/Post-Editors) who worked in the pilot:

- **Four** senior-level linguists with 10+ year-experience in E2J technical translation
- Past experience in PE was **not** required (though two of them did have some PE experience)
- Each of them translated/PE’d the same 5,000-word sample document
- They focused on achieving sufficient (HT-level) quality in PE; never forced to use MT outputs or “hurry up” in PE

**METHODOLOGY AND ASSUMPTIONS (CONT'D.)**

**Linguistic reference:**
Made linguistic reference as **simple** as possible to see the pure impact of MT on quality and productivity:

- No Translation Memory (TM)
- No Terminology Database (TD)
- No Style Guidelines (SG)
METHODOLOGY AND ASSUMPTIONS (CONTD.)

Pilot project for Productivity evaluation

- Each linguist produces a translation of each segment, either by
  - **HT**: translating the source segment without referring to any MT outputs, or
  - **PE**: editing MT output of the source segment
  - To do HT or PE is randomly chosen by the system so that the total # of HT/PE’d segments will be equal

METHODOLOGY AND ASSUMPTIONS (CONTD.)

Pilot project for Productivity evaluation (contd.)

- For **PE**, either **GNMT** or **NICT** engine applied
  - Randomly chosen by the system so that the total # of the segments from each engine will be equal
  - Not make it visible to the linguist which engine was used (to avoid any bias)
- We used **TAUS DQF tools** for productivity evaluation
METHODOLOGY AND ASSUMPTIONS (CONTD.)

Post-editing on TAUS DQF tools:

Quality evaluation of raw MT outputs

- Evaluated quality of raw MT outputs of GNMT and NICT engines
  - Randomly chosen by the system so that the total # of the segments from each engine will be equal
  - Not make it visible to the evaluator which engine was used (to avoid any bias)
**Methodology and Assumptions (Contd.)**

- We used **TAUS DQF tools** and their evaluation criteria for quality evaluation
  - Fluency
    - **Flawless** (4) — refers to a perfectly flowing text with no errors.
    - **Good** (3) — refers to a smoothly flowing text even when a number of minor errors are present.
    - **Disfluent** (2) — refers to a text that is poorly written and difficult to understand.
    - **Incomprehensible** (1) — refers to a very poorly written text that is impossible to understand.
  - Adequacy
    - **Everything** (4) — All the meaning in the source is contained in the translation, no more, no less.
    - **Most** (3) — Almost all the meaning in the source is contained in the translation.
    - **Little** (2) — Fragments of the meaning in the source are contained in the translation.
    - **None** (1) — None of the meaning in the source is contained in the translation.

**Results—Productivity**

Overall Productivity (wph)

<table>
<thead>
<tr>
<th>Words per hour</th>
<th>PE GNMT</th>
<th>PE NICT</th>
<th>PE avg.</th>
<th>HT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>587</td>
<td>466</td>
<td>518</td>
<td>360</td>
</tr>
</tbody>
</table>

Proceedings of MT Summit XVI, Vol.2: Users and Translators Track  
Nagoya, Sep. 18-22, 2017 | p. 137
RESULTS—PRODUCTIVITY (CONTD.)

Productivity (wph) by Translator

<table>
<thead>
<tr>
<th>Words per hour</th>
<th>Overall</th>
<th>Words per hour</th>
<th>Translator SJ</th>
<th>Words per hour</th>
<th>Translator KH</th>
<th>Words per hour</th>
<th>Translator TH</th>
<th>Words per hour</th>
<th>Translator SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>587</td>
<td>620</td>
<td>608</td>
<td>677</td>
<td>496</td>
<td>466</td>
<td>392</td>
<td>370</td>
<td>380</td>
</tr>
<tr>
<td>Translator SJ</td>
<td>518</td>
<td>312</td>
<td>512</td>
<td>287</td>
<td>344</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Translator KH</td>
<td>360</td>
<td>723</td>
<td>661</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Translator TH</td>
<td>446</td>
<td></td>
<td>466</td>
<td></td>
<td></td>
<td></td>
<td>287</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

RESULTS—PRODUCTIVITY (CONTD.)

Key findings:
- **PE w/ GNMT**
  - **Highest** productivity
  - **63%** faster than HT on average
- **PE w/ NICT**
  - **30%** faster than HT on average
RESULTS—PRODUCTIVITY (contd.)

- **PE GNMT > PE NICT > HT**
  —The same tendency observed almost independent of the translator

- **Correlation ratio** between Productivity and MT Engine/HT:
  \( \eta = 0.57 \)
  \[
  \eta := \frac{\sum_{i=1}^{n_i} n_i (x_i - \overline{x})^2}{\sum_{i=1}^{n_i} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2}
  \]
  \((0 \leq \eta \leq 1)\)

RESULTS—PRODUCTIVITY (contd.)

Other observations:
- **No** apparent correlation observed between Productivity and Segment Length (word count of each segment)
- In particular, in HT, SL does not seem to affect Productivity at all
- **GNMT** seems to show a slight tendency that the longer SL, the **higher productivity**, but it’s not significant
RESULTS—QUALITY: FLUENCY

Fluency Evaluation Results

Fluency: GNMT

- Flawless: 42%
- Good: 28%
- Disfluent: 1%
- Incomprehensible: 1%

Fluency: NICT

- Flawless: 34%
- Good: 26%
- Disfluent: 6%
- Incomprehensible: 6%

RESULTS—QUALITY: ADEQUACY

Adequacy Evaluation Results

Adequacy: GNMT

- Everything: 1%
- Most: 9%
- Little: 32%
- None: 50%

Adequacy: NICT

- Everything: 28%
- Most: 28%
- Little: 32%
- None: 12%
RESULTS—QUALITY

○ Key findings
  • GNMT had better scores overall, where
    ○ +85% segments had Flowless (4) or Good (3) fluency
    ○ +90% segments had Everything (4) or Most (3) adequacy

○ Other observations
  • Almost no correlation observed between Segment Length and Quality of MT outputs in our pilot:

    Scatter plots of the evaluation data:

    ![Scatter plots](image)

CONCLUSION

Productivity gains
○ We observed 63% average productivity gains in PE w/ GNMT as well as 30% gains in PE w/ NICT.
○ This strongly suggests that significant improvement in efficiency can be achieved in most E2J technical localization projects by utilizing the latest MT engines, in particular, GNMT, in the translation process.

Other findings
○ In our pilot, we didn’t observe the tendency “Longer sentences give worse MT outputs, thus result in lower PE productivity”, which may be a myth.
The Impact of MT Quality Estimation on Post-Editing Effort

Carlos Teixeira  Sharon O’Brien
Carlos.teixeira@dcu.ie sharon.obrien@dcu.ie

Centre for Translation and Textual Studies (CTTS)
ADAPT Centre for Digital Content Technology
Dublin City University (Ireland)

MOTIVATION

- Professional translators edit suggestions coming from translation memories (TM) and machine translation (MT)
- Handling those two *types of linguistic support* requires different strategies
- TM suggestions incorporate metadata to increase efficiency and quality (e.g. Fuzzy Match scores)
- QE scores are an attempt to provide *relevant metadata* for MT suggestions

**Novelty:** Despite recent advances in QE research, little is known about the *real impact of QE scores* on the translation process.
POTENTIAL IMPACT

- Improve translators’ efficiency when working with MT
- Reduce cognitive strain on translators
- Increase translators’ trust in MT output by offering accurate QE
- Reduce their need to search for validation from additional, external resources (cf. Bundgaard 2017, Daems et al. 2016)

EXPERIMENT DESIGN

- Online post-editing tool (HandyCAT)

Only source text and MT displayed

- Participants: 20 professional translators
- Materials: 4 texts (WMT13 news material)
- Languages: English → Spanish
- Four different QE modes (more details below)
EXPERIMENT DESIGN (cont.’d)

QE mode consists of two parts:

- **Score Type:**
  - No QE: the QE box is hidden in HandyCAT
  - Accurate QE: scores obtained from the automatic scoring system that ranked best in the WMT13 shared task (automatic, accurate)
  - Inaccurate QE: ‘random’ scores (automatic, inaccurate)
  - Human QE: scores obtained using a human evaluation method (human, accurate) (Graham et al. 2015)

- **Score Level:** Percentage (between ~20% and 99%)

EXPERIMENT DESIGN (cont.’d)

Research question:

- What is the impact of the different modes of QE scores on:
  - temporal effort (time spent)
  - physical effort (number of keystrokes)
  - cognitive effort (gaze behaviour)

Data collection:

- activity logging
- screen recording
- eye tracking
EXPERIMENT DESIGN (cont.’d)

Full range of variables being considered:

<table>
<thead>
<tr>
<th>Role</th>
<th>Name</th>
<th>Type</th>
<th>Measurement / Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Physical – Cognitive –</td>
<td>Translation time</td>
<td>numeric</td>
<td>seconds per word</td>
</tr>
<tr>
<td></td>
<td>Amount of typing</td>
<td></td>
<td>keys per word</td>
</tr>
<tr>
<td></td>
<td>Fixation count</td>
<td></td>
<td>n per word</td>
</tr>
<tr>
<td></td>
<td>Fixation duration</td>
<td></td>
<td>seconds per word</td>
</tr>
<tr>
<td></td>
<td>Pupil dilation</td>
<td></td>
<td>mm (variance)</td>
</tr>
<tr>
<td>Independent (Fixed effects)</td>
<td>Primary QE score type</td>
<td>categorical</td>
<td>No_QE, Acc_QE, Inacc_QE, Human_QE</td>
</tr>
<tr>
<td></td>
<td>QE score level</td>
<td></td>
<td>N/A (No_QE condition)</td>
</tr>
<tr>
<td></td>
<td>Document</td>
<td></td>
<td>SRC1, SRC2, SRC5, SRC7</td>
</tr>
<tr>
<td></td>
<td>Task order</td>
<td></td>
<td>T01, T02, T03, T04</td>
</tr>
</tbody>
</table>

RESULTS - Temporal effort

(time spent per word)

Fixed Effects

<table>
<thead>
<tr>
<th>Source</th>
<th>F</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected Model</td>
<td>29.680</td>
<td>12</td>
<td>1.027</td>
<td>.000</td>
</tr>
<tr>
<td>Score_Type</td>
<td>0.035</td>
<td>2</td>
<td>1.027</td>
<td>.965</td>
</tr>
<tr>
<td>Score_Level_Ordinal</td>
<td>14.049</td>
<td>3</td>
<td>1.027</td>
<td>.000</td>
</tr>
<tr>
<td>Document</td>
<td>34.544</td>
<td>3</td>
<td>1.027</td>
<td>.000</td>
</tr>
<tr>
<td>Task_Order</td>
<td>1.850</td>
<td>3</td>
<td>1.027</td>
<td>.120</td>
</tr>
</tbody>
</table>

Primary variables

Secondary variables

Probability distribution: Normal
Link function: Identity
RESULTS - Temporal effort

Effects found for Document

Estimated Means: Document

Target: log Time

Pairwise Contrasts

Significant contrasts are shaded gold. The sequential Bonferroni adjusted significance level is .05.

C. Teixeira & S. O’Brien

MT Summit XVI

10
RESULTS – Physical effort

(# of keys typed per word)

Results are similar to the ones found for Temporal effort:

- **No significant** differences in average # of keys according to Score Type
- **Significant** differences in average # of keys according to Score Level

---

RESULTS – Cognitive effort

Fixation duration per Score Level – No significant effects found

![Box plot showing fixation duration per score level](image-url)
RESULTS – Cognitive effort

Fixation duration per Score Type – No significant effects found

![Box plot showing fixation duration per Score Type](image)

Pupil diameter per Score Level – No significant effects found

![Box plot showing pupil diameter per Score Level](image)
RESULTS – Cognitive effort

Pupil diameter per Score Type – No significant effects found

![Box plots showing pupil diameter for different score types.](image)

C. Teixeira & S. O’Brien  
MT Summit XVI  
17

SUMMARY

Our results indicate:

- No significant effect of **Score Type** on either time or edits.

- A significant effect of **Score Level** on both time and edits:
  
  The higher the score level the less time is spent and the fewer keys are typed (regardless of how the scores were calculated!)

- Displaying QE scores (even if they are accurate) is not necessarily better than displaying no scores.

- No significant variations in the number of fixations, fixation duration or pupil size that could be associated with the display of QE scores.
DISCUSSION

- In our experiment, only a QE percentage was displayed.
- Perhaps the same results would have been found for TM if we had removed the *diff* indication and just left the Match percentages?
DISCUSSION (cont.’d)

- Our results point toward the need to combine QE scores with the display of phrase-level or word-level QE indication.

This is what we displayed:

```
The Army intelligence analyst, arrested in June 2010, is accused of stealing thousands of classified documents while serving in Iraq.
```

```
El ejército analista de inteligencia, detenido en junio de 2010, es acusado de robar miles de documentos clasificados aunque sirven en Iraq.
```

DISCUSSION (cont.’d)

- Our results point toward the need to combine QE scores with the display of phrase-level or word-level QE indication.

This might be the way forward to make QE more effective:

```
The Army intelligence analyst, arrested in June 2010, is accused of stealing thousands of classified documents while serving in Iraq.
```

```
El ejército analista de inteligencia, detenido en junio de 2010, es acusado de robar miles de documentos clasificados aunque sirven en Iraq.
```
FUTURE RESEARCH

- Test the effect of word-level or phrase-level QE indicators
- Test different layouts for the presentation of QE information
- Try more fine-grained buckets of QE score levels to identify ideal cut-off point
- Assess the effect of QE on the Quality of the final translations
- Study the impact of QE if translators learned to trust the information (longitudinal)

REFERENCES


Thank you!

ありがとうございます

Carlos Teixeira  carlos.teixeira@dcu.ie
Sharon O’Brien  sharon.obrien@dcu.ie
Agenda

- Evaluation
  - Raw Output Quality
  - Throughput PE vs HT
- Challenges / Best Practices
- Key Takeaways
- Q&A
Raw Output Quality Evaluation – Details of sample

- Human assessment
- Language pair: English-Japanese
- Translation volume: 3786 words
- Content type: Software manual

Raw Output Quality Evaluation – Human assessment / how to score

<table>
<thead>
<tr>
<th>Meaning and Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Perfectly Understandable</td>
</tr>
<tr>
<td>2. Fully Understandable</td>
</tr>
<tr>
<td>3. Barely Understandable</td>
</tr>
<tr>
<td>4. Not Understandable</td>
</tr>
</tbody>
</table>

Better

Worse
Raw Output Quality Evaluation – Results

-Total-

-strings containing 20 words or more-
Raw Output Quality Evaluation – Analysis

- GNMT is extremely good in translating software manuals
- The quality of GNMT is high even with long sentences
- Reasons (speculations):
  - There are large amount of software manuals on the Internet
  - Google crawls the Internet for its training corpus
  - GNMT is like an MT system with a huge translation memory from multiple software vendors

Throughput Evaluation - Details

- Localization projects of various document types
- Not in production but completely simulated

<table>
<thead>
<tr>
<th>(As of July 28, 2017)</th>
<th>Source volume</th>
<th>Number of projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>English-Japanese</td>
<td>49,883 weighted words</td>
<td>36</td>
</tr>
<tr>
<td>Japanese-English</td>
<td>10,057 weighted characters</td>
<td>2</td>
</tr>
</tbody>
</table>
Throughput Evaluation - Context levels

- Translation depends on the information outside the sentence
- Other sentences in the document
- Basic knowledge of the products / services
- Common sense

- LOW: A sentence provides all the information for translation.
- MEDIUM
- HIGH: Information from other sources is needed.

<table>
<thead>
<tr>
<th>Weighted word count</th>
<th>PostEdit time (hr)</th>
<th>Speed (w/hr)</th>
<th>Content type</th>
<th>Context level</th>
</tr>
</thead>
<tbody>
<tr>
<td>2174</td>
<td>11</td>
<td>197.6</td>
<td>Training role play scripts</td>
<td>HIGH</td>
</tr>
<tr>
<td>1229</td>
<td>4</td>
<td>307.3</td>
<td>Resource file</td>
<td></td>
</tr>
<tr>
<td>1108</td>
<td>3.5</td>
<td>316.8</td>
<td>FAQ (web services)</td>
<td></td>
</tr>
<tr>
<td>3175</td>
<td>10</td>
<td>317.5</td>
<td>Product information</td>
<td></td>
</tr>
<tr>
<td>1682</td>
<td>4</td>
<td>420.5</td>
<td>FAQ (web services)</td>
<td>LOW</td>
</tr>
<tr>
<td>1482</td>
<td>3</td>
<td>494.0</td>
<td>Service description</td>
<td>LOW</td>
</tr>
<tr>
<td>1023</td>
<td>2</td>
<td>615.0</td>
<td>Software manual</td>
<td>LOW</td>
</tr>
</tbody>
</table>

Faster than human translation (~250w/hr)
### Throughput Evaluation - Results: English-Japanese

**Throughput average by context level**

- **High (1 editor)**: TP Avg
- **Med (2 editors)**: TP Avg, # of prj
- **Low (7 editors)**: TP Avg, # of prj

![Graph showing throughput averages for different context levels](image)

### Throughput Evaluation - Results: Japanese-English

<table>
<thead>
<tr>
<th>Weighted char count</th>
<th>PostEdit time (hr)</th>
<th>Speed (ch/hr)</th>
<th>Content type</th>
<th>Context level</th>
</tr>
</thead>
<tbody>
<tr>
<td>9352</td>
<td>4</td>
<td>2338.0</td>
<td>Whitepaper</td>
<td>LOW</td>
</tr>
<tr>
<td>705</td>
<td>0.33</td>
<td>2350.0</td>
<td>Developer page (UGC)</td>
<td>MEDIUM</td>
</tr>
</tbody>
</table>

Needs to collect more data, but much faster than manual translation (~500ch/hr) + good for UGC
Challenges – Fun Fact

There are 24 spelling patterns for the translation of **User Interface**:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ユーザーインターフェース</td>
<td>ユーザーインタフェース</td>
<td>ユーザーインターフェイス</td>
<td>ユーザーインタフェイス</td>
</tr>
<tr>
<td>ユーザインターフェース</td>
<td>ユーザインタフェース</td>
<td>ユーザインタフェース</td>
<td>ユーザインタフェース</td>
</tr>
<tr>
<td>ユーザーインターフェース</td>
<td>ユーザーインターフェース</td>
<td>ユーザーインターフェイス</td>
<td>ユーザーインターフェイス</td>
</tr>
<tr>
<td>ユーザインターフェース</td>
<td>ユーザインターフェース</td>
<td>ユーザインターフェース</td>
<td>ユーザインターフェース</td>
</tr>
<tr>
<td>ユーザーインターフェース</td>
<td>ユーザーインターフェース</td>
<td>ユーザーインターフェイス</td>
<td>ユーザーインターフェイス</td>
</tr>
<tr>
<td>ユーザインターフェース</td>
<td>ユーザインターフェース</td>
<td>ユーザインターフェース</td>
<td>ユーザインターフェース</td>
</tr>
</tbody>
</table>

▲ stands for a single-byte space.

Challenges (1) – Following rules in style guides (1)

Most of companies have their own style guides and the rules are slightly different, such as spacing rules, brackets, long vowels (*cho-on*), etc.

<table>
<thead>
<tr>
<th>Spacing rules</th>
<th>Company A</th>
<th>Company B</th>
<th>Company C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Katakana words</td>
<td>ユーザーインターフェース</td>
<td>ユーザーインターフェース</td>
<td>ユーザーインターフェイス</td>
</tr>
<tr>
<td></td>
<td>User interface</td>
<td>User interface</td>
<td>User interface</td>
</tr>
<tr>
<td></td>
<td>ユーザーインターフェース</td>
<td>ユーザーインターフェース</td>
<td>ユーザーインターフェイス</td>
</tr>
<tr>
<td>Between single-byte and double-byte characters</td>
<td>From Sept. 19 to 21 9月19日～9月21日</td>
<td>From Sept. 19 to 21 9月19日～9月21日</td>
<td>From Sept. 19 to 21 9月19日～9月21日</td>
</tr>
</tbody>
</table>
Challenges (1) – Following rules in style guides (2)

- Most of companies have their own style guides and the rules are slightly different, such as spacing rules, brackets, long vowels (*cho-on*), etc.

<table>
<thead>
<tr>
<th></th>
<th>Company A</th>
<th>Company B</th>
<th>Company C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brackets</td>
<td>Use [ ] (single-byte) for user interface terms</td>
<td>Use 「」 (double-byte) for user interface terms</td>
<td>Use 「」 (double-byte) for user interface terms</td>
</tr>
<tr>
<td></td>
<td>Use 『』 for book titles and use 「」 for chapter/section titles</td>
<td>Use 『』 for book, chapter and section titles</td>
<td>Use 『』 for book, chapter and section titles</td>
</tr>
<tr>
<td>Long</td>
<td>User … ユーザー Printer … プリンター Programmer … プログラマ</td>
<td>User … ユーザー Printer … プリンター Programmer … プログラマ</td>
<td>User … ユーザー Printer … プリンター Programmer … プログラマ</td>
</tr>
<tr>
<td>vowels</td>
<td>(cho-on) (depends of numbers of syllables)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Challenges (2) – Tone (*de-aru vs desu-masu* (常体/敬体))

- There are two major writing styles in Japanese, *de-aru* style vs *desu-masu* style. These styles should be applied appropriately to match the context.

<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Raw MT</th>
<th>Post edited</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>de-aru</em> style (常体)</td>
<td>(This course helps you to: )</td>
<td>• ABC 製品の新しいサービスと機能を使用して、最新の技術を学ぶことができます。</td>
<td>• ABC 製品の新しいサービスと機能を使用して、最新の技術について学習する。 (e.g., bullet items)</td>
</tr>
<tr>
<td></td>
<td>• Use new services and features from the ABC product to learn about modern technologies.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>desu-masu</em> style (敬体)</td>
<td>Use new services and features from the ABC product to learn about modern technologies.</td>
<td>ABC 製品の新しいサービスと機能を使用して、最新の技術を学ぶことができま</td>
<td>ABC 製品の新しいサービスと機能を使用すると、最新の技術を学ぶことができます。 (e.g., normal texts)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>•</td>
<td></td>
</tr>
</tbody>
</table>
Challenges (3) – Glossary (UI terms / client specific / titles of references)

- Most companies have UI glossaries and terminologies so the post editors need to apply the appropriate terms.

<table>
<thead>
<tr>
<th>Source</th>
<th>Raw MT</th>
<th>Post edited</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Example 1</strong></td>
<td>Click on the “Continue” button.</td>
<td>「続行」ボタンをクリックします。</td>
</tr>
<tr>
<td><strong>Example 2</strong></td>
<td>a getting started guide</td>
<td>スタートガイド</td>
</tr>
<tr>
<td><strong>Example 3</strong></td>
<td>詳細については、「APIを使用した展開」を参照してください。</td>
<td>For details, see “Deployment using API”.</td>
</tr>
</tbody>
</table>

Challenges (4) – General terms (Contexts/Inconsistencies)

- Post editors need to apply the correct translations to context-sensitive terms.
- Even the translations are correct, they must be consistent.

<table>
<thead>
<tr>
<th>Source</th>
<th>Raw MT</th>
<th>Consider when post editing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Example 1</strong></td>
<td>available</td>
<td>利用可能 (able to use) ご利用いただけます (polite “able to use”) あります (exists / be in stock)</td>
</tr>
<tr>
<td><strong>Example 2</strong></td>
<td>question</td>
<td>質問 (an act of asking) 問題 (a problem) 疑わしいこと (a doubt)</td>
</tr>
<tr>
<td><strong>Example 3</strong></td>
<td>server-side</td>
<td>サーバーサイド (server side) サーバー側 (server side)</td>
</tr>
</tbody>
</table>
### Challenges (4) – General terms (new words/buzzwords)

- Some new words may not be translated correctly sometimes.

<table>
<thead>
<tr>
<th>Source</th>
<th>Raw MT</th>
<th>Post edited</th>
</tr>
</thead>
<tbody>
<tr>
<td>deep dive</td>
<td>The XXX Conference is a one-day <strong>deep dive</strong> into new technology.</td>
<td>XXX Conference は、新たな技術についての深いダイビングです。 (<em>a recreational diving</em>)</td>
</tr>
<tr>
<td>DevOps</td>
<td><strong>DevOps</strong> focuses on improving automation.</td>
<td>開発部門は自動化の改善に重点を置いています。 (<em>Development Dept.</em>)</td>
</tr>
</tbody>
</table>

### Challenges (5) – Tags / variables

- In most cases, tags are not properly treated. Also, tags can cause poor translation.

<table>
<thead>
<tr>
<th>Source</th>
<th>Raw MT</th>
<th>Post edited</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;br&gt; tag</td>
<td>Cover&lt;br&gt;letter</td>
<td>Coverletter (<em>the tag is omitted</em>)</td>
</tr>
<tr>
<td>variable tag</td>
<td>Please ¥{0¥} to try again. (<em>unnecessary spaces</em>)</td>
<td>再試行するには¥▲¥{0▲¥}してください。</td>
</tr>
</tbody>
</table>
### Challenges and solutions

<table>
<thead>
<tr>
<th>Issues</th>
<th>Solutions</th>
<th>Can be fixed automatically?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Client-specific style specifications</td>
<td>Apply the rules with regular expression</td>
<td>Some yes, others no</td>
</tr>
<tr>
<td>Tone</td>
<td>Check and replace manually in Post Edit</td>
<td>No</td>
</tr>
<tr>
<td>Terminology (UI / client-specific / ref mat titles)</td>
<td>Apply some translations from terminology file automatically, and then replace manually in Post Edit (if necessary)</td>
<td>Some yes, others no</td>
</tr>
<tr>
<td>Terminology (general/new terms)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tags / variables</td>
<td>Delete before MT and insert manually in Post Edit</td>
<td>No</td>
</tr>
</tbody>
</table>

### Best Practices

- Decide the content type to be machine-translated
  - Manuals, user interface, FAQ, UGC, marketing contents

- Align the final expectations between client and LSP
  - Final quality of translation, TATs, costs, content cycles

- Then, support and train post editors
  - Appropriate allocation of post editors by content type and final quality expectation, pre-process with SW components, continuous feedback loop
Takeaways - Neural MT for Commercial Use

- NMT makes the translation hours 1.36x faster and the productivity 1.48x higher (evaluation average)
- Usable in production both in English-Japanese and Japanese-English pairs in IT localization (incl. UGC)
- There are issues to be solved manually in Post Edit, but some can be automatically processed with software components

Q&A

t-shishido@science.co.jp
www.science.co.jp
+81-3-5321-3111
Comparative Evaluation of NMT with Established SMT Programs

Lena Marg
Naoko Miyazaki
Elaine O’Curran
Tanja Schmidt

AGENDA

1. Objective
2. Scope of the evaluation
   - Language pairs
   - Content types
   - Size and integrity of the test sets
3. Evaluation methodologies
   - Human evaluations
   - Automatic scoring
4. Results
5. Conclusions
WHO + WHERE WE ARE

WORDS TRANSLATED 2015: 1.15 BILLION
LANGUAGES TRANSLATED: 175+
EMPLOYEES: 1000+
GLOBAL OFFICES: 21
7TH LARGEST PROVIDER IN THE WORLD
4TH LARGEST LSP IN THE US
*2016 Common Sense Advisory

Objective

Proceedings of MT Summit XVI, Vol.2: Users and Translators Track
Nagoya, Sep. 18-22, 2017 | p. 167
Objective

- Compare the performance of two public NMT systems with a customized SMT solution that is applied in production for two enterprise-level clients.
- Evaluate how generic NMT performs out-of-the-box for different languages and content types that are in high demand in our industry.
- Enable us to make well-founded business decisions as we move forward with our MT strategy.
- Provide data-driven advice and support to our clients.

Scope of the Evaluation
Sampling and Sample Size

<table>
<thead>
<tr>
<th>Evaluation Type</th>
<th>Sample Size (TUs)</th>
<th>Sample Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autoscoring (HT)</td>
<td>Approx. 2500</td>
<td>This is the randomized, blind test set taken from the customized SMT engine.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The segments in the test set are not included in the engine's training data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>and originate from production TMs.</td>
</tr>
<tr>
<td>Side-by-side engine</td>
<td>200</td>
<td>The 200 segments for human evaluation are randomly selected from the 2500 TU</td>
</tr>
<tr>
<td>ranking</td>
<td></td>
<td>test set described above</td>
</tr>
<tr>
<td>Adequacy and Fluency</td>
<td>100</td>
<td>From the 200 segments above, we randomly selected 100 segments for the more</td>
</tr>
<tr>
<td>scoring</td>
<td></td>
<td>detailed human analysis and post-editing sample</td>
</tr>
<tr>
<td>Strength and Weaknesses</td>
<td>100</td>
<td>Same sample as above</td>
</tr>
<tr>
<td>Assessment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autoscoring (PE)</td>
<td>100</td>
<td>Same sample as above is post-edited and scored</td>
</tr>
</tbody>
</table>

Scope Overview

<table>
<thead>
<tr>
<th>Evaluation Type</th>
<th>MT Systems</th>
<th>Content Type</th>
<th>Language Pairs</th>
<th>Evaluators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Side-by-side engine</td>
<td>Customized SMT, Generic1 NMT, Generic2 NMT</td>
<td>Light Marketing, Technical Documentation</td>
<td>de-DE, fr-FR, ja-JP, pt-BR, ru-RU, zh-CN</td>
<td>Two evaluators: one account translator, one experienced MT evaluator</td>
</tr>
<tr>
<td>ranking</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scoring</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autoscoring (PE)</td>
<td>Customized SMT, Generic1 NMT</td>
<td>Light Marketing</td>
<td>de-DE, ja-JP, pt-BR</td>
<td>One evaluator: account translator</td>
</tr>
</tbody>
</table>
Methodology

Side-by-Side Engine Ranking

- The TAUS DQF tool used for this evaluation randomizes the order in which the target segments from the engines being compared are presented. This means the evaluator(s) do not get conditioned into giving anticipated rankings.
- Ranking (1,2,3) of the 3 engines, from best to worst.
- Allows equal ranking of two or three outputs.
Adequacy and Fluency Scoring

<table>
<thead>
<tr>
<th>Adequacy Score Evaluation Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 All meaning expressed in the source appears in the translation. You do not need to refer to the source to understand the meaning.</td>
</tr>
<tr>
<td>4 Most of the source meaning is expressed in the translation. You can understand most of the meaning without referring to the source.</td>
</tr>
<tr>
<td>3 Much of the source meaning is expressed in the translation. Roughly half the MT output can be understood without referring to the source.</td>
</tr>
<tr>
<td>2 Little of the source meaning is expressed in the translation. Although you can guess fractions of the MT output, you cannot understand it without referring to the source.</td>
</tr>
<tr>
<td>1 None of the meaning expressed in the source is expressed in the translation. You cannot make any sense of the MT output alone AND/OR the MT output says exactly the opposite of the source.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fluency Score Evaluation Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Native language fluency. No grammar errors, good word choice and syntactic structure. No PE required.</td>
</tr>
<tr>
<td>4 Near native fluency. Few terminology or grammar errors which don’t impact the overall understanding of the meaning. Little PE required.</td>
</tr>
<tr>
<td>3 Not very fluent. About half of translation contains errors and requires PE.</td>
</tr>
<tr>
<td>2 Little fluency. Wrong word choice, poor grammar and syntactic structure. A lot of PE required.</td>
</tr>
<tr>
<td>1 No fluency. Absolutely ungrammatical and for the most part doesn’t make any sense. Translation has to be re-written from scratch.</td>
</tr>
</tbody>
</table>

Ranking Strengths and Weaknesses

<table>
<thead>
<tr>
<th>WHICH TRANSLATION IS BETTER WITH REGARD TO:</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuracy (accurate rendition of source meaning)</td>
</tr>
<tr>
<td>fluency &amp; style</td>
</tr>
<tr>
<td>general domain terminology</td>
</tr>
<tr>
<td>client-specific terminology &amp; instructions</td>
</tr>
<tr>
<td>completeness (all key information from source is rendered)</td>
</tr>
<tr>
<td>redundancy (translation contains additional information not contained in the source)</td>
</tr>
<tr>
<td>syntax</td>
</tr>
<tr>
<td>grammar</td>
</tr>
<tr>
<td>localization (correct format of punctuation; spacing; dates &amp; time, units measurement)</td>
</tr>
<tr>
<td>tags &amp; placeholders</td>
</tr>
<tr>
<td>spelling</td>
</tr>
<tr>
<td>Other</td>
</tr>
</tbody>
</table>

we localize doing things differently
Autoscoring

- BLEU
- NIST
- METEOR
- GTM
- Precision
- Recall
- TER
- PE Distance*

*In our analysis we focus on PE distance, which applies the Levenshtein algorithm and is character-based. Compared to word-based scoring, this method captures morphological post-edits, such as fixing word forms, and we have found it to correlate well with human judgment.

Results
Engine Ranking Results for Light Marketing

**ENGINE RANKED BEST**

<table>
<thead>
<tr>
<th></th>
<th>PTBR</th>
<th>FR</th>
<th>DE</th>
<th>RU</th>
<th>ZHCN</th>
<th>JA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customized SMT</td>
<td>32.02%</td>
<td>32.77%</td>
<td>30.93%</td>
<td>31.18%</td>
<td>28.38%</td>
<td>27.72%</td>
</tr>
<tr>
<td>Generic2 NMT</td>
<td>31.37%</td>
<td>32.49%</td>
<td>32.12%</td>
<td>31.68%</td>
<td>32.66%</td>
<td>31.91%</td>
</tr>
<tr>
<td>Generic1 NMT</td>
<td>36.61%</td>
<td>34.74%</td>
<td>36.95%</td>
<td>37.14%</td>
<td>38.95%</td>
<td>40.37%</td>
</tr>
</tbody>
</table>

---

Engine Ranking Results for Technical Documentation

**ENGINE RANKED BEST**

<table>
<thead>
<tr>
<th></th>
<th>PTBR</th>
<th>FR</th>
<th>DE</th>
<th>ZHCN</th>
<th>JA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customized SMT</td>
<td>30.33%</td>
<td>29.30%</td>
<td>28.56%</td>
<td>27.25%</td>
<td>27.35%</td>
</tr>
<tr>
<td>Generic2 NMT</td>
<td>32.68%</td>
<td>34.11%</td>
<td>35.94%</td>
<td>35.10%</td>
<td>34.79%</td>
</tr>
<tr>
<td>Generic1 NMT</td>
<td>36.99%</td>
<td>36.59%</td>
<td>35.94%</td>
<td>37.65%</td>
<td>37.86%</td>
</tr>
</tbody>
</table>
### German Results

<table>
<thead>
<tr>
<th>Locale</th>
<th>Evaluation</th>
<th>Light Marketing Content</th>
<th>Technical Documentation Content</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Generic1 NMT</td>
<td>Generic2 NMT</td>
</tr>
<tr>
<td></td>
<td>Ranking</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Adequacy</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Fluency</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Fluency &amp; Style</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Syntax</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Grammar</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Terminology</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Completeness</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Localization</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Edit Distance (HT)</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Edit Distance (PE)</td>
<td>√</td>
<td>2</td>
</tr>
</tbody>
</table>

### Japanese Results

<table>
<thead>
<tr>
<th>Locale</th>
<th>Evaluation</th>
<th>Light Marketing Content</th>
<th>Technical Documentation Content</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Generic1 NMT</td>
<td>Generic2 NMT</td>
</tr>
<tr>
<td></td>
<td>Ranking</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Adequacy</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Fluency</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Fluency &amp; Style</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Syntax</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Grammar</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Terminology</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Completeness</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Spelling</td>
<td>√</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Edit Distance (HT)</td>
<td>√</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Edit Distance (PE)</td>
<td>√</td>
<td>2</td>
</tr>
</tbody>
</table>
## Brazilian Results

<table>
<thead>
<tr>
<th>Locale</th>
<th>Evaluation</th>
<th>Light Marketing Content</th>
<th>Technical Documentation Content</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Generic1 NMT</td>
<td>Generic2 NMT</td>
</tr>
<tr>
<td>pt-BR</td>
<td>Ranking</td>
<td>√</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>√</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>Fluency</td>
<td>√</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>Fluency &amp; Style</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Completeness</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Redundancy</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Syntax</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grammar</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Terminology</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Localization</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tags &amp; Placeholders</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Edit Distance (HT)</td>
<td>2 2</td>
<td>√</td>
</tr>
<tr>
<td></td>
<td>Edit Distance (PS)</td>
<td>2 3</td>
<td>√</td>
</tr>
</tbody>
</table>

## French, Russian, Simplified Chinese Results

<table>
<thead>
<tr>
<th>Locale</th>
<th>Evaluation</th>
<th>Light Marketing Content</th>
<th>Technical Documentation Content</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Generic1 NMT</td>
<td>Generic2 NMT</td>
</tr>
<tr>
<td>fr-FR</td>
<td>Ranking</td>
<td>√</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Edit Distance (HT)</td>
<td>2 2</td>
<td>√</td>
</tr>
<tr>
<td>zh-CN</td>
<td>Ranking</td>
<td>√</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Edit Distance (HT)</td>
<td>1 2</td>
<td>√</td>
</tr>
<tr>
<td>ru-RU</td>
<td>Ranking</td>
<td>√</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Edit Distance (HT)</td>
<td>2 3</td>
<td>√</td>
</tr>
</tbody>
</table>
SUMMARY

- All evaluators prefer generic NMT during side-by-side ranking, the first evaluation task.
- NMT also wins Adequacy & Fluency scoring with the exception of German Adequacy for Light Marketing.
- Evaluators for JA, DE, PTBR overall prefer customized SMT for terminology and localization-related issues, but NMT for fluency, style, grammar and syntax. JA also prefers NMT for accuracy.
- NMT outperforms SMT more consistently on Technical Documentation than on Light Marketing.
- For Technical Documentation the autoscores favor NMT, while they show mixed results for Light Marketing.
- After completing the post-editing task on Light Marketing, the German and Brazilian translators had a slight preference for SMT, contradicting the previous human evaluation results and indicating that the autoscores may be more accurate.

SUMMARY

- The most significant quality improvement with NMT are for Chinese and Japanese
- For the other languages, the quality differences between NMT and SMT are less pronounced
Conclusions

- Generic NMT is a suitable alternative for generic domains across all the language pairs.
- In the technology domain, generic NMT is a suitable alternative for some language pairs, such as Chinese and Japanese, where we see a substantial increase in performance compared to customized SMT.
- Because most of our enterprise-level programs rely on accurate terminology, we recommend waiting for customized NMT for the remaining language pairs.
- Post-edit distance on actual post-edited content proved to be the most reliable metric in our evaluation. Ranking and Adequacy & Fluency scoring from the same resource was not always consistent. Autoscores (HT) did not correlate with human evaluations in several cases.
NEXT STEPS

We are running several follow-up pilots:

1) Comparing the performance of customized NMT against customized SMT.

2) Comparing Post-edit distance in live production using customized SMT and generic NMT. We would like to see if more extensive production data will confirm our initial findings.

THANK YOU
Neural Machine Translation

- **Neural Machine Translation (NMT)** is an end-to-end learning approach for automated translation, with the potential to overcome many of the weaknesses of conventional phrase-based translation systems.

- The strength of NMT lies in its ability to learn directly, in an end-to-end fashion, the mapping from input text to associated output text.

Neural Machine Translation (GNMT), an end-to-end learning framework that learns from millions of examples, and provided significant improvements in translation quality. It was enabled for eight languages: to and from English and French, German, Spanish, Portuguese, Chinese, Japanese, Korean and Turkish. (1)
"Google Neural Machine Translation (GNMT), an end-to-end learning framework that learns from millions of examples, and provided significant improvements in translation quality. It was enabled for eight languages: to and from English and French, German, Spanish, Portuguese, Chinese, Japanese, Korean and Turkish" (1)

"Microsoft Translator launched Neural Network based translations for all its speech languages. Using the compute power offered by Microsoft’s AI supercomputer and Cognitive Toolkit, the service has released the neural offering across a range of languages: Arabic, Chinese Mandarin, English, French, German, Italian, Brazilian Portuguese, Russian and Spanish – as well as Japanese text" (2)

Neural Machine Translation timeline

September 2016

November 2016

April 2017


© 2017 Autodesk | Localization Solutions
Goal

Assess quality of *Neural MT* versus *Autodesk MT*

Assumptions: MT systems
Assumptions: MT systems
Assumptions: MT systems
Assumptions: MT systems

OLD

• Outdated Moses version
• Lot of Pre/post processing operations
• TCP socket infrastructure
• Not scalable

NEW

• Newer Moses version
• Less pre/post processing - Tokenization and word casing
• REST Api infrastructure
• Scalable

© 2017 Autodesk | Localization Solutions
Assumptions: Products

© 2017 Autodesk | Localization Solutions

Assumptions: Products

© 2017 Autodesk | Localization Solutions
Assumptions: ADSK legacy product

- Human Translation for these products started from the OLD ADSK MT (translation is now post-editing)
- For some portions of Infraworks and Dynamo final Human Translation was then used to retrain the engines ADSK MT, OLD and NEW
- The nature of Autodesk content favors higher matches even on non-trained engines (i.e. Architecture, 3D and so on)
- For these products it looks like there isn't much difference whether an engine was retrained or not, therefore we will not make a distinction in the conclusions

Assumptions: Products

- Cases which shouldn't give any advantage to ADSK MTs
- It was not easy to find content for which we haven't trained our engines. But looking at the results it is clear that we would benefit from more languages at least for the identified content.

For example we don't have such samples for German and Simplified Chinese.
### Assumptions: Scope

<table>
<thead>
<tr>
<th>PRODUCT</th>
<th>CATEGORY</th>
<th>German</th>
<th>French</th>
<th>Spanish</th>
<th>Japanese</th>
<th>Simplified Chinese</th>
<th>Portuguese Brazilian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamo</td>
<td>SW</td>
<td>45k</td>
<td>45k</td>
<td>45k</td>
<td>12k</td>
<td>45k</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DOC</td>
<td>51k</td>
<td>51k</td>
<td>51k</td>
<td>12k</td>
<td>51k</td>
<td></td>
</tr>
<tr>
<td>AUTODESK INFRAWORKS</td>
<td>SW</td>
<td>45k</td>
<td>57k</td>
<td>56k</td>
<td>18k</td>
<td>17k</td>
<td>55k</td>
</tr>
<tr>
<td></td>
<td>DOC</td>
<td>374k</td>
<td>437k</td>
<td>288k</td>
<td>89k</td>
<td>119k</td>
<td>427k</td>
</tr>
<tr>
<td>Knowledge Network</td>
<td>DOC</td>
<td>166k</td>
<td>164k</td>
<td>151k</td>
<td>50k</td>
<td>43k</td>
<td></td>
</tr>
<tr>
<td>AUTODESK 360</td>
<td>DOC</td>
<td>5k</td>
<td>6k</td>
<td>7k</td>
<td>2k</td>
<td>1.5k</td>
<td>6k</td>
</tr>
<tr>
<td>Delcam</td>
<td>DOC</td>
<td>244k</td>
<td>57k</td>
<td></td>
<td></td>
<td></td>
<td>658k</td>
</tr>
<tr>
<td>FreeCAD</td>
<td>DOC</td>
<td>397k</td>
<td>282k</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Used to train ADSK MT

### Approach

© 2017 Autodesk | Localization Solutions
Approach

AUTOMATIC

- Automatic quality evaluation comparing machine’s output and human translation

MANUAL

- Human review, involving internal native speakers and external reviewers
Automatic: mt-eval system

Upload TMX file and additional metadata. This information plus source segment, language and human translation are saved in the DB.
Automatic: mt-eval system

1. Upload TMX file and additional metadata. This information plus source segment, language and human translation are saved in the DB.

2. Machine translate all the entries in the DB. Results are saved in the same DB.

3. The system creates input files reference, hypotheses baseline and hypotheses for each sys. Entries from the translation DB are tokenized, lowercased, space-delimited sentences in UTF-8 encoding, one sentence per line. mt-eval automatically calculate* standard and Autodesk's MT quality metrics.

Ref. *
https://git.autodesk.com/LocalizationServices/multeval
https://github.com/jhclark/multeval
Automatic: mt-eval system

1. Upload TMX file and additional metadata. This information plus source segment, language and human translation are saved in the DB.

2. Machine translate all the entries in the DB. Results are saved in the same DB.

3. The system creates input files reference, hypothesis baseline and hypotheses for each sys. Entries from the translation DB are tokenized, lowercased, space-delimited sentences in UTF-8 encoding, one sentence per line. mt-eval automatically calculate* standard and Autodesk’s MT quality metrics.

4. Use output to create reports.

5. Generate survey for human review activity.

* Ref. https://git.autodesk.com/LocalizationServices/multeval
https://github.com/jhclark/multeval

© 2017 Autodesk | Localization Solutions
Automatic: MT quality metrics

**COMMON**

**BLEU - Bilingual Evaluation Understudy**
- Quality is considered to be the correspondence between a machine’s output and that of a human. The closer a machine translation is to a professional human translation, the better it is [1].

**METEOR - Metric for Evaluation of Translation with Explicit Ordering**
- The metric evaluates translation hypotheses by aligning them to reference translations and calculating sentence-level similarity scores. It uses stemming and synonymy matching, along with the standard exact word matching. The metric was designed to fix some of the problems found in BLEU [2].

**TER - Translation Error Rate**
- A method to determine the amount of Post-Editing required for machine translation jobs. The automatic metric measures the number of actions required to edit a translated segment inline with one of the reference translations [3].

**Length**
- Machine's output length over professional human translation length as a percent. If it is 100%, machine and human translation output have the same length [4].

**CFS - Character-based Levenshtein distance**
- Levenshtein distance on character level

**WFS - Word-based Fuzzy Score**
- Levenshtein distance on word level

**JFS - Joint Fuzzy Score**
- It is a combination of the two above, taking the worse of the two scores for each segment and computing a joined score like this for the whole test set

All three below are based on the Levenshtein distance between the output and the reference translation, the higher the score the better.

Levenshtein distance between two words is the minimum number of single-character edits (i.e. insertions, deletions or substitutions) required to change one word into the other.

Ref:

Manual: Human review rating

**Adequacy**
How much of the meaning expressed in the source is also expressed in the target translation
- **None:** Completely nonsense translation
- **Little:** Sentence preserves some of the meaning of the source sentence but misses significant parts
- **Most:** Sentence retains most of the meaning of the source sentence, but may have some grammar mistakes
- **Everything:** Perfect translation: the meaning of the translation is completely consistent with the source, and the grammar is correct

**Fluency**
Readability and naturalness of the translated text
- **Incomprehensible:** The content is not fluent nor natural in the target language. The translated text is a word by word translation, therefore it is hard to read and understand.
- **Disfluent:** The content reads like it was translated. Some sentence structures don’t seem to be natural in the target language or are not idiomatic. It contains some literal translations.
- **Good:** The content reads like it was originally written in the target language. It uses proper sentence structure and idiomatic expressions. But a few minor improvements might be necessary.
- **Flawless:** The content reads like it was originally written in the target language. It uses proper sentence structure and idiomatic expressions.

*OLD ADSK not rated*
### Manual: Survey

<table>
<thead>
<tr>
<th>Source</th>
<th>Accuracy Score</th>
<th>Frequency Score</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group by pressing SHIFT until you have finished selecting the objects with mouse clicks.</td>
<td>2</td>
<td>2</td>
<td>Group by pressing SHIFT until you have finished selecting the objects with mouse clicks.</td>
</tr>
<tr>
<td>Shows values greater than the condition.</td>
<td>4</td>
<td>4</td>
<td>Shows values greater than the condition.</td>
</tr>
<tr>
<td>If your use of Autodesk software subject to an educational license as part of the apprentice program will be part of commercial projects. The terms and conditions for Autodesk Educational licenses restrict the use exclusively to teaching and exercising activities.</td>
<td>4</td>
<td>4</td>
<td>If your use of Autodesk software subject to an educational license as part of the apprentice program will be part of commercial projects. The terms and conditions for Autodesk Educational licenses restrict the use exclusively to teaching and exercising activities.</td>
</tr>
<tr>
<td>The email was sent but not delivered.</td>
<td>4</td>
<td>4</td>
<td>The email was sent but not delivered.</td>
</tr>
</tbody>
</table>

**Internal ~ 250 segments | External ~ 2500 segments**

© 2017 Autodesk | Localization Solutions

---

### Manual: Survey

<table>
<thead>
<tr>
<th>Source</th>
<th>Accuracy Score</th>
<th>Frequency Score</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group by pressing SHIFT until you have finished selecting the objects with mouse clicks.</td>
<td>2</td>
<td>2</td>
<td>Group by pressing SHIFT until you have finished selecting the objects with mouse clicks.</td>
</tr>
<tr>
<td>Shows values greater than the condition.</td>
<td>4</td>
<td>4</td>
<td>Shows values greater than the condition.</td>
</tr>
<tr>
<td>If your use of Autodesk software subject to an educational license as part of the apprentice program will be part of commercial projects. The terms and conditions for Autodesk Educational licenses restrict the use exclusively to teaching and exercising activities.</td>
<td>4</td>
<td>4</td>
<td>If your use of Autodesk software subject to an educational license as part of the apprentice program will be part of commercial projects. The terms and conditions for Autodesk Educational licenses restrict the use exclusively to teaching and exercising activities.</td>
</tr>
<tr>
<td>The email was sent but not delivered.</td>
<td>4</td>
<td>4</td>
<td>The email was sent but not delivered.</td>
</tr>
</tbody>
</table>

**Internal ~ 250 segments | External ~ 2500 segments**

© 2017 Autodesk | Localization Solutions
Manual: Survey

Results: Automatic
Results: Automatic

ADSK legacy product

ADSK MTs are better than Neural, which matches the assumptions

© 2017 Autodesk | Localization Solutions

*METEOR only for FR and DE – not in the graph*
Results: Automatic

ADSK legacy product

ADSK MTs are better than Neural, which matches the assumptions

ADSK new product or External product

Google NMT is best in all cases

* METEOR only for FR and DE – not in the graph
Results: Manual

Average of Adequacy Score
Average of Fluency Score

ADSK legacy product

© 2017 Autodesk | Localization Solutions
Human Translation is always best,
Google NMT is always second

Results: Manual
ADSK legacy product

Human Translation is always best,
Google NMT is always second

Results: Manual
ADSK new product or External product

© 2017 Autodesk | Localization Solutions
Results: Manual

ADSK legacy product

Human Translation is always best, Google NMT is always second

ADSK new product or External product

Google NMT is very close to Human, sometimes surpassing

Average of Adequacy Score
Average of Fluency Score

Average of Adequacy Score
Average of Fluency Score

© 2017 Autodesk | Localization Solutions

Conclusions

- Commercial Neural MT are **viable**
- **Moses Engines** are still useful on legacy products
- Next Steps:
  - Explore **Open source solutions** (i.e. OpenNMT)
  - Use the best **MT system** that matches current context (i.e. product, language, content type, etc.)
Result: Breakdown

<table>
<thead>
<tr>
<th>Approach</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ADSK legacy product</strong></td>
<td><strong>AUTOMATIC</strong></td>
</tr>
<tr>
<td></td>
<td><em>NEW and OLD ADSK MTs are clearly better than Neural - which matches the assumptions</em></td>
</tr>
<tr>
<td></td>
<td><em>NEW and OLD ADSK MTs tend to have very similar results, except for German</em></td>
</tr>
<tr>
<td></td>
<td><em>Between Neural MTs, only Japanese shows better results with Microsoft than Google</em></td>
</tr>
<tr>
<td><strong>MANUAL</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Human Translation is always best except one case only for Portuguese where Google Neural is a little bit better</em></td>
</tr>
<tr>
<td></td>
<td><em>Google Neural is always second</em></td>
</tr>
<tr>
<td></td>
<td><em>Hard to say whether ADSK or Microsoft are best, it varies between languages but globally they are quite a bit lower than the others and close together</em></td>
</tr>
<tr>
<td><strong>ADSK new product or External product</strong></td>
<td><strong>AUTOMATIC</strong></td>
</tr>
<tr>
<td></td>
<td><em>Google Neural tends is best in all cases except Japanese</em></td>
</tr>
<tr>
<td></td>
<td><em>For Japanese Microsoft Neural is the best</em></td>
</tr>
<tr>
<td></td>
<td><em>Neural is better than ADSK MT, NEW and OLD</em></td>
</tr>
<tr>
<td><strong>MANUAL</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Google Neural is very close to Human, sometimes surpassing</em></td>
</tr>
<tr>
<td></td>
<td><em>Microsoft and ADSK are often close alternating third position</em></td>
</tr>
<tr>
<td></td>
<td><em>For OPENOFFICE we had to ignore Human Translation scores</em></td>
</tr>
</tbody>
</table>
### Breakdown: ADSK legacy product (1/2)

<table>
<thead>
<tr>
<th>Language</th>
<th>Approach</th>
<th>Ranking</th>
<th>Notes</th>
</tr>
</thead>
</table>
| German   | AUTOMATIC| 1. NEW ADSK  
           |          | 2. OLD ADSK  
           |          | 3. Google Neural / Microsoft Neural |
|          | MANUAL   | 1. Human Translation  
           |          | 2. Google Neural  
           |          | 3. Microsoft Neural |
|          |          | 4. NEW ADSK |
| German   | AUTOMATIC| 1. NEW ADSK  
           |          | 2. OLD ADSK  
           |          | 3. Google Neural / Microsoft Neural |
|          | MANUAL   | 1. Human Translation  
           |          | 2. Google Neural  
           |          | 3. Microsoft Neural |
|          |          | 4. NEW ADSK |

### Breakdown: ADSK legacy product (2/2)

<table>
<thead>
<tr>
<th>Language</th>
<th>Approach</th>
<th>Ranking</th>
<th>Notes</th>
</tr>
</thead>
</table>
| Portuguese| AUTOMATIC| 1. NEW ADSK / OLD ADSK  
           |          | 2. Google Neural  
           |          | 3. Microsoft Neural |
|          | MANUAL   | Adequacy  
           |          | Human is best  
           |          | Google Neural second quite a bit lower |
|          |          | Fluency   
           |          | Google Neural is best, Human is close |
| Japanese | AUTOMATIC| 1. NEW ADSK / OLD ADSK  
           |          | 2. Google Neural  
           |          | 3. Microsoft Neural |
|          | MANUAL   | 1. Human Translation  
           |          | 2. Google Neural  
           |          | 3. Microsoft Neural |
|          |          | 4. NEW ADSK |

© 2017 Autodesk | Localization Solutions

Proceedings of MT Summit XVI, Vol.2: Users and Translators Track  
Nagoya, Sep. 18-22, 2017 | p. 203
# ADSK legacy product: Trained VS Not-Trained

<table>
<thead>
<tr>
<th>Approach</th>
<th>TRAINED: DYNAMO (SW), INFRAWORKS (SW/DOC)</th>
<th>NOT-TRAINED: DYNAMO (DOC), AKN (DOC), ADSK MIX (DOC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MANUAL</td>
<td>*Human Translation is always best&lt;br&gt;•Google Neural is second in most of the languages&lt;br&gt;•NEW ADSK is close to or a little bit better than Google Neural in French, Spanish and Portuguese&lt;br&gt;•Microsoft Neural is worst in most of the languages except Japanese and German Fluency</td>
<td>*Human Translation is always best except Portuguese where Google Neural is best&lt;br&gt;•Google Neural is second and close to Human Translation in most of the languages&lt;br&gt;•Microsoft Neural is third in most of the languages except Spanish&lt;br&gt;•NEW ADSK is worst not far away from Microsoft Neural</td>
</tr>
<tr>
<td>AUTOMATIC</td>
<td>•NEW ADSK is always best&lt;br&gt;•OLD ADSK is always second except Japanese&lt;br&gt;•Google Neural and Microsoft Neural are close in most of the languages except Simplified Chinese where Google is clearly better than Microsoft Neural&lt;br&gt;•Japanese where Microsoft Neural is clearly better than Google Neural</td>
<td>•OLD ADSK is always best&lt;br&gt;•NEW ADSK is always second except CFS in Japanese and Simplified Chinese&lt;br&gt;•Google Neural is third&lt;br&gt;•Microsoft Neural is fourth, very close to Google Neural in most of the languages</td>
</tr>
</tbody>
</table>

---

### Breakdown: ADSK new product or External product (1/2)

<table>
<thead>
<tr>
<th>Product</th>
<th>Language</th>
<th>Approach</th>
<th>Ranking</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>DELCAM</td>
<td>French</td>
<td>AUTOMATIC</td>
<td>1. Google Neural / Microsoft Neural&lt;br&gt;2. NEW ADSK&lt;br&gt;3. OLD ADSK</td>
<td>*Google Neural and Microsoft Neural are the best and very close&lt;br&gt;NEW ADSK is a bit lower than Neural, and quite a bit better than OLD ADSK</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MANUAL</td>
<td>1. Human Translation&lt;br&gt;2. Google Neural&lt;br&gt;3. Microsoft Neural&lt;br&gt;4. NEW ADSK</td>
<td>*Human is best&lt;br&gt;Google is second not too far from Human&lt;br&gt;Microsoft Neural is third quite a bit lower&lt;br&gt;NEW ADSK last not too far from Microsoft Neural</td>
</tr>
<tr>
<td></td>
<td>Japanese</td>
<td>AUTOMATIC</td>
<td>1. Microsoft Neural&lt;br&gt;2. Google Neural&lt;br&gt;3. Human Translation / Microsoft Neural&lt;br&gt;4. NEW ADSK</td>
<td>*Google Neural is the best and quite a bit better than Google Neural&lt;br&gt;NEW and OLD ADSK are lower and very close</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MANUAL</td>
<td>1. Google Neural&lt;br&gt;2. Human Translation / Microsoft Neural&lt;br&gt;3. NEW ADSK</td>
<td>*Google Neural is best&lt;br&gt;Followed by Human and Microsoft Neural being very close together&lt;br&gt;NEW ADSK last a bit lower</td>
</tr>
<tr>
<td></td>
<td>Portuguese</td>
<td>AUTOMATIC</td>
<td>1. Google Neural&lt;br&gt;2. Microsoft Neural&lt;br&gt;3. NEW ADSK / OLD ADSK</td>
<td>*Google Neural is the best&lt;br&gt;Google and Microsoft Neural are the best and close&lt;br&gt;NEW and OLD ADSK are lower and very close</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MANUAL</td>
<td>1. Human Translation&lt;br&gt;2. Google Neural&lt;br&gt;3. Microsoft Neural&lt;br&gt;4. NEW ADSK</td>
<td>*Human is best, Google Neural second but very close&lt;br&gt;Fluency: *Opposite, Google Neural best with Human very close&lt;br&gt;Third is Microsoft Neural followed closely by NEW ADSK</td>
</tr>
</tbody>
</table>
## Breakdown: ADSK new product or External product (2/2)

<table>
<thead>
<tr>
<th>Product</th>
<th>Language</th>
<th>Approach</th>
<th>Ranking</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>OPENOFFICE</td>
<td>French</td>
<td>AUTOMATIC</td>
<td>1. Google Neural</td>
<td>*Google Neural is the best</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Microsoft Neural</td>
<td>*Google Neural and Microsoft Neural are the best and close</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. NEW ADSK</td>
<td>*NEW and OLD ADSK are lower and close</td>
</tr>
<tr>
<td></td>
<td>MANUAL</td>
<td></td>
<td>1. Google Neural</td>
<td>*Google Neural is best</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. Microsoft Neural</td>
<td>*Microsoft Neural is second</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. NEW ADSK</td>
<td>*NEW ADSK last not too far</td>
</tr>
<tr>
<td>Japanese</td>
<td>AUTOMATIC</td>
<td></td>
<td>1. Microsoft Neural (except BLEU)</td>
<td>*Microsoft Neural is the best except for BLEU where OLD ADSK wins</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. OLD ADSK</td>
<td>*OLD ADSK is generally higher than Google Neural</td>
</tr>
<tr>
<td></td>
<td>MANUAL</td>
<td></td>
<td>1. Google Neural / Microsoft Neural</td>
<td>*Google Neural and Microsoft Neural are best very close</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. NEW ADSK</td>
<td>*Adequacy: Microsoft Neural a little better, opposite for Fluency</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*NEW ADSK is quite a bit lower</td>
</tr>
<tr>
<td>Spanish</td>
<td>AUTOMATIC</td>
<td></td>
<td>1. Google Neural</td>
<td>*Google Neural is the best</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. OLD ADSK / NEW ADSK / Microsoft Neural</td>
<td>*The rest is lower and quite similar results</td>
</tr>
<tr>
<td></td>
<td>MANUAL</td>
<td></td>
<td>1. Google Neural</td>
<td>*Google Neural is best</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. NEW ADSK</td>
<td>*NEW ADSK is second</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Microsoft Neural</td>
<td>*Microsoft Neural is last</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*All very close</td>
</tr>
</tbody>
</table>
A Reception Study of Machine Translated Subtitles for MOOCs

Ke Hu, Sharon O’Brien, Dorothy Kenny
ADAPT Centre, SALIS, Dublin City University

Overview

§ Context
  § Why MOOCs?
  § Why MT?
  § Why reception?
§ Research question and hypothesis
  § Methodologies
  § Reception model
  § Sub-hypotheses
§ Pilot study
§ What’s next?
Why MOOCs?

<table>
<thead>
<tr>
<th>Year</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>“MOOC” coined by Dave Cormier (2008, online) Massive Open Online Courses E.g.: Coursera, Udacity, edX...</td>
</tr>
<tr>
<td>2013</td>
<td>Coursera has over 30 university partners, 2.8 million registered students, 1.4 million course enrolments every month (Cusumano, 2013)</td>
</tr>
</tbody>
</table>

**In China**

<table>
<thead>
<tr>
<th>Year</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early 2013</td>
<td>Ø Chinese universities started to join MOOCs Ø 4 universities joined edX, 6 universities joined Coursera</td>
</tr>
<tr>
<td>2014</td>
<td>Ø 2 universities joined FutureLearn Ø Over 50 MOOCs offered by Chinese universities on international platforms (Yuan &amp;Liu, 2014)</td>
</tr>
<tr>
<td>Now</td>
<td>Around 20 Chinese MOOC platforms (unclear)</td>
</tr>
</tbody>
</table>

Developed by Tsinghua University, largest Chinese MOOC platform, offers 504 MOOCs to 1,290,000 registered students from 126 countries (Ma, 2015)

Why MT?

Survey by MOOC学院 (mooc.guokr.com) in 2014:

- 3300 responses
- 74% are users, 47% of users
- 26% are non-users, 17.5% of non-users

Language was a barrier to learning via MOOCs
Why reception?

“It is not the software but the human side of the implementation cycle that will block progress in seeing that delivered systems are used effectively.”

-- Peter G. W. Keen (1991:1249)

Questions:

What are the needs of MT users?
What can affect user experience of MT?
How well do end users receive MT content?
...

Machine Translation!!
Research question and hypothesis

**Main research question:**

Is there a difference in reception between participants who are offered raw MT subtitles and those who are offered full PEMT subtitles?

**Main hypothesis:**

Participants who are offered full PEMT subtitles will score higher on our reception metrics compared with those who are offered raw MT subtitles.

Methodologies

**A mixed-methods approach**

- Quantitative method → Eye-tracking → Objective validity
- Qualitative method → Questionnaires → Subjective experience

Sufficient data → Robust method
### Reception model

<table>
<thead>
<tr>
<th>Element</th>
<th>Related to</th>
<th>Reflected in</th>
<th>Measured by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response</td>
<td>Perceptual decoding</td>
<td>Attentional processes</td>
<td>Eye-tracking</td>
</tr>
<tr>
<td>Reaction</td>
<td>Psycho-cognitive issue</td>
<td>Processing effort and comprehension</td>
<td>Eye-tracking and comprehension testing</td>
</tr>
<tr>
<td>Repercussion</td>
<td>Attitudinal issues and sociocultural dimensions</td>
<td>Attitudes and beliefs</td>
<td>Background survey and attitude questions</td>
</tr>
</tbody>
</table>

Based on Gambier’s model (Gambier, 2009)

### Sub-hypotheses

**Response:**

**Hypothesis 1:** Fewer subtitles are skipped when participants are watching full PEMT subtitles. *(measured by visit count)*

**Hypothesis 2:** Relatively more attention is allocated to the image area when full PEMT subtitles are displayed than when raw MT subtitles are displayed. *(measured by fixation count and visit duration)*

**Reaction:**

**Hypothesis 3:** Comprehension score is higher with full PEMT subtitles. *(measured by comprehension testing)*

**Hypothesis 4:** Mean fixation duration is shorter when full PEMT subtitles are displayed.

**Repercussion:**

**Hypothesis 5:** Attitudes toward machine translation are better among participants shown full PEMT subtitles. *(measured by attitude questions)*
Pilot study

Ø DCU, May 2017
Ø Video: “What is physical activity?” (6”59’) under the MOOC “Sit Less, Get Active” on Coursera.
Ø MT system: Google Translate (EN-ZH)
Ø Two versions of subtitles (Number: 114 & 115)

Raw MT subtitles (BLEU: 42.05%) VS Full PEMT subtitles (TER: 19.69%)

Participants

Ø MOOCs: university students, 18-25 years old
Ø China: 50.94 out of 100, English Proficiency Index 2016 by EF
Ø Ideal participants: Chinese undergraduates with low English level

Ø Four Chinese participants (two groups)

<table>
<thead>
<tr>
<th>Gender</th>
<th>1 female, 3 male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>22-33</td>
</tr>
<tr>
<td>Education</td>
<td>2 PhD students, 1 Post-doc, 1 final-year undergraduate</td>
</tr>
<tr>
<td>English level</td>
<td>1 intermediate</td>
</tr>
<tr>
<td></td>
<td>3 upper intermediate</td>
</tr>
</tbody>
</table>
Step 1: Pre-recruitment questionnaire & Online English test (Cambridge English Language Assessment)

Step 2: Watching MOOC video with eye-tracker (SMI REDn Scientific)

Step 3: Post-task questionnaire: comprehension testing (multiple choice) and attitude survey (five-point Likert scale)

Results

😢 All hypotheses were NOT supported by the results.
      —— Tiny sample

😢 A few questions could be answered by common sense.
      —— Questionnaire needs to be modified

😢 Vagaries of participants’ memories and concentration.
      —— Irresistible force
Main experiment in China!

**Larger sample:** over 30 Participants  
**One more group:** human translated text added!  
**Statistical methods:** ANOVA and t-test

Ke Hu: [ke.hu2@mail.dcu.ie](mailto:ke.hu2@mail.dcu.ie)  
Sharon O’Brien: [sharon.obrien@dcu.ie](mailto:sharon.obrien@dcu.ie)  
Dorothy Kenny: [dorothy.kenny@dcu.ie](mailto:dorothy.kenny@dcu.ie)
Recent Developments

Joss Moorkens & Yota Georgakopoulou
MT Summit XVI

Table of contents

• The TraMOOC Project
• NMT systems for TraMOOC
• Comparative Evaluation of Neural MT and Phrase-Based SMT
• Crowdsourced evaluations (explicit & implicit)
• Task-based evaluations

Joss Moorkens, Sheila Castilho, Federico Gaspari, Andy Way (DCU/ADAPT) – Ireland
Yota Georgakopoulou, Maria Gialama (Deluxe Media) – Greece/United Kingdom
Rico Sennrich, Antonio Valerio Miceli Barone (University of Edinburgh) - United Kingdom
Valia Kordoni, Markus Egg, Maja Popović (Humboldt University of Berlin) - Germany
Vilelmini Sosoni (Ionian University, Corfu) - Greece
Iris Hendrickx (Radboud University Nijmegen) – The Netherlands
Menno van Zaanen (Tilburg University) – The Netherlands
**Our Project**

- **Reliable Machine Translation (MT) for Massive Open Online Courses (MOOCs)**
- The main expected outcome is a **high-quality semi-automated machine translation service** for educational text data on a MOOC platform
- Open educational platform for MT and a replicable process for creating such a service

**Our Project**

- Create domain-specific **SMT NMT engines** – 3 iterations
- Crowdsourced evaluation of MT quality
- Explicit and implicit evaluation stages
- Task-based evaluations
- Free and premium platform due 2018
Make existing monolingual educational material available to speakers of other languages
  - multi-genre and heterogeneous textual course material
  - Subtitles – video lectures
  - assignments
  - tutorial text
  - social web text posted on MOOC blogs and fora (questions/answers/comments)

Reusing existing linguistic infrastructure and MT resources extending existing models

Test on a MOOC platform and on the VideoLectures.Net digital video lecture library

The Target Audience

Users who want access to open online education that is not constrained by language barriers.

MOOC providers, who wish to offer high-quality, integrated multilingual educational services.

Machine Translation developers, who need a platform for promoting, testing and comparing their solutions.

Language Technology Engineers, who want access to accurate and wide-coverage linguistic infrastructure, even for less widely spoken languages.
The Consortium

• 10 partners from 6 European countries
  o Humboldt University (Coordinator)
  o Dublin City University
  o University of Edinburgh
  o Ionian University
  o Radboud University
  o Tilburg University
  o Deluxe Media Europe LTD
  o Knowledge 4 All Foundation LTD
  o EASN Technology Innovation Services
  o (Iversity) HPI

Which MT paradigm?

• Project had originally planned to compare Syntax-Based and Phrase-Based SMT
• Comparative Evaluation of Neural MT (Nematus) and Phrase-Based SMT (Moses)
• English to German, Greek, Portuguese, and Russian
• MT engines trained on open and educational data
Main strength of NMT is grammatical improvements, but possible degradation in lexical transfer (Neubig, Morishita, Nakamura 2015).

Output conditioned on full source text and target history.

Some problems:
- Networks have fixed vocabulary → poor translation of rare/unknown words.
- Models are trained on parallel data; how do we use monolingual data?
- Recent solutions:
  - Subword models allow translation of rare/unknown words (Sennrich, Birch, Haddow 2016a)
  - Train on back-translated monolingual data (Sennrich, Birch, Haddow 2016b)
NMT vs. PB-SMT

- 4 datasets (250 segments) from EN MOOC data translated into German, Greek, Portuguese, and Russian using TraMOOC engine prototype 2
- PB-SMT/NMT mixed, random task order
- 2-4 professional translators in Deluxe Media
- Detailed results presented by Sheila Castilho in Research Track and in proceedings of MT Summit XVI

Machine Translation Systems

- PBSMT
  - Moses, MGIZA is used to train word alignments, and KenLM is used for language model training and scoring (Huck and Birch 2015)
- NMT Tools Used:
  - Nematus: https://github.com/rsennrich/nematus
  - Amun: https://github.com/amunmt/amunmt (for deploying the models)
- Domain adaptation:
  - Models initially trained on all available data, then continually trained on in-domain data, which effectively adapts the system to the domain NMT
## NMT/SMT Fluency

- For all 4 language pairs:

<table>
<thead>
<tr>
<th>FLUENCY</th>
<th>EN-DE</th>
<th>EN-EL</th>
<th>EN-PT</th>
<th>EN-RU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. No fluency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Little fluency</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Near native</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Native</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| % scores assigned 3-4 fluency value (SMT, NMT) | 54.2 | 67.6 | 65 | 75 | 73.8 | 79.5 | 60.2 | 75.1 |
| % scores assigned 1-2 fluency value (SMT, NMT) | 45.8 | 32.4 | 35 | 25 | 26.2 | 20.5 | 39.8 | 24.9 |

---

## NMT/SMT Adequacy

- For all 4 language pairs:

<table>
<thead>
<tr>
<th>ADEQUACY</th>
<th>EN-DE</th>
<th>EN-EL</th>
<th>EN-PT</th>
<th>EN-RU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. None of it</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Little of it</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Most of it</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. All of it</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| % scores assigned 3-4 adequacy value (SMT, NMT) | 73.5 | 66.4 | 89 | 89 | 94.7 | 97.1 | 72.8 | 77.5 |
| % scores assigned 1-2 adequacy value (SMT, NMT) | 26.5 | 33.6 | 11 | 11 | 5.3 | 2.9 | 27.2 | 22.5 |
### NMT/SMT PE Temporal Effort

<table>
<thead>
<tr>
<th>Words per second (all PEs)</th>
<th>SMT</th>
<th>NMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>0.21</td>
<td>0.22</td>
</tr>
<tr>
<td>Greek</td>
<td>0.22</td>
<td>0.24</td>
</tr>
<tr>
<td>Portuguese</td>
<td>0.29</td>
<td>0.30</td>
</tr>
<tr>
<td>Russian</td>
<td>0.14</td>
<td>0.14</td>
</tr>
</tbody>
</table>

Previous work by Moorkens & O’Brien (2015) found an average speed of 0.39 WPS for EN-DE professional PE.

<table>
<thead>
<tr>
<th>SMT, NMT</th>
<th>German</th>
<th>Greek</th>
<th>Portuguese</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>POST-EDITED SENTENCES (CHANGED)</td>
<td>940</td>
<td>813</td>
<td>928</td>
<td>874</td>
</tr>
<tr>
<td>UNCHANGED SMT, NMT</td>
<td>60</td>
<td>187</td>
<td>72</td>
<td>137</td>
</tr>
</tbody>
</table>

### NMT/SMT Summary

- In this study, using these language pairs, in this domain...
- Fluency is improved, word order errors are fewer using NMT
- Fewer segments require editing using NMT
- NMT produces fewer morphological errors
- No clear improvement for omission or mistranslation using NMT
- NMT for production: no great improvement in post-editing throughput
  - “Errors are more difficult to spot”
- Based on the pace of improvement of NMT however, TraMOOC moved to NMT exclusively
Underway: Crowdsourced explicit evaluations

Using the Crowdflower platform for all 11 language pairs:

- **Clear instructions available during the entire translation procedure.**
- **Test Questions to validate the accuracy of the participants’ input.**
- **Post-editing question should be displayed first, hiding the rest of the questions to avoid influencing the contributors’ judgment.**
- **Fluency for ST and TT, adequacy and error mark-up for TT**
- **Multiple error mark-up supported.**

For QA and language coverage, 5-10% expert evaluation by DME

Crowdsourced explicit evaluations: post-editing

- **Post-editing (expert and crowd): “Make changes in the translation if there are errors in grammar, meaning or spelling”**
  - Basic rules regarding spelling apply. If there are any typos or slight grammatical/syntactic mistakes in the original, please fix them in the translation
  - Do not implement corrections for stylistic reasons only
  - No need to restructure sentences only to improve the natural flow of the text

---

**Joss Moorkens & Yota Georgakopoulou**

Proceedings of MT Summit XVI, Vol.2: Users and Translators Track

Nagoya, Sep. 18-22, 2017 | p. 222
Crowdsourced explicit evaluations: error annotation

- Change the mark-up error type list (for expert group) so as to map onto DQF-MQM typology: **Addition, Mistranslation, Omission, Untranslated, Function Words, Word Form, and Word Order**

Crowdsourcing – Issues & solutions

<table>
<thead>
<tr>
<th>Crowd behaviour issue</th>
<th>Solution(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malicious behaviour</td>
<td>Constant monitoring, manual and automated</td>
</tr>
<tr>
<td>Use of Google Translate</td>
<td>Source language text is an image. Manual check with Google Translate feature in Chrome.</td>
</tr>
<tr>
<td>BR performing EU-PT tasks</td>
<td>Target specific countries</td>
</tr>
<tr>
<td>No change, yet low score on quality</td>
<td>Popup alerts</td>
</tr>
<tr>
<td>Poor coverage/ low contributor flow</td>
<td>Increase HIT payment; expand geographical reach &amp; channel; decrease contributor level; decrease text question difficulty</td>
</tr>
</tbody>
</table>
### Malicious behaviour | Solutions
---|---
Blank translations | Change tactics for test questions, binary evaluation answers, distributed randomly
Random symbols | Increase the minimum time per page
Repetitive answers | Increase contributors’ level
Other language characters | Constant manual and script-based (automated) monitoring: Python scripts for blanks, Latin characters in non-Latin languages, etc.
Multiple malicious accounts | Customised alerts scripts (blanks, length, time per page, etc.); flag malicious contributors; ban specific channels

---

**Implicit evaluation**: Annotation of entities, topics and terms in the source and target texts

- Generate a thesaurus of tag-sets that allows for the **implicit evaluation** of MT output through the comparison of the source and target tag-sets

**Activities**:
1. Entity annotation via Wikification
2. Topic & sentiment annotation
To come: Task-based evaluations

- openHPI - European MOOC platform plus TraMOOC API
  - Launched by the Hasso Plattner Institute (HPI) for Digital Engineering in Potsdam, Germany
- Users will be able to switch between the original course language and automatically translated content
- Users will be able to request translation for specific forum contributions
- Feedback via surveys on the translation content and the integration of the translation tools into the openHPI platform

Joss Moorkens & Yota Georgakopoulou

Nagoya, Sep. 18-22, 2017 | p. 225