AMTA
Association for Machine Translation in the Americas

October 30 - November 3, 2015
MIAMI, FLORIDA, USA

Machine Translation Summit XV

http://www.amtaweb.org/mt-summit-xv

Proceedings of MT Summit XV

Vol. 2: MT Users' Track

Steve Richardson, Mike Dillinger, Jen Doyon,
& Patti O'Neill-Brown, Editors

MT Summit XV

October 30 – November 3, 2015 -- Miami, FL, USA

Proceedings of

MT Summit XV,
Vol. 2: MT Users' Track

Steve Richardson, Mike Dillinger, Jen Doyon & Patti O'Neill-Brown, Eds.

Association for Machine Translation in the Americas
http://www.amtaweb.org

Introduction

The Commercial MT Users and Translators Track at MT Summit XV features presentations from various groups within the translation and language technology industry including Language Service Providers, commercial machine translation technology developers, and a wide range of machine translation practitioners coming from organizations and enterprises worldwide. This year's presentations reflect the evolving variety of MT applications and usage scenarios, with heavy emphasis on the evaluation of MT quality and resulting productivity improvements. Industry experts and MT enthusiasts will cover topics relevant to all involved in the rapidly growing adoption of MT across industries, including fast and effective ways of adapting machine translation engines to specific domains, assessing the final quality of the post-edited content, analyzing and encouraging acceptance of MT by professional translators, developing and deploying machine translation engines for less common languages, integrating terminology systems and processes with MT, and applications of speech translation. Presentations on these and other topics will demonstrate how essential machine translation technology has become for the success of the localization and translation programs around the world.

The Commercial MT Users and Translators Track Co-Chairs

Steve Richardson
Mike Dillinger

The 2015 MT Summit Government MT Users Track brings together government, commercial and academic translation and language technology experts from around the world to present their work in the areas of design, development, enhancement, integration, evaluation and use of machine translation, translation memory and terminology management data resources and engines. This year's presentations reflect governments' constant struggle to produce accurate translations of untold amounts of foreign language data in a timely and cost-effective manner with a shortage of human translators. Presenters will share how they have integrated various translation tools and post-editing techniques into government environments and translation workflows; developed and enhanced translation engines to specific domains and other topics that demonstrate how essential machine translation technology has become for the success of government missions.

The Government MT Users Program Chair

Jen Doyon

Contents

Page

1 MT at NetApp – This is how we do it

Edith Bendermacher, Pablo Vazquez

11 Machine Translation in Mobile Games – Augmenting Social Media Text Normalization with Incentivized Feedback

Nikhil Bojja, Arun Nedunchezhian, Pidong Wang

17 Why are we (still) waiting? What premium translators need to use MT effectively.

Robin Bonthrone, Konstantin Lakshin

19 Machine Translation and Terminology Management

Jennifer DeCamp

20 MT Quality Estimation for E-Commerce Data

José G. C. de Souza, Marcello Federico, Hassan Sawaf

30 Yandex.Translate approach to the translation of Turkic languages

Irina Galinskaya, Farkhat Aminov

31 Solving Specific Content Challenges with Flexible Machine Translation

Quinn Lam

65 Quality Evaluation of Four Translations of a Kidney Document: focus on reliability

Alan K. Melby

69 A Survey of Usage Environment of Machine Translation by Professional Translators

Tomoki Nagase, Tatsuhiro Kudoh, Katsunori Kotani, Wenjun Ye, Takeshi Mori, Yoshiyuki Sakamoto, Nobutoshi Hatanaka, Takamitsu Takeda, Shu Hirata, Hiromi Nakaiwa

92 Machine Translation for enterprise technical communications – a journey of discovery

Morgan O'Brien, Ana Duarte

106 MT Quality Evaluations: From Test Environment to Production

Elaine O'Curran

144 Enterprise Application of MT: Progress and Challenges

Craig Plesco, Nestor Rychtyckyj

146 Machine Translation Quality Estimation - A Linguist's Approach

Juan Rowda

163 Industry Shared Metrics with the TAUS Dynamic Quality Dashboard and API

Achim Ruopp

192 Accurately Predicting Post-editing Time & Labor for Cost-Management

Carla Schelfhout

199 Adjusting Interaction Levels in a Speech Translation System for Healthcare

Mark Seligman, Mike Dillinger

219 Beyond Text, Machine Translation and NLP for e-discovery

Jean Senellart, Denis Gachot, Joshua Johanson

220 Designing User Experience for Machine Translated Conversations

Tanvi Surti

224 How Much Cake is Enough: The Case for Domain-Specific Engines

Alex Yanishevsky

248 Productivity Promotion Strategies for Collaborative Translation on Huge-Volume Technical Documents

Guiping Zhang, Na Ye, Fang Cai, Chuang Wu, Xiangkui Sun, Jinfu Yuan, Dongfeng Cai

Government MT Users

259 A Machine Assisted Human Translation System for Technical Documents

Vishwajeet Kumar, Ashish Kulkarni, Pankaj Singh, Ganesh Ramakrishnan and Ganesh Arnaal

MT at NetApp – This is how we do it

Edith Bendermacher
Pablo Vázquez

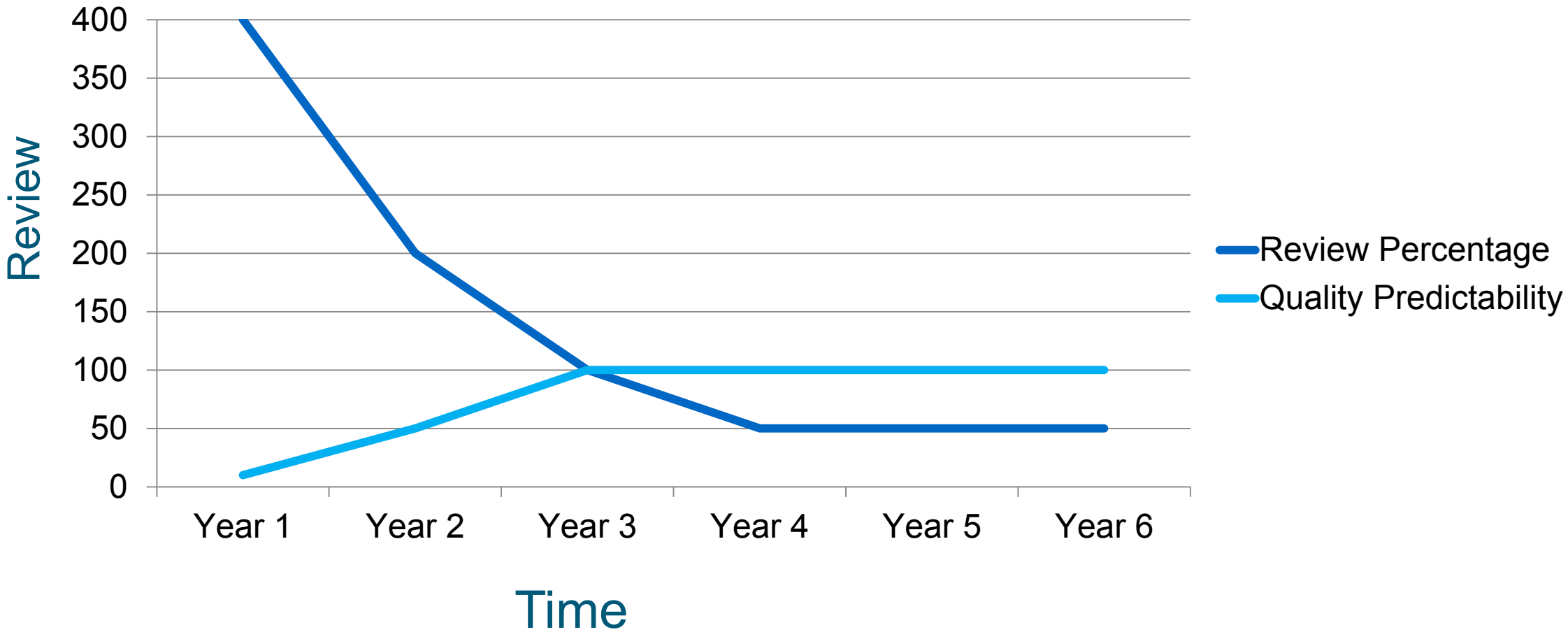
October 2015

Agenda

1) The journey of quality at NetApp
2) How do we leverage MT at NetApp - NetApp's Content Classification model
3) MT infrastructure and Post Edit process
4) NetApp's QA Process
5) Automation

NetApp

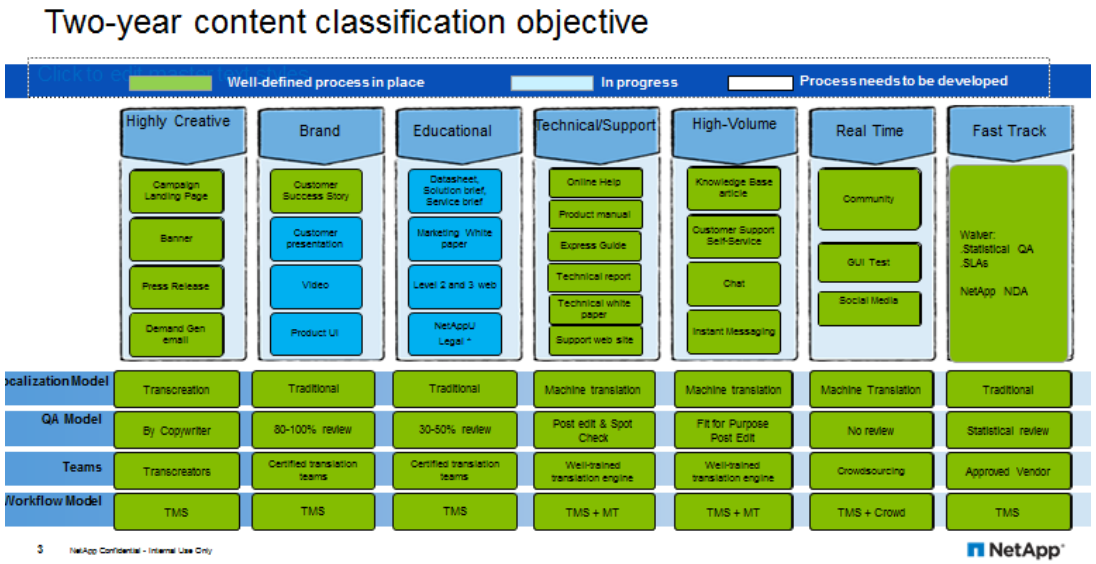
The history of NetApp and its quality expectation



NetApp

Why do we need MT at NetApp - NetApp's Content Classification model

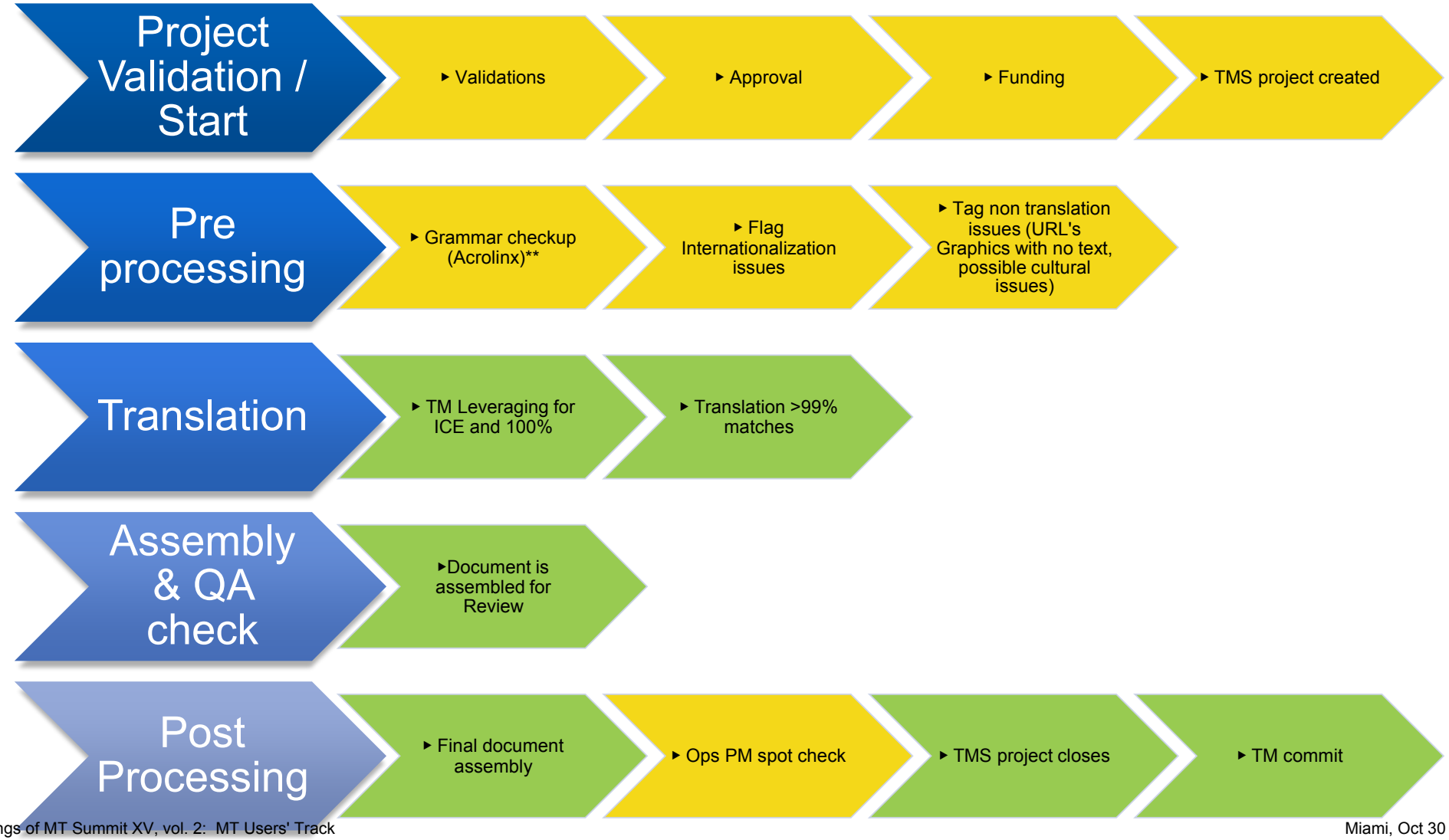
1) NetApp has different types of content
2) Not all content is created equally and requires same processing
3) Product manuals can leverage TMs better then highly creative marketing content
4) Process more for less



Decision matrix:

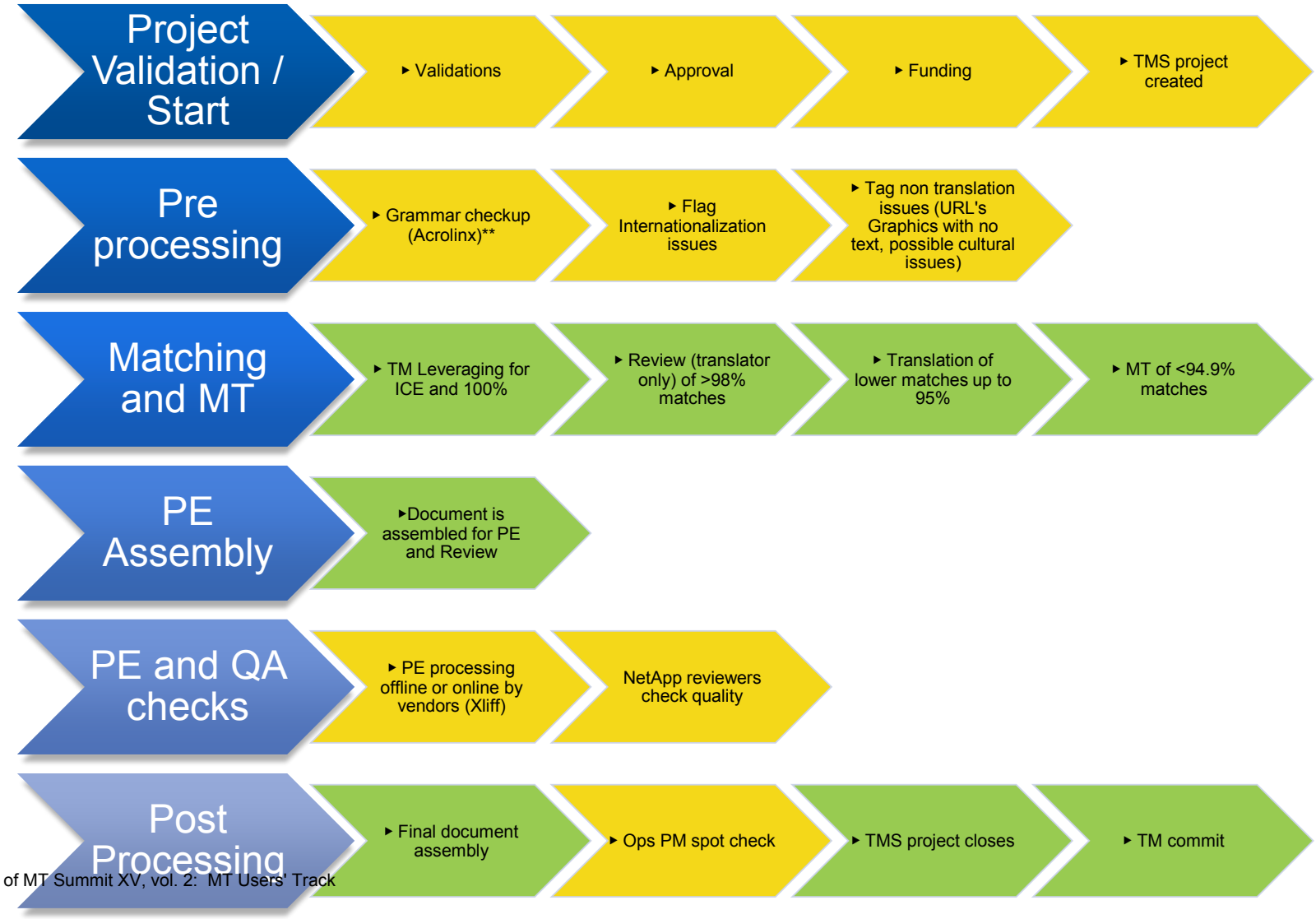
Content survey	Is OK to RAW MT	Is OK to PE	Is HT
Content repository	Yes	Yes	Yes
File format and pre and post processing needs	Only if no DTP is needed	Only if is a small DTP effort	All levels of DTP
Languages	All* (7 Trained)	7 Languages	12 Core languages
MT engine quality (based on TMs leveraging) and reuse	Only High confidence	7 languages	All
Content type based on the content classification	Technical, low visibility, low traffic	Any technical	High visibility, banners, High touch Marketing, etc.,

Traditional workflows



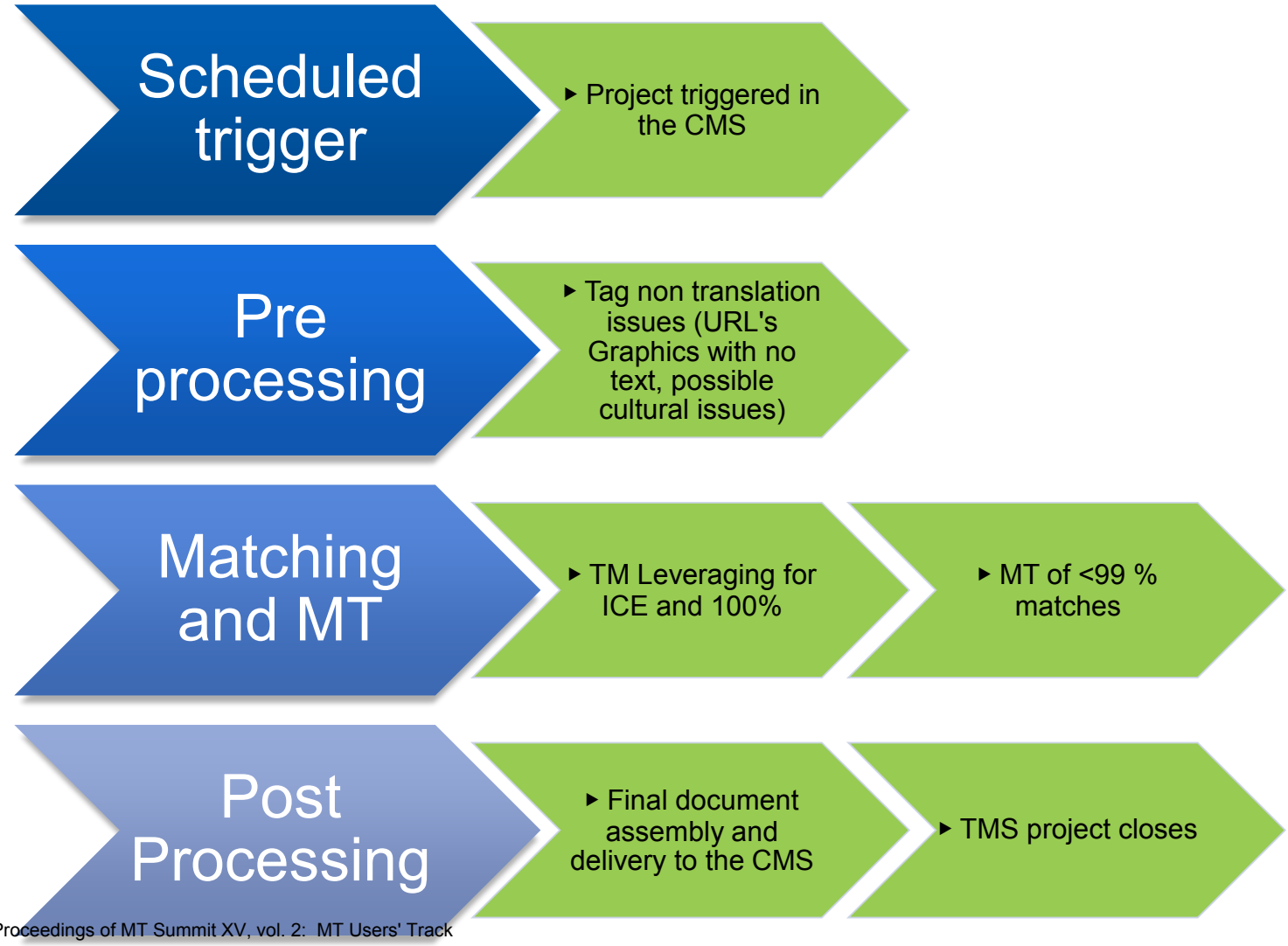
NetApp

PE workflows



NetApp

Raw MT Process



NetApp

NetApp’s QA model

1) To comply with NetApp’s high quality standards, additional step was added after PE
2) NetApp’s GCMs conduct review and compile feedback
3) Feedback is being categorized and submitted to engine training team for retraining

Thank You!

Machine Translation in Mobile Games: Augmenting Social Media Text Normalization with Incentivized Feedback

Nikhil Bojja nbojja@machinezone.com
Arun Nedunchezhian arun@machinezone.com
Pidong Wang pwang@machinezone.com
Machine Zone Inc., Palo Alto, CA, USA

Abstract

Machine Translation across languages is made difficult in the context of Mobile games where slang or ungrammatical language reduces the effectiveness of open domain translation systems. We describe a system here that improves translation systems by normalizing user slang with an active learning system. A crowsourcing system is created by incentivizing players to normalize slang through a game feature that rewards participants with in-game currency rewards. The rewards ensure active participation from players and the feedback is in turn used to train a phrase-based Text Normalization System that is relevant to the domain of the data, thereby improving Machine Translation accuracy.

1 Introduction

Advances in Machine Translation techniques have enabled people from across the globe to communicate with each other beyond language boundaries. Online texts such as news articles can be translated on demand with commercial translation service providers. These providers have reasonable translation accuracy with texts under various domains. The problem of accuracy in Machine Translation is made severe when we target general purpose translation systems on domain specific data, especially when this domain specific data is not very grammatical. Applying domain specific data to re-train and adapt translation systems is a potential solution for this problem. However, it is not easy to obtain Social Media or Mobile Game text in a format that can be used to train translation systems.

For our experiments, we selected *Game of War: Fire Age*, a popular Massively Multiplayer Online Role Playing Game (MMORPG) that is primarily played on mobile devices on a global scale. This game has the ability to let players from around the world communicate in real-time with each other and across languages with the help of an in-built translation module. In this paper, we describe this system and the problem of acquiring data for improving machine translation output in the context of slang-speak in mobile game interactions.

In the following sections we will talk about some of the related work in this field, describe the system, showcase improvements brought in by this system and discuss future possibilities.

2 Previous work

Statistical Machine Translation (Brown et al., 1993) has made it easy for people around the world to access webpages in foreign languages. Its applications help make more information

available for those seeking it. Phrase-based Statistical Machine Translation (Koehn et al., 2003) has been a popular choice for building machine translation systems between language pairs. Parallel corpora between source and target languages are used to build phrase level alignment tables, which are then used in conjunction with a language model to generate target translations. This makes the model sensitive to the data that it is trained on and specifically the domain of the data supplied.

When it comes to specific domains like mobile games, players communicate with each other in a highly informal setting. Text generated from such a setting tends to have slang words and chats that are not necessarily structured well grammatically, and could have a lot of misspellings. It is known that should we attempt to apply Machine Translation on texts with a lot of informal slang in them, the translation output is less than optimal (Ling et al., 2013). Attempts have been made in the Machine Translation community to normalize the effect of such slang by using slang dictionaries. Aw et al. (2006) have shown that building a Statistical Machine Translation system just for the purpose of normalizing slang can have an overall improvement in translation quality. Another work (Wang and Ng, 2013; Wang, 2013) has presented a novel text rewriting decoder for slang text normalization that could enhance overall translation accuracy of the system.

3 Normalization system

The translation system in *Game of War: Fire Age* lets players chat with each other in realtime. To make this possible chats from a source language are run through *MZ Transformer*, an ensemble normalization system which employs a combination of slang dictionaries, abbreviation lists, spell checkers and most importantly a phrase based text normalization system. To develop the phrase based text normalization system, we prepared a slang corpus made up of player chats extracted from the Mobile Game logs. The data was noted to contain slang used by players in the game and reflected the informal tone of the domain. The slang corpus was then manually normalized to a grammatical equivalent corpus of sentences. The eventual parallel corpus of slang and normalized sentences served as training data for building a Phrase-based Statistical Text Normalization system using Moses (Koehn et al., 2007).

The resulting system *translated* slang text to grammatical text within the same source language. *MZ Transformer* could now handle the transformation of most of player slang used in the game and convert it to a grammatically better version. This grammatical version was then fed to a hybrid translation system which comprised of an internal cross language translation system and commercial translation service providers [1][2]. The overall quality and readability of output translations obtained was observed to be significantly better. More importantly, the system could now make sense of slang used by players than just delegating them as Out of Vocabulary words (OOV's).

4 The Data problem

The initial parallel text used for creating the prototype in Section 3 was manually created. This is of course expensive and not feasible when we want to build a more robust system with a larger training dataset or similar systems for languages other than English. Various methods have been suggested for accumulating bilingual training data for building Statistical Machine Translation systems for instant messaging systems (Bangalore et al., 2002) or for microblogs (Ling et al., 2013; Xu et al., 2013).

Though the vocabulary of these domains can be assumed to be similar to the language

[1]Microsoft Translator. `http://bing.com/translator`
[2]Google Translate. `http://translate.google.com`

used in Mobile games, we noticed that this domain uses a more specific vocabulary tied to in-game actions and events. We also noticed that the slang used in games contained many more abbreviations and variations than that of microblogs. The length of source sentences in Mobile games tended to be smaller than microblog messages such as those from Twitter. On identical sample sizes, Twitter messages averaged 73.51 characters per source sentence compared to 34.43 characters per source sentence in the Mobile game dataset (Wang et al., 2015). The length and perplexity of the Mobile game data is hence contextually limited that indicates sub-domain level differences. It should be noted that the limited data per input sentence further exacerbates the lower translation accuracy problem.

Crowdsourcing techniques could be a good way to obtain parallel data in these cases. Platforms such as Amazon's Mechanical Turk *(mTurk)* could be used to obtain data (Zaidan and Callison-Burch, 2011). Apart from the monetary cost associated with it, getting data in languages other than English came up as an issue with using *mTurk*. Thus it became necessary for us to create a novel system to create the dataset of parallel slang and grammatical data at a low cost.

4.1 Game Economy

Most Multiplayer Role playing strategy games have an in-game economy that is critical to its functioning. *Game of War* too has such an economy with in-game currency on one side, and various items available for sale on the other side. The items available are bought by players to be used in the game. There is a huge variety in the types of items available as well as the quantities in which they are offered. In-game currency is used to monitor the pricing of these items and game designers have the flexibility to offer sales and discounted prices on the items available. Needless to say, these in-game purchases are highly sought after by active players who want to get ahead of their competitors in the game. Control over such a lucrative game economy can be leveraged for our purpose of collecting data needed for training our models.

5 Crowdsourcing System

To solve the data problem, we created a Rewards based Normalization module within the game. In this module, players are presented with slang words or phrases that need to be normalized. Along with each such input, in-game items are presented as rewards in remuneration for normalizing the data. This way we provide an engaging feature within the game where players can earn in-game items in exchange for spending some effort normalizing slang text.

Text is injected into the module based on language and number of unknown slang words in the corpus. Each phrase is presented to multiple players concurrently and normalized outputs are accumulated. To ensure high quality output, we setup a two-step process. One set of players type in normalized versions of input slang words/phrases, and another set of players are shown a multiple-choice style visualization of input slang phrase and candidate normalized phrases with an option for users to choose from the normalized output versions. Automated Quality control is put in place by use of text similarity techniques to remove entries entered by users that are irrelevant to the input word/phrase.

5.1 Task Instructions

The only paragraph of instruction that appears to all players participating in the crowsourced task is: *Select the best correction for the misspelled words and earn rewards. We select the top, most accurate entry submitted by users like you and approve rewards for them. Note that there could be cases where theres no correction needed too*.

Given that the feedback system is connected to an online game with a highly active chat system the users of the system discussed the feature and have evolved into a user-base that

agreed upon the right way to do the job. We did put in checks to avoid collusion and have been successful in making the system efficient.

5.2 Creating the Parallel Corpus

Source Phrase	Response Received	Num. Users
yo wasup zack .. i just wakey	Yo, what's up Zack? I just woke up.	**1013**
	Hi, what's up Zack? I just woke up.	327
	Hey, what's up Zack? I just woke up.	133
	What's up Zack? I just woke up.	61
	To what's up Zack? I just woke up.	12
	Yo what's up Zack. I just awoke	3

Table 1: Sample of Data collected

The player base in *Game of War: Fire Age* is numerous enough for us to choose a 1-best hypothesis that has been agreed upon by a multitude of players for a given input sentence. There is of course an option to obtain n-best hypotheses - ranked by number of players agreeing on the same normalized output. Rewards are given out to players at the end of selecting the 1-best hypothesis for inputs. A sample of the data collected per phrase can be seen in Table 1. One can see that the top hypothesis is significantly ahead of the remaining hypotheses which validates the use of a 1-best hypothesis. We note that this trend is consistent with data collected across other remaining input phrases too.

Hence, we now have a feedback loop from players who can help improve the normalization process and in turn improve translation accuracy. Such a feedback loop is a desired feature in every Machine Translation system. The lack of incentives could be attributed to users seldom providing feedback on translation quality in traditional translation systems. Due to the game economy based incentives, we have a feedback loop that is assured to gather feedback in a timely manner from a willing player base.

6 Experiments and Discussions

Using the Rewards based Crowdsourcing system we were able to collect normalized data across languages such as English, French, Spanish, German, Portuguese and Russian. Translation systems augmented with *MZ Transformer* as described in Section 3 were built for each of these languages using the data collected.

To measure the impact of the normalization system on translation quality, a separate held out test set was created with manually translated messages in various language pairs. Each language pair had 1000 samples in the test set. The number of tokens in each of the test sets approximately averaged 6500 in number. The test set for each language pair was built through a random selection of chats from a database.

6.1 Results

The test set for each language pair was translated with a commercial translation provider[3] and translated with a Translation system that gets normalized inputs from *MZ Transformer*. We used the BLEU metric (Papineni et al., 2002) to measure the translation quality. The results are shown in Table 2.

The results show a clear improvement in translation quality for all language pairs when Normalization was used as a pre-step before translation. Manual analysis of the outputs showed

[3]Microsoft Translator: `http://bing.com/translator`

Source Lang.	Target Lang.	w/o Normalization	w/ Normalization
Spanish	English	37.82	**39.77**
English	Spanish	31.29	**32.87**
French	English	46.30	**47.73**
English	French	31.90	**33.19**
German	English	41.02	**43.98**
English	German	26.92	**26.96**
Portuguese	English	50.94	**52.13**
English	Portuguese	38.09	**38.12**
Russian	English	38.64	**40.17**
English	Russian	24.80	**25.43**

Table 2: BLEU score improvement

that even in language pairs where the improvement in BLEU scores was minimal, the readability of the sentences improved greatly with normalization. Normalization targets tokens that tend to have a higher degree of occurrence in player chats. As an example, *lol* in English (laugh out loud) is the most frequently occurring token in the player chat database. However, this does not occur as frequently in the test set. *MZ Transformer* however ensures that *lol* is translated to *mdr* when translating to French. *mdr* (mort de rire) is the equivalent of *lol* in French slang. Readability greatly improves in a player chat session with such translations on high frequency slang words, but such gains don't necessarily translate to BLEU score improvements.

A learning from these results is that an improvement in translation quality correlates with the number of normalization layers and the quantity of training data in *MZ Transformer*. Also, each language seems to have a different degree of slang usage and hence we deduce that perplexity correlates with translation improvement too. Do note that this was only one round of feedback addition to *MZ Transformer's* training data. After collecting some more data we could check for further improvements in translation quality.

We used 10-best hypotheses from the data collection process (Table 1) as an alternate training dataset for the Phrase-based text Normalization system. This system had a lower BLEU score compared to the system trained with 1-best hypotheses. This could be attributed to overfitting because of the high degree of similarity in training hypotheses.

7 Future work

The Mobile game economy and the demand for in-game items from players creates an ideal ecosystem where getting crowdsourced data becomes easy. With the growing popularity of Mobile games around the world, getting data on resource poor languages can be made easy through a crowdsourced ecosystem like this where we have access to native speakers of various languages globally. We have started collecting data in languages such as Bulgarian, Malay, Ukrainian, Slovak among others and hope to build similar normalization systems in these languages.

The system could be further utilized to collect data of any kind, be it text normalization, text translation or even speech transcription. The speed at which crowdsourcing is done could be modulated with the number of rewards announced for each task. This will ensure speedy output from the system should we need data urgently. As the number of players outnumbers the amount of data needed, we can get multiple hypotheses for each input, thus ensuring a high quality crowdsourced output.

References

Aw, A., Zhang, M., Xiao, J., and Su, J. (2006). A phrase-based statistical model for sms text normalization. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 33–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bangalore, S., Murdock, V., and Riccardi, G. (2002). Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system. In *COLING*.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.*, 19(2):263–311.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ling, W., Xiang, G., Dyer, C., Black, A., and Trancoso, I. (2013). Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics*, ACL '13. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wang, P. (2013). *A Text Rewriting Decoder with Application to Machine Translation*. PhD thesis, National University of Singapore.

Wang, P., Bojja, N., and Kannan, S. (2015). A language detection system for short chats in mobile games. In *Proceedings of the third International Workshop on Natural Language Processing for Social Media*, pages 20–28, Denver, Colorado. Association for Computational Linguistics.

Wang, P. and Ng, H. T. (2013). A beam-search decoder for normalization of social media text with application to machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 471–481, Atlanta, Georgia. Association for Computational Linguistics.

Xu, W., Ritter, A., and Grishman, R. (2013). *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, chapter Gathering and Generating Paraphrases from Twitter with Application to Normalization, pages 121–128. Association for Computational Linguistics.

Zaidan, O. F. and Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1220–1229, Stroudsburg, PA, USA. Association for Computational Linguistics.

Session title: Why are we (still) waiting? What premium translators need to use MT effectively

Presenters: Robin Bonthrone (Fry & Bonthrone Partnerschaft, Mainz, Germany) and Konstantin Lakshin (Russian Link LLC, Golden/CO)

Abstract:

This presentation is given by two very experienced professional translators with wide-ranging knowledge of and expertise in the potential benefits that can be obtained from using MT, as well as the practical constraints. They examine the reasons for the continued gap between what MT developers offer premium translators and the solutions that translators expect. They then examine some of the technical issues and propose a list of requirements that MT developers need to meet so that premium translators can deploy MT systems effectively and productively as a key component of a holistic expert environment that combines state-of-the-art translation support tools with the unique expertise of professional human translators. As such, the presentation combines both a strategic business case review and a more bottom-up analysis of specific technical requirements.

The past decade has seen the emergence of a split in the translation industry between the high-volume mass-market business on the one hand, and the high-end premium segment on the other, although it is rarely possible to identify a point where one stops and the other begins. It certainly appears to be the case that the mass translation market has successfully attracted much of the attention (and investment) of MT vendors up to now, whereas there is little evidence of any consistent approach to reflecting the MT-related needs of premium translators and their (equally premium) clients. A key question to be addressed is therefore whether the fact that the premium segment has often been ignored to such an extent by MT vendors is because they are actually unaware of its existence, its structures, and/or its requirements. Among other things, this presentation seeks to remedy this deficit by outlining the defining characteristics of the premium segment (or rather, segments) and what distinguishes it (them) from the more familiar mass-market, high-volume translation business.

Many premium translators working today are highly tech-savvy when it comes to a broad spectrum of translation technologies. They readily embrace state-of-the-art translation memory and terminology management suites, for example, and appreciate (and indeed demand) the tangible and sustainable productivity gains they can now leverage from the intelligent, integrated deployment of these and other systems.

They are in many cases convinced that integrating advanced MT solutions with their existing translation tools would enable them to achieve considerably greater productivity gains and economic benefits. These would not only offer sustainable solutions for the translation industry as a whole, but would also provide MT developers with a rewarding new market channel for their products. However, a number of often serious constraints—including data security, system size and scalability, interoperability, integration at both a technical and a workflow level, MT vendor longevity, and ROI—appear to come together to prevent them from doing so in practice.

Many translators active in the premium segment have—or have access to—the knowledge and skills needed to address at least some of these constraints, but they still face considerable hurdles when it comes to integrating best-of-breed MT into their workflows efficiently, or even to experimenting with it, as the existing MT HMI largely ignores the needs of professional premium translation providers. As a result, considerable potential is left unused, to the detriment of both premium translators and MT developers.

Before addressing the technical issues that need to be resolved to enable seamless interaction between premium translators and MT, there is a need to emphasize that, in most cases, specialist translation as a professional activity relies to a critical extent on various forms of knowledge—from general or language-related to domain-specific—that is not readily available or accessible in machine-readable form for use by MT. In many cases, however, translators are in a position to add some representation of such knowledge to actual source texts and training corpora for use in, for example, factored or class-based SMT models or for fine-tuning RBMT behaviors.

Moving toward such a level of integration without setting unrealistic goals would be beneficial for both translators—who would be a step closer to obtaining usable MT—and MT developers—who would gain an additional market channel that would also allow them to explore potential improvements to their systems with direct backing from the users.

To start this process, we need to rethink the concepts and architectures of user interaction with both frontend and backend resources that are already part of at least some MT systems. For the frontend, this includes such things as interactive and selective application of MT from within the traditional translation memory tools, as well as tighter integration with grammar checkers and QA tools. For the backend, it implies a relatively straightforward interface/toolset for restructuring the existing data and adding the metadata and, in many cases, direct access to engines and model settings, so that translators have the option of venturing into the uncharted waters of personalized small-scale MT systems at their own risk, or of commissioning MT experts.

The presenters expect that the facts and arguments highlighted in this presentation will contribute to a greater understanding of the needs of premium translators, as well as the opportunities for MT vendors that this segment offers. Equally, there is a hope that the conversation between the two sides that emerges from the presentation will deepen the dialog and accelerate the cooperation between translators and MT developers that will, ultimately, help to shape market-ready translation support ecosystems for premium translators that incorporate premium MT capabilities.

Machine Translation and Terminology Management

Jennifer DeCamp jdecamp@mitre.org
MITRE Corporation

Abstract

In the past few years, developers in companies such as SDL and Microsoft have focused on how to improve the quality of fully automated machine translation (FAMT) by leveraging tools for machine assisted human translation (MAHT). They have also focused on how to improve the quality of MAHT by leveraging FAMT capabilities and on how to leverage interactive MT by leveraging terminologies and dictionaries. This paper describes ways developers and users have found to leverage tools across FAMT, interactive MT, and MAHT to provide increased translation coverage, greater agility, and better quality. It also identifies and describes areas where feedback loops are non-existent or broken, resulting in diverging translations.

MT Quality Estimation for E-Commerce Data

José G. C. de Souza desouza@fbk.eu
Marcello Federico federico@fbk.eu
Fondazione Bruno Kessler, Trento, Italy

Hassan Sawaf hsawaf@ebay.com
Human Language Technology unit, eBay Inc., San Jose, USA

Abstract

In this paper we present a system that automatically estimates the quality of machine translated segments of e-commerce data without relying on reference translations. Such approach can be used to estimate the quality of machine translated text in scenarios in which references are not available. Quality estimation (QE) can be applied to select translations to be post-edited, choose the best translation from a pool of machine translation (MT) outputs, or help in the process of revision of translations, among other applications. Our approach is based on supervised machine learning algorithms that are used to train models that predict post-editing effort. The post-editing effort is measured according to the translation error rate (TER) between machine translated segments against their human post-edits. The predictions are computed at the segment level and can be easily extended to any kind of text ranging from item titles to item descriptions. In addition, our approach can be applied to different kinds of e-commerce data (e.g. different categories of products). Our models explore linguistic information regarding the complexity of the source sentence, the fluency of the translation in the target language and the adequacy of the translation with respect to its source sentence. In particular, we show that the use of named entity recognition systems as one source of linguistic information substantially improves the models' performance. In order to evaluate the efficiency of our approach, we evaluate the quality scores assigned by the QE system (predicted TER) against the human post-editions (real TER) using the Pearson correlation coefficient.

1 Introduction

Approaches to machine translation (MT) quality estimation (QE) are used in situations in which a quality score about the translation is required but no references translations are available. In MT QE, automatically translated sentences have their quality estimated without using references. Such scenarios include supporting the work of translators in a CAT scenario (Turchi et al., 2015), informing readers of the translation whether the translation is reliable or not (Turchi et al., 2012), selection of the best translation generated by a pool of MT systems (Specia et al., 2010), or filtering out low-quality translation suggestions that should be rewritten from scratch (Specia et al., 2009).

QE is usually cast as a classification, regression or ranking problem that is modelled using supervised learning techniques. The different forms of supervision used to train the models imply different ways of perceiving the quality of a translation. The choice of the supervision label depends on the envisaged application scenario. For example, for regression and ranking, previous work employed either the time required to post-edit the translations or the minimum number of modifications required to make the translation acceptable as measured by the human

translation error rate (HTER[1], see Snover et al. (2009)). Another required information that must be defined a priori is the kind of linguistic cues that are going to be used to predict quality. Such indicators are extracted from the source and the translated sentence and aim to serve as a proxy for the complexity of translating the source sentence, the fluency of the translated sentence and the adequacy of the translation in function of the source.

In this work we present the first approach to MT QE geared towards e-commerce user-generated data. Our challenge is two-fold: (i) the data have been generated by many users and therefore are not necessarily composed of grammatically well-formed sentences, and (ii) they belong to a domain composed of very diverse topics (read different categories of products). We propose new features designed to deal with the characteristics of these data and evaluate our models against post-edits produced by humans.

2 Background

eBay is a marketplace platform in which sellers can advertise items and buyers can search for items, electronically bid and eventually buy them. To enable trade between buyers and sellers with different languages, at least four types of texts need to be translated: queries, item titles, descriptions, and item specifics. Machine translation has been recently introduced in eBay's platform with the objective of fostering cross-border trade between sellers and buyers that speak different languages (Guha and Heger, 2014). In this work we predict the quality of translation of item titles, which are concise and usually very informative descriptions of items put on sale. One item title example is given below:

> *Universal 12000mAh Backup External Battery USB Power Bank Charger for Cell Phone*

It specifies several characteristics of a product ranging from more generic information (i.e. "Backup External Battery") to more specific characteristics (i.e. 12000 mAh). Common challenges in the translation of eBay's user generated content in general, and of titles (Sanchez and Badeka, 2014) are the correct rendering of proper names and the translation of words which can have multiple senses, depending on the context in which they appear. Furthermore, words can appear in a relative free-order in the title without damaging its meaning. This presents a challenge for MT QE models because they assume that source sentences are well-formed and grammatical in the source language.

3 Related Work

Most of the work for MT QE has been developed using well-formed and grammatical sentences belonging to different domains such as legal (transcription of political speeches), or news-wire texts covering different topics (Callison-Burch et al., 2012; Bojar et al., 2013, 2014). Likewise, all the features designed in previous work assume that source sentences are grammatical and that the MT systems were trained over parallel data with fluent and well-formed segments.

To the best of our knowledge, the first MT QE approach to consider user-generated data was presented by Rubino et al. (2013a). In this work, the authors present regression and classification models trained and evaluated on two different language pairs and two different domains. In particular, they developed a QE classification model for English-French information technology forums data described in Roturier and Bensadoun (2011). The approach explores features based on topic models that focus on the adequacy aspect of the translations (i.e. check whether the meaning of the source sentence is present in its translation).

[1]The translation error rate is calculated as the number of edits (word insertions, deletions, substitutions, and shifts) divided by the number of words in the reference. Lower HTER values indicate better translations.

The same English-French dataset is used by Rubino et al. (2013b) to develop QE classification models with features that were tailored to be more discriminative on user-generated data. The features specific to user-generated data explore inconsistent use of character case, non-standard punctuation, spelling mistakes and sentence splitting problems. [talk about performance and best feature sets]

Previous work differs from our work in two main aspects: quality label and data domain. The quality labels used by Rubino et al. (2013a,b) describe whether a translation is adequate or not. Binary classification models are developed aiming to predict the adequacy of translations. In this work, instead, we focus on predicting post-edition effort as a proxy for quality by training regression models. Furthermore, the data domain of previous work is information technology forums whereas our focus is on e-commerce data that spans several products in different categories.

4 MT QE for E-Commerce

In this section we describe our approach to MT QE for e-commerce data. We first describe the features we extract from both source and translation sentences and then we move to the description of the learning algorithms used for training QE models.

4.1 Features

We use a combination of features that mixes general-purpose linguistic cues with features designed specifically for the kind of data we are dealing with. We assume that the QE system does not have access to the MT system and therefore we do not use any kind of feature extracted from the MT system translation process. Such assumption allows the features presented here to be used with any MT system.

4.1.1 Domain-independent features

We extract a set of 79 domain-independent features implemented in the QuEst feature extractor framework Specia et al. (2013). These features have been proposed in previous work for MT QE and span three translation aspects: source complexity, translation fluency, and translation adequacy.

The source complexity refers to the difficulty of translating the source sentence. Longer sentences or sentences with more than one clause tend to be more complex to understand and more difficult to translate. Examples of complexity-oriented features are (computed only in the source sentence):

- number of punctuation marks;
- average token length;
- number of tokens.

The translation fluency dimension regards the correct use of grammar in the translation in the target language. The more fluent is the translation generated in the target language, the better the translation is. Examples of fluency features are (computed only in the translation sentence):

- language model log probability for the whole translation sentence;
- language model perplexity for the whole sentence;
- percentage of nouns.

Adequacy-oriented features approximate how much of the meaning of the source sentence is found in its translation. Adequacy features are computed with the source and translation sentences at the same time. Examples of features are:

- ratio of nouns in the source and translation sentences;
- absolute difference between the number of punctuation marks between the source and the translation normalised by translation length.

For a list with descriptions for all 79 features please refer to `http://www.quest.dcs.shef.ac.uk/quest_files/features_blackbox`. These features are referred to as "BB79".

4.1.2 Item title embeddings

For item titles it is more important to have translations that convey the meaning of the source title than fluent discourse in the target language. For this reason, focusing on adequacy features is important: because they can capture the meaning of the title instead of language correctness. Following the recent popularity of word embeddings in the NLP literature, we experiment with `paragraph2vec` (Le and Mikolov, 2014) to obtain embeddings that encode the meaning of a title. The embeddings are trained for both the source and the target side of the available item titles parallel corpus. Both source and target embeddings of a given title are then concatenated and used as features in our regression models. The number of features varies according to the number of dimensions of the embeddings. We experiment with several dimensions and the best results are reported in Section 6. We train the embeddings with the `paragraph2vec` implementation of `gensim`[2] (Řehůřek and Sojka, 2010). These features are referred to as "DM" (distributional memory).

4.1.3 NER-based features

Item titles segments present many word or expressions (formed by more than one word) that are proper names and that should not be translated (such as brand names or technical expressions like USB). A MT QE system could benefit of named entity recognition (NER) system that outputs whether a given token is a named entity. If the token is marked as a named entity it should not be translated and therefore it is possible to check whether the brand or name is preserved in the translation.

We developed a set of three features that verify whether the tokens marked as named entities or "do-not-translate" are in fact not translated in the MT output. The three features are:

- number of "do-not-translate" tokens found in the source sentence;
- number of tokens in the translation segment that exactly match the items marked as "do-not-translation" in the source sentence;
- a ratio between the second and first features above.

Such features rely on a in-house NER system that produces binary tags for each token in a sentence. These features could be considered adequacy-oriented features and are tailored specifically for user-generated e-commerce data that inherent to eBay's platform. These are called "NER" hereafter.

[2] `https://radimrehurek.com/gensim/index.html`

4.2 Learning algorithms

We train our models with two different non-linear ensemble learning algorithms: extremely randomized trees (Geurts et al., 2006) and AdaBoost regression trees (). Both are batch non-linear learning algorithms that also provide the importance of each feature in the final fitted model.

Extremely randomized trees (ET) is a learning algorithm based on an ensemble of decision trees (Breiman et al., 1984). ET is an ensemble of randomized trees in which each decision tree can be parameterized differently. When a tree is built, the node splitting step is done at random by picking the best split among a random subset of the input features. All the trees are grown on the whole training set and the results of the individual trees are combined by averaging their predictions. We explore this model after successful results in MT QE (de Souza et al., 2014a).

The second learning algorithm we use to train our models is AdaBoost Regression (Drucker, 1997) (ADA). This algorithm fits a sequence of weak learners (very small decision trees in our case) on several iterations of modified versions of the data. Training examples receive weights according to the difficulty the model has at predicting them, forcing the algorithm to focus on examples that were incorrectly predicted by previous iterations. The final prediction consists of a weighted majority vote (or sum) of all iterations.

5 Experimental Settings

5.1 MT system

The MT system was trained with in-domain (item titles and item descriptions of e-commerce data) and out of domain parallel data (legal and news-wire texts) for training the word alignments. Translation models were trained using the standard Moses pipeline. Due to the nature of the item titles, no lexicalized reordering model is used. On the target side, trigram language models are trained. The parallel data used to train the system comes from various publicly available collections, proprietary repositories and in-house translated item titles. In particular, in-house translated items, descriptions, and specifics are here considered as in-domain data while all the rest is regarded as out of domain data. A summary of the data used to train the MT system is given in Table 1.

	Train (Out-Domain)	**Train (In-Domain)**
Segments no.	5.28M	336K
Tokens EN	69M	2M
Tokens PT no.	70M	2M

Table 1: Statistics of English-Portuguese parallel data.

5.2 Data

We train and evaluate our models on item titles translated from English to Portuguese with the MT system described in Section 5.1. All the translations were post-edited by professional translators following a conservative post-edition guideline (i.e. the post-editors should focus on the minimum modifications necessary to make the translation acceptable). We worked on a translation job of approximately 11,000 translation units comprising more than 200 product categories. For our experiments, we focus on the three more frequent categories, namely: “Cellphones & Accessories” (CPA), “Cellphones & Smartphones” (CPS) and “Women’s Clothing” (WC).

We compute the HTER scores between the translations and their post-editions[3] for each category. The scores are clipped between 0 and 1. The distributions are very different across the

[3]The HTER scores are computed with the tercom tool available at

	CPA	CPS	WC
Segments no.	854	1,031	834
Tokens EN no.	11,632	11,807	10,118
Tokens PT no.	13,318	13,552	11,686

Table 2: Summary statistics for the data used to train and evaluate the QE models.

different categories, showing a big discrepancy in translation quality. WC has a large mass of segments with HTER close to 1 (which means almost all translations are rewritten from scratch) whereas CPS and CPA are centered around the range that goes from 0.4 to 0.7 HTER. In general, the translation quality for all the categories is low, with most of the segments presenting HTER higher than 0.3. The HTER distributions are shown in Figure 1.

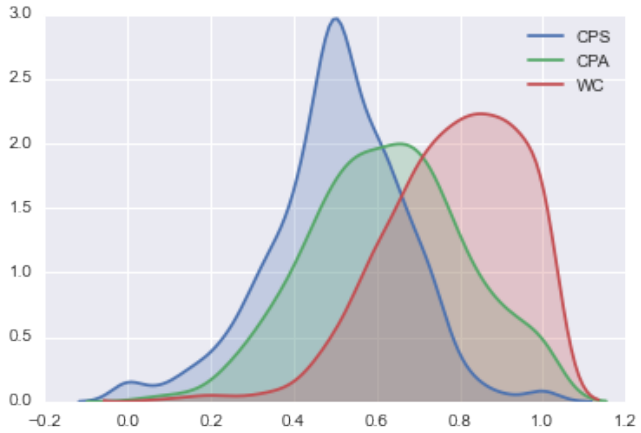


Figure 1: HTER distributions for the segments of the CPA, CPS and WC categories.

5.3 Evaluation metrics

The performance of our regression models is evaluated in terms of two metrics. The first is the mean absolute error (MAE), a standard error measure for regression problems commonly used also for QE Callison-Burch et al. (2012). The MAE is the average of the absolute errors $e_i = |\hat{y}_i - y_i|$, where $\hat{y}_i$ is the prediction of the model and y_i is the true value for the i^{th} instance. As it is an error measure, lower values indicate better performance (↓).

The second is the Pearson correlation, a measure of the linear dependence between two variables. Pearson correlation is defined as the covariance of two variables divided by the product of their standard deviations: $\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$. Higher values indicate better performance (↑).

6 Results and Discussion

In this section we report the results of our models. Hyper-parameters were found with 100 iterations of randomized search (Bergstra and Bengio, 2012) on 5-fold cross-validation over the training data. The final models were trained over the whole training data with the best parameters found during the randomized search procedure.

Results for the three categories evaluated are in Table 3 (CPA), Table 4 (CPS) and Table 5. The first row of each column is a simple baseline that applies the training set HTER mean as

CPA				
	ET		ADA	
Features	**MAE ↓**	**Pearson ↑**	**MAE ↓**	**Pearson ↑**
Mean	15.4	0	15.4	0
BB79	14.29	47.32	13.64	50.27
DM+BB79	14.29	47.6	13.83	46.35
BB79+NER	**13.78**	**50.37**	**13.07**	**55.97**
DM+BB79+NER	13.84	49.87	13.46	51.93

Table 3: Results for the category "Cellphones & Accessories" (CPA).

a prediction for every segment in the test set. This baseline is a lower bound above which our models should perform. Any model with results lower than the "Mean" baseline do not learn anything with the data. All three categories "Mean" baseline presents the highest MAE and correlation equal to zero.

CPS				
	ET		ADA	
Features	**MAE ↓**	**Pearson ↑**	**MAE ↓**	**Pearson ↑**
Mean	12.86	0	12.86	0
BB79	12.42	39.56	11.68	45.57
DM+BB79	12.5	38.72	12.18	41.59
BB79+NER	**12.19**	**44.17**	**11.11**	**53.51**
DM+BB79+NER	12.29	43.42	11.8	49.28

Table 4: Results for the category "Cellphone & Smartphones" (CPS).

Overall, the best feature set is the combination of BB79 and NER for both ET and ADA. For both CPA and CPS this combination presents the best MAE and Pearson correlation. The NER feature set seem to help in particular for the CPA and CPS categories but not so much for WC. The main reason are the characteristics of item titles in the WC category. They contain less brand names and technical concepts about the product and more generic descriptions about clothes, making the named entity information less efficient. Furthermore, the general performance of the QE models for WC is much lower than for the other two categories. The most likely reason is the distribution of HTER labels (Figure 1), which is almost in its entirety composed of bad translations (close to 1 HTER) and very few examples of good translations (close to zero HTER).

WC				
	ET		ADA	
Features	**MAE ↓**	**Pearson ↑**	**MAE ↓**	**Pearson ↑**
Mean	12.99	0	12.99	0
BB79	**12.83**	13.2	13.11	6.75
DM+BB79	12.93	10.04	**12.55**	**11.27**
BB79+NER	12.84	**12.15**	12.93	10.8
DM+BB79+NER	12.93	7.24	12.72	4.14

Table 5: Results for the category "Women's clothing" (WC).

The features based on the title embeddings (DM) do not seem to help the overall performance for predicting post-edition effort. It presents the best results when combined with BB79 and trained with ADA for the WC category, however, the final Pearson correlation is very low if compared with the best models for the other two categories (Pearson correlation of 11.27).

Regarding the learning algorithms, ADA outperforms ET for both CPA and CPS categories. For CPS the results are substantially higher (approximately 1 MAE point and 9 Pearson correlation points). AdaBoost's shortcoming, however, is the time required to train the models. In our experiments, it was as much as 15 times slower than ET.

6.1 Feature analysis

In order to better understand what are the most predictive features for the e-commerce domain we analyze what are the ten most important features according to the models trained with ET for each category. Here we present the features that appear in the intersection of the top-10 most important features for each pair of categories. In the following list, features are sorted by their importance score (for CPA and CPS):

- number of named entities marked as do-not-translate found in the translation;
- number of named entities found in the source sentence (do-not-translate terms);
- ratio of named entities matches found in the translation divided by total number of named entities in the source;
- average number of translations per source word in the sentence (threshold in IBM1: prob > 0.01) weighted by the frequency of each word in the source corpus
- average word frequency: on average, each type (unigram) in the source sentence appears N times in the corpus (in all quartiles);

The most predictive features are the ones related to adequacy and the features developed specifically for eBay's data are the most predictive (NER-based features). For WC, on the contrary, they were not helpful:

- language model log probability of part-of-speech tags in the translation sentence
- language model log probability of the translation sentence

The most predictive features for the WC category are the ones that model fluency in the automatically-generated translation. One interesting avenue of research is to analyse how similar are categories taking into consideration only the features extracted and exploring their similarities and discrepancies in order to build more robust QE models (similarly to de Souza et al. (2014b)).

7 Conclusion

In this paper we presented an approach to MT QE for e-commerce data. We train and evaluate models that predict post-edition effort (HTER) on products from three different categories in the inventory of eBay's marketplace platform. Our models use a combination of domain-independent and domain-specific features and reach approximately 55% correlation when evaluated against post-edit scores produced by professional translators.

As future work, we would like to test our QE system in a localization application scenario. We envisage that such a system could be used to sample segments to be sent for post-edition or to revise post-editions produced by a language service provider (LSP). Many companies and LSPs still rely on a random sampling process that could be improved quality and time-wise by a more informed method that uses MT QE to score translations.

Acknowledgements

The first author received support through a financial gift by eBay Inc. to FBK. The second author received support by the EU H2020 funded MMT project (grant agreement No 645487), by eBay Inc., and by FBK's Mobility programme.

References

Bergstra, J. and Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13:281–305.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Eighth Workshop on Statistical Machine Translation*, pages 1–44.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2012). Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 10–51, Montr{é}al, Canada. Association for Computational Linguistics.

de Souza, J. G. C., González-Rubio, J., Buck, C., Turchi, M., and Negri, M. (2014a). FBK-UPV-UEdin participation in the WMT14 Quality Estimation shared-task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 322–328.

de Souza, J. G. C., Turchi, M., and Negri, M. (2014b). Machine Translation Quality Estimation Across Domains. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.*, pages 409–420.

Drucker, H. (1997). Improving regressors using boosting techniques. In *14th International Conference on Machine Learning*, pages 107–115.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Guha, J. and Heger, C. (2014). Machine Translation for Global E-Commerce on eBay. In *Proceedings of the AMTA*, volume 2: MT Users, pages 31–37.

Le, Q. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *International Conference on Machine Learning - ICML 2014*, 32:1188–1196.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Roturier, J. and Bensadoun, A. (2011). Evaluation of MT Systems to Translate User Generated Content. In *Proceedings of the Thirteenth Machine Translation Summit*, pages 244–251.

Rubino, R., de Souza, J. G. C., and Specia, L. (2013a). Topic Models for Translation Quality Estimation for Gisting Purposes. In *Machine Translation Summit XIV*, pages 295–302.

Rubino, R., Foster, J., Samad Zadeh Kaljahi, R., Roturier, J., and Hollowood, F. (2013b). Estimating the Quality of Translated User-Generated Content. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, number October, pages 1167–1173.

Sanchez, J. and Badeka, T. (2014). Linguistic QA for MT of user-generated content at eBay. In *Proceedings of the AMTA*, volume 2: MT Users, pages 1–24.

Snover, M., Madnani, N., and Dorr, B. (2009). Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, number March, pages 259–268.

Specia, L., Cancedda, N., Dymetman, M., Turchi, M., and Cristianini, N. (2009). Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of the 13th Annual Conference of the EAMT*, number May, pages 28–35.

Specia, L., Raj, D., and Turchi, M. (2010). Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

Specia, L., Shah, K., de Souza, J. G. C., and Cohn, T. (2013). QuEst–A translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 79–84.

Turchi, M., Negri, M., and Federico, M. (2015). MT Quality Estimation for Computer-assisted Translation: Does it Really Help ? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 530–535.

Turchi, M., Steinberger, J., and Specia, L. (2012). Relevance Ranking for Translated Texts. In *Proceedings of th 16th International Conference of the European Association for Machine Translation (EAMT)*, number May, pages 153–160.

Yandex.Translate approach to the translation of Turkic languages

Irina Galinskaya
Farkhat Aminov
Yandex, LLC.

The Turkic languages are spoken by about 170 million people from Southeastern Europe to Asia. They are also official languages in many countries and national autonomies and, therefore, supported by governments and public institutions. With the development of Internet in Asian region the need of online translation for these languages is constantly growing and companies engaged in the machine translation development show a great interest in this language group. The creation of high-quality machine translation for Turkic languages is not only an important business task and a very interesting scientific problem, but also a strong challenge for both developers and researchers.

At first glance, Turkic languages may seem to be a simple task for machine processing due to regularity of Turkic morphology as well as a lexical similarity between languages of the group. However, Turkic languages pose some problems for statistical machine translation, as they are agglutinative and represent a huge variety of word forms, translation of which cannot be fully covered by the available data. Another problem is related to the syntactic structure of the Turkic languages, which is different from most European languages. Finally, there are very few documents available in the Web for many of the Turkic languages.

To solve the task, we conducted a series of experiments. First, for a large group of Turkic languages we collected parallel corpuses from the documents available in the web and then evaluated the quality of the resulting machine translation. Second, for the language pair with the best translation quality, which was Turkish-English, we took the initial version as a baseline and made the following improvements: integrated Turkish morphology analyser (to relieve lexical scarcity); developed a pre-ordering mechanism (to provide a coherent translation). These helped us to significantly increase the quality of Turkish-English translation and allowed to enter the Turkish market with a line of translation services and products. The presentation will show how the audience was growing along with the increase of the translation quality.

Approaches proven in the experiments with Turkish language can be applied to all other Turkic languages, primarily for the Kazakh and Azerbaijani, languages with many documents in the web and the high basic translation quality. However, we don’t have enough documents for other Turkic languages spoken in the former USSR (Uzbek, Turkmen, Kirghiz) and in Russian Federation (Tatar, Bashkir, Chuvash) so the morphological analysis and pre-ordering cannot significantly improve the translation quality for these languages.

We describe our approach to developing the translation for low-resourced Turkic language by the example of Tatar-Russian translation. The achieved translation quality allowed us to release the first version of the machine translation for the Tatar language. This event sparked great interest in Tatarstan not only to the Tatar translation, but also to the translation service itself.

Solving Specific Content Challenges with Flexible Machine Translation

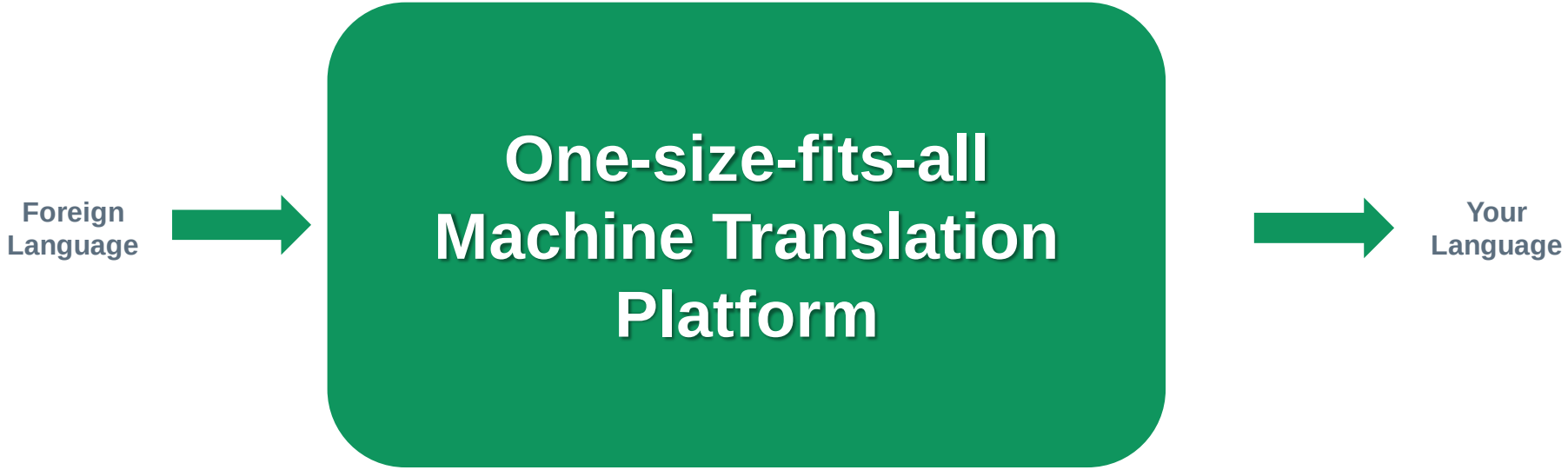
MT Summit XV

Quinn Lam, Senior Program Manager, Machine Translation

November 2015

What is Flexible Machine Translation?

In the past …

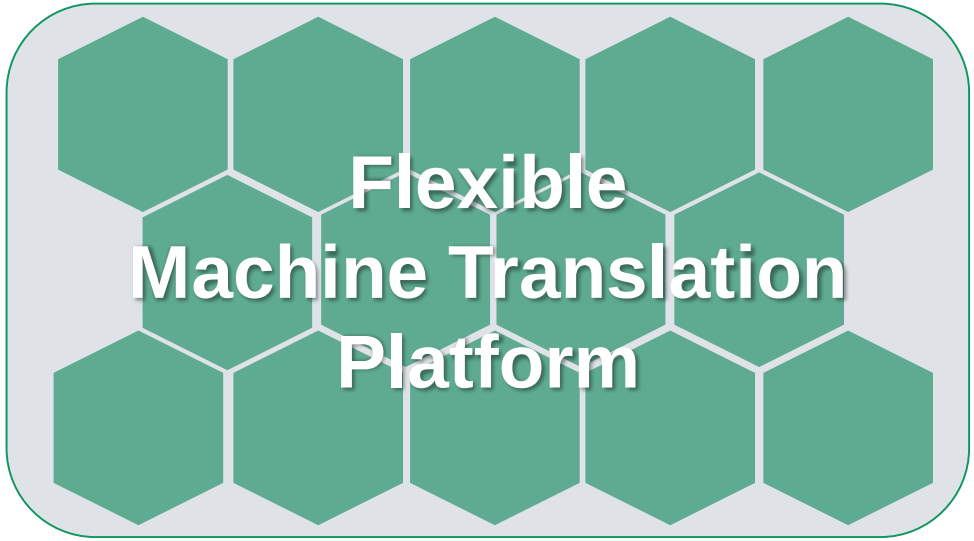


Translation go through the decoding & training pipeline, regardless of

- **The source & target language**
- **The content domain**
- **The translation use case**

Our approach

Foreign Language



Your Language

Modules could be added/removed/modified from the decoding & training pipeline, depending on

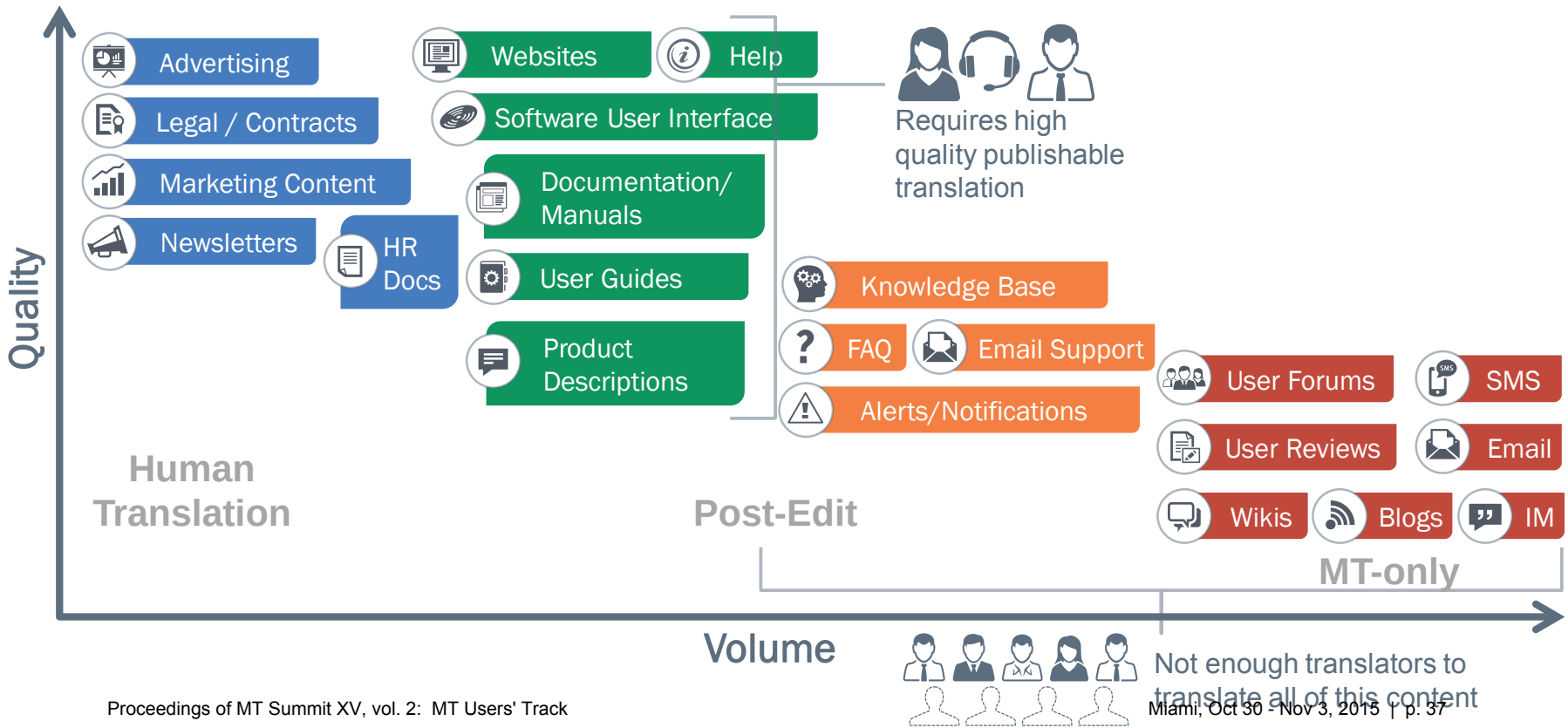
- **The source & target language**
- **The content domain**
- **The translation use case**
- **etc**

Flexible Machine Translation in Action

- Specific challenges our customers encountered with MT
- Current MT shortfall in resolving those challenges
- SDL MT solution

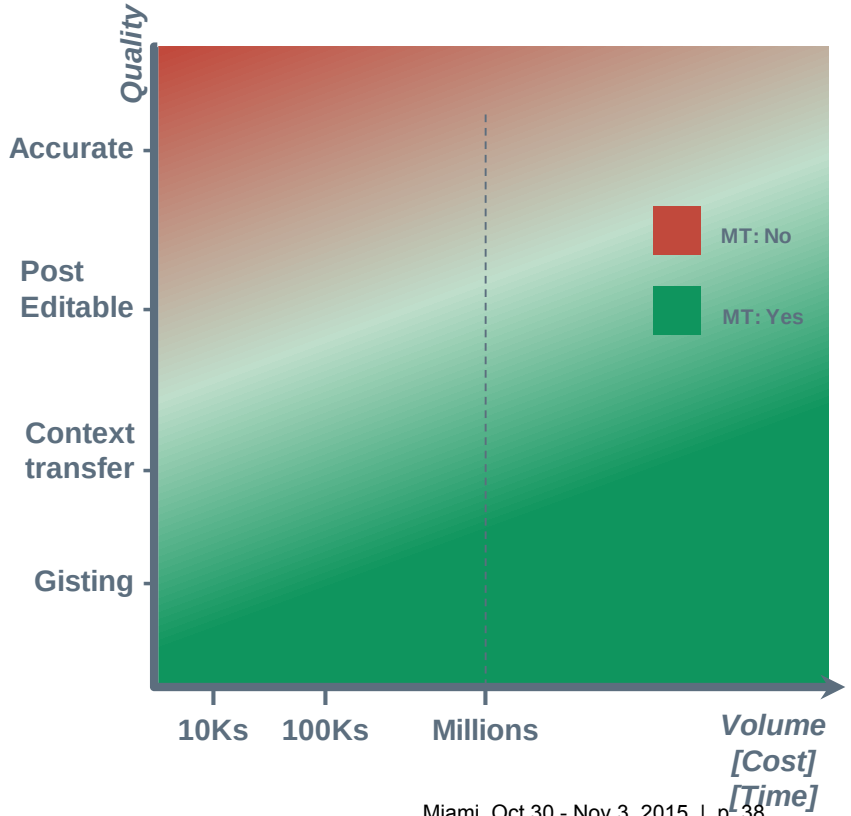
MT qualification

MT Use-Cases


Quality
Advertising
Legal / Contracts
Marketing Content
Newsletters
HR Docs
Websites
Help
Software User Interface
Documentation/ Manuals
User Guides
Product Descriptions
Requires high quality publishable translation
Knowledge Base
FAQ
Email Support
Alerts/Notifications
User Forums
SMS
User Reviews
Email
Wikis
Blogs
IM
Human Translation
Post-Edit
MT-only
Volume
Not enough translators to translate all of this content

Qualifying profiles for MT adoption

- While there are many important criteria, there are three key qualifiers for MT adoption:
  1. **Speed**: Where content needs to be translated at a pace that humans can not match.
  2. **Volume**: Where volume exceeds what can reasonably be accomplished (time and cost) by humans.
  3. **Quality**: Ability to produce translations at a compelling quality. MT does not deliver perfect translations ("perfect" is subjective), but translation that are actionable.



Improving Number Translation Accuracy

SDL

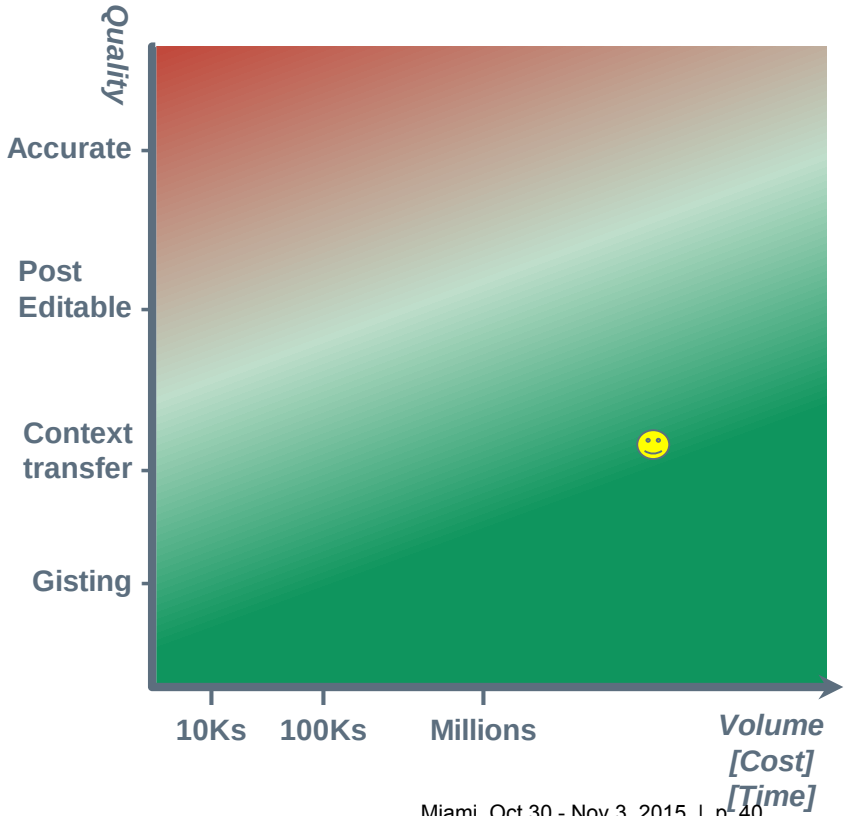
Customer A

Domain: Financial

Use-Case: Analysis of financial news and press releases

MT Solution: Custom MT engines using customer's domain specific-data

- Higher Bleu scores when compared to baseline engines
- Higher Likert (Human Reviewer) scores through blind evaluation



Examples of Number Variations

FAKTORZY	IQ 2008 obroty w mln PLN	IQ 2009 obroty w mln PLN	% udział w rynku	% rok do roku
ING Commercial Finance	3127.0	**2118.0**	30.26%	**-32.3%**
Pekao Faktoring	1275.0	**1104.0**	15.78%	**-13.4%**
Coface Poland Factoring Sp. z o.o.	548.1	**954.0**	13.63%	**74.0%**
Polfactor	775.0	**782.2**	11.18%	**0.9%**

Aktualne notowania NBP (2015-08-27)			Przed miesiącem (2015-07-27)		
Waluta	Kurs (PLN)	Zmiana	Waluta	Kurs (PLN)	Zmiana mies.
1 EUR	4,2255	- 0.3 % ↓	1 EUR	4,1495	+ 1.83 % ↑
1 CHF	3,9260	+ 0.2 % ↑	1 CHF	3,9150	+ 0.28 % ↑
					1 % ↓

Nazwa	Otwarcie dnia	Kurs	Zmiana	Zmiana %	Minimum	Maksimum	Ostatnia zmiana
AUD / CAD	0,9477	0,9477	↑ +0,0001	↑ +0,01%	0,9429	0,9493	21:14
AUD / CHF	0,6802	0,6899	↑ +0,0098	↑ +1,44%	0,6760	0,6931	21:14
AUD / JPY	85,6900	86,3470	↑ +0,6570	↑ +0,77%	85,0750	86,9400	21:14
AUD / NZD	1,1046	1,1094	↑ +0,0048	↑ +0,43%	1,1015	1,1099	21:14

Financial news are "digits heavy"

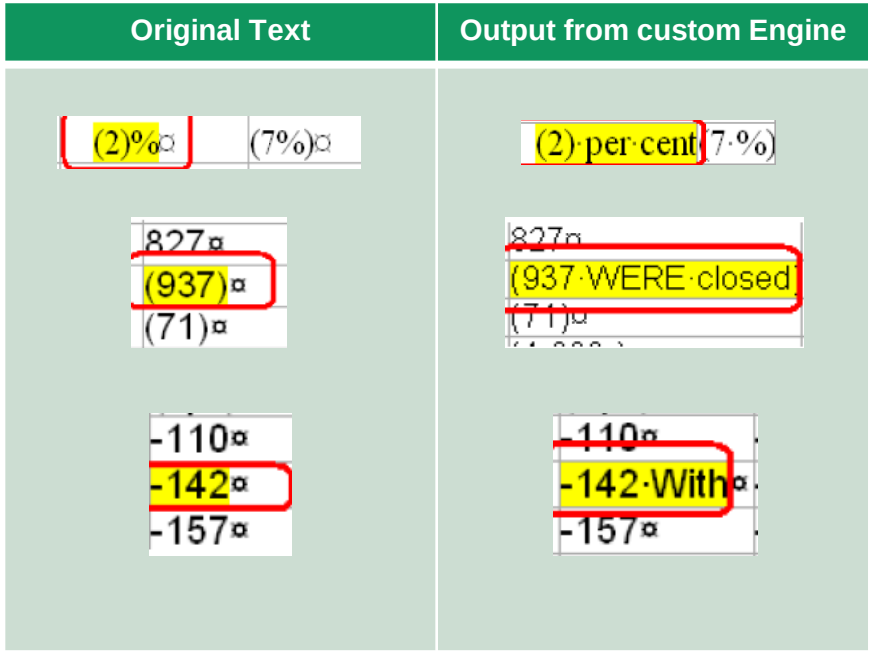
Challenge for MT: Numbers, currencies, dates, etc. need to be translated with high accuracy and consistency

- Dates need to stay in original language format
- Numbers need to be in original language format
- Currency symbols need to be correct
- Negative losses (-) and positive gains (+) need to be respected

Number Translation Issues

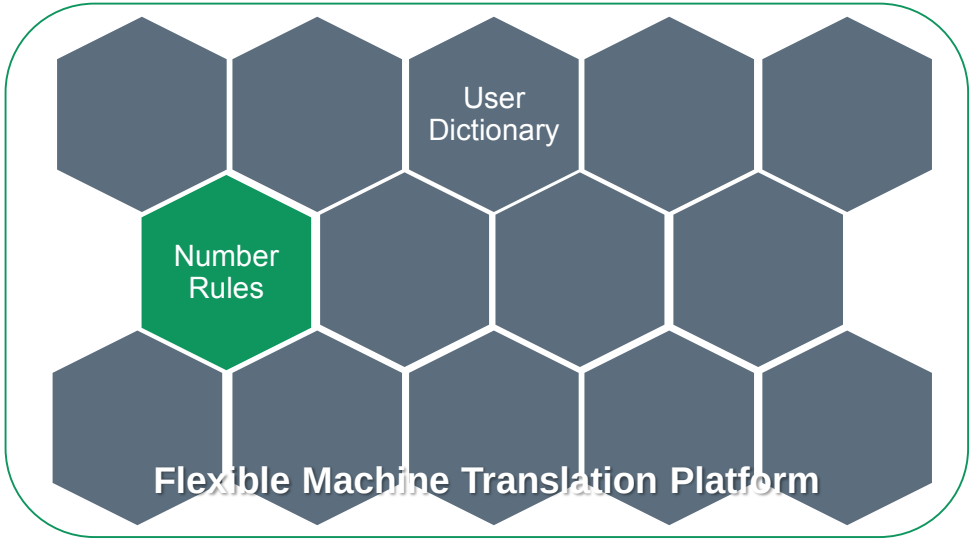
With use, customer found “surprising” issues related to how some numbers were translated

The statistical nature of SMT creates unpredictable number translation when SMT engine had learned incorrect phrase-pairs with numbers



Solution with Flexible MT Platform

Foreign Language



Your Language

- 109/109 Number-related issues fixed
- Number Rules can be added as new pattern of issues occur
- No retraining of customer engines required
- Number Rules are compatible with user's dictionary usage

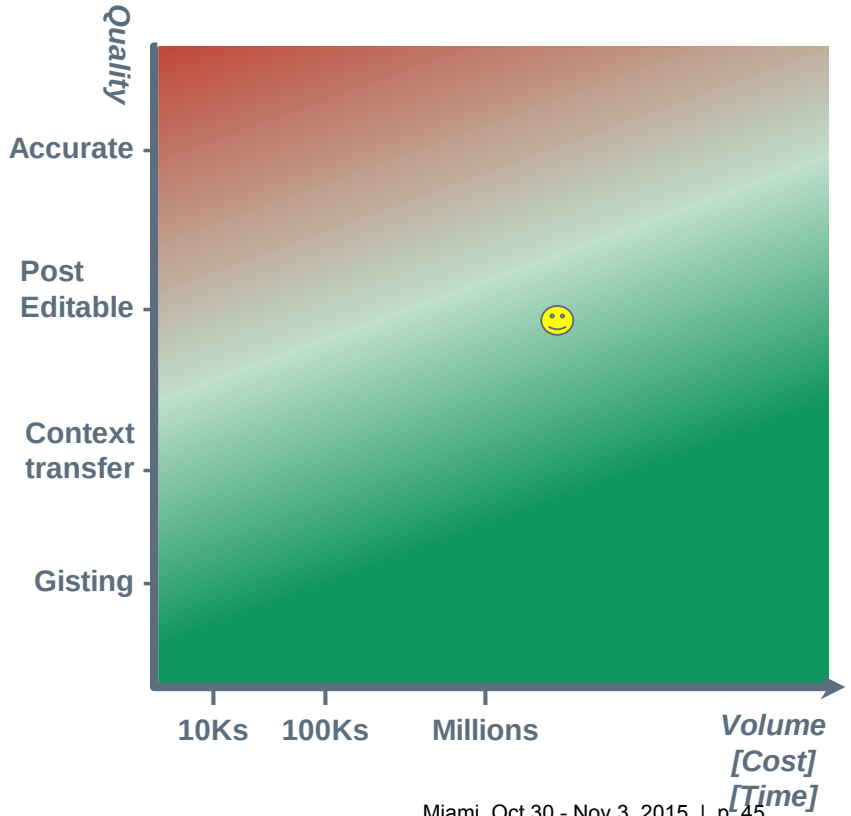
Translating Unstructured Addresses to Structured Form

Customer B

Domain: Mass Media

Use-Case: Detection and analysis of organized events

MT Solution: Generic MT engines to translate externally generated press releases and event details



Examples of South Korea Addresses

South Korea Addresses	Romanized Addresses
153-014 서울 금천구 시흥대로 378	378, Siheung-daero, Geumcheon-gu, Seoul 153-014
152-050 서울 구로구 구로동 1128-1	1128-1 Guro-dong Guro-Gu, Seoul 152-050
서울 마포구 상암동 1587	1587 Sangam-dong, Mapo-gu, Seoul
215-852 강원도 양양 강현면 물치리 16-3	16-3, Mulchi-ri, Ganghyeon-myeon, Yangyang-gun, Gangwon-do 215-852
손양면 선사유적로 678	678, Seonsayujeok-ro, Sonyang-myeon, Yangyang-gun

Challenge for MT: Not enough "good" data for the MT engines to learn from.

- South Korea addressing systems continue to reform overtime
- Official address system is not the most common address system used

Korea Addresses Through MT

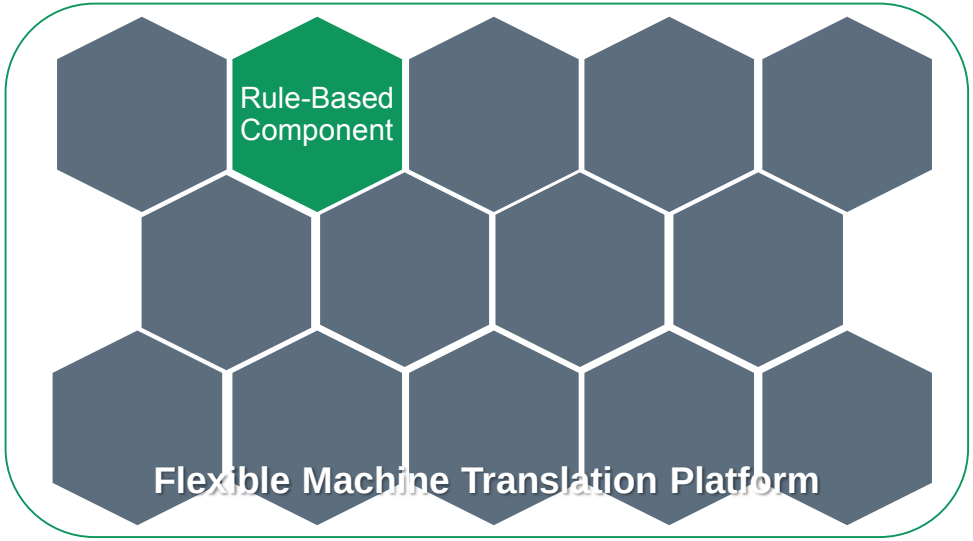
Korea Addresses	Romanized Addresses	Baseline MT Engine
153-014 서울 금천구 시흥대로 378	378, Siheung-daero, Geumcheon-gu, Seoul 153-014	Dong Geumcheon-gu, Seoul 153-014 378
152-050 서울 구로구 구로동 1128-1	1128-1 Guro-dong Guro-Gu, Seoul 152-050	152-050 Seoul Guro-dong, 1128-1
서울 마포구 상암동 1587	1587 Sangam-dong, Mapo-gu, Seoul	The Sangam, Mapo-gu in Seoul, 1587
215-852 강원도 양양 강현면 물치리 16-3	16-3, Mulchi-ri, Ganghyeon-myeon, Yangyang-gun, Gangwon-do 215-852	215-852 mulchiri in Gangwon Province on Mt. Yang Yang 16-3
손양면 선사유적로 678	678, Seonsayujeok-ro, Sonyang-myeon, Yangyang-gun	International Seonsayujeogro 678

Solution with Flexible MT Platform

Korea Addresses	Romanized Addresses	Baseline MT Engine	Rule-based Component Enhancement
153-014 서울 금천구 시흥대로 378	378, Siheung-daero, Geumcheon-gu, Seoul 153-014	Dong Geumcheon-gu, Seoul 153-014 378	153-014 378, siheung-daero, Geumcheon-gu in Seoul.
152-050 서울 구로구 구로동 1128-1	1128-1 Guro-dong Guro-Gu, Seoul 152-050	152-050 Seoul Guro-dong, 1128-1	152-050 1128-1, guro-dong, Guro-gu in Seoul
서울 마포구 상암동 1587	1587 Sangam-dong, Mapo-gu, Seoul	The Sangam, Mapo-gu in Seoul, 1587	1587, sangam-dong, Manan-gu in Seoul.
215-852 강원도 양양 강현면 물치리 16-3	16-3, Mulchi-ri, Ganghyeon-myeon, Yangyang-gun, Gangwon-do 215-852	215-852 mulchiri in Gangwon Province on Mt. Yang Yang 16-3	215-852 Gangwon-do16-3, mulchi-ri, Ganghyeon-myeon, Yangyang-gun
손양면 선사유적로 678	678, Seonsayujeok-ro, Sonyang-myeon, Yangyang-gun	International Seonsayujeogro 678	678, Seonsayujeok-ro, Sonyang-myeon

Solution with Flexible MT Platform

Foreign Language



Your Language

Before:
- Exact match - 1465/26110 (5.6%)

After:
- Exact match – 8248/26110 (31.6%)

463% improvement in exact translation

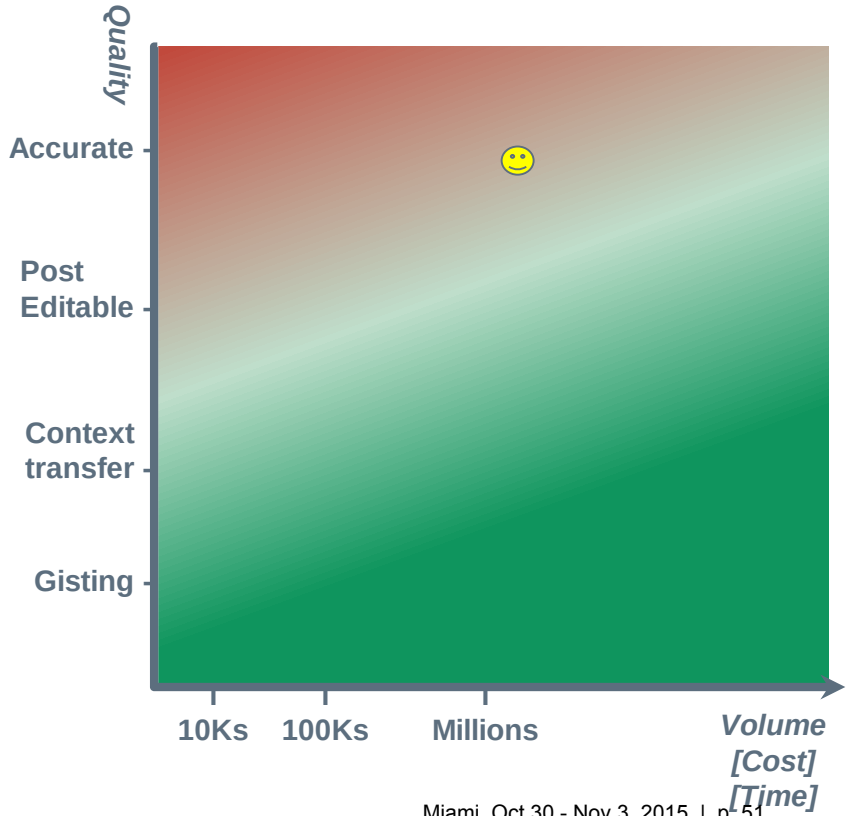
Raising MT Quality of Dissimilar Syntax Language Pairs

Customer C

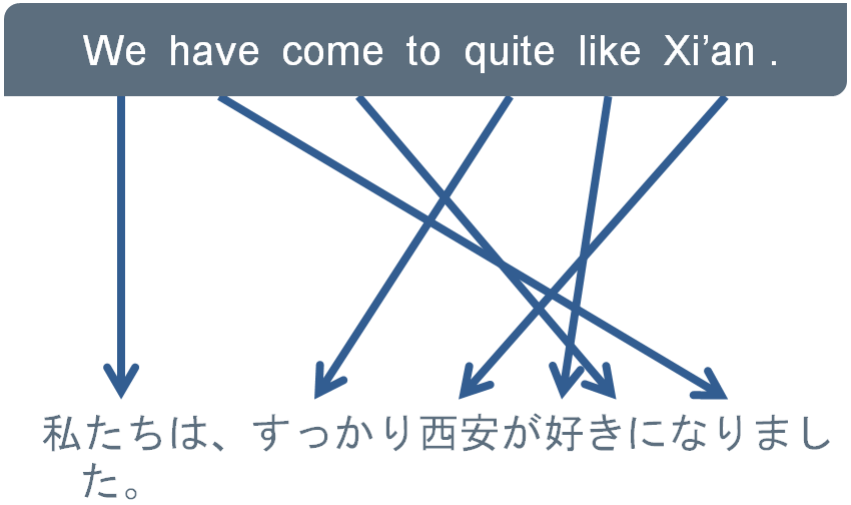
Domain: Language Service Provider (LSP)

Use-Case: Translation productivity

MT Solution: Generic Japanese <> English engines to be used in projects whenever possible



Example of English to Japanese Translation

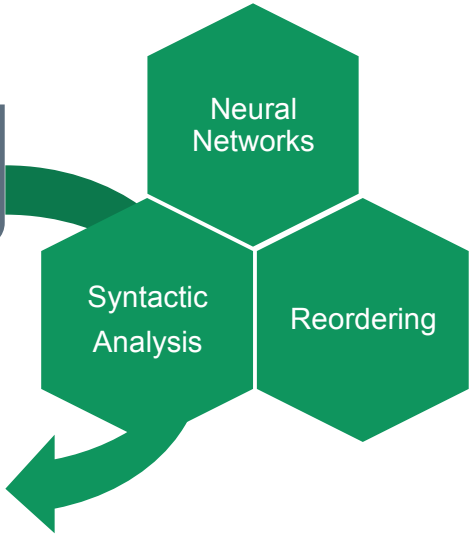


Challenge for MT: English and Japanese are syntactically very different

- Makes reordering of words and phrases to well formed sentences difficult for MT

Solution with Flexible MT Platform

We have come to quite like Xi'an .

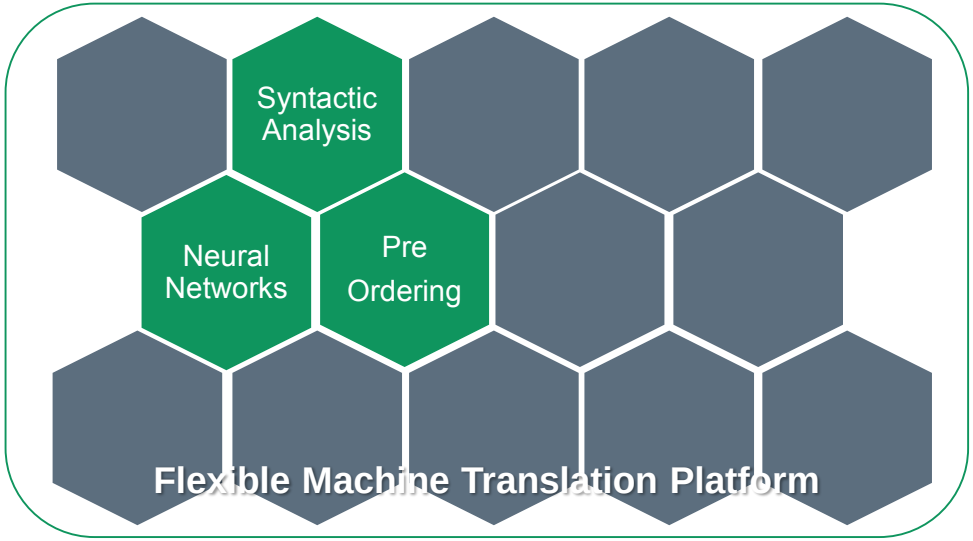


We quite Xi'an like to come have .

私たちは、すっかり西安が好きになりました。

Solution with Flexible MT Platform

Foreign Language



Your Language

Modules added to Eng<>Jpn engines during the training of the engines and are used during decoding time

- **5% Bleu score improvement**
- **+0.08 Likert score improvement**

Translating Social Media Morphed Content

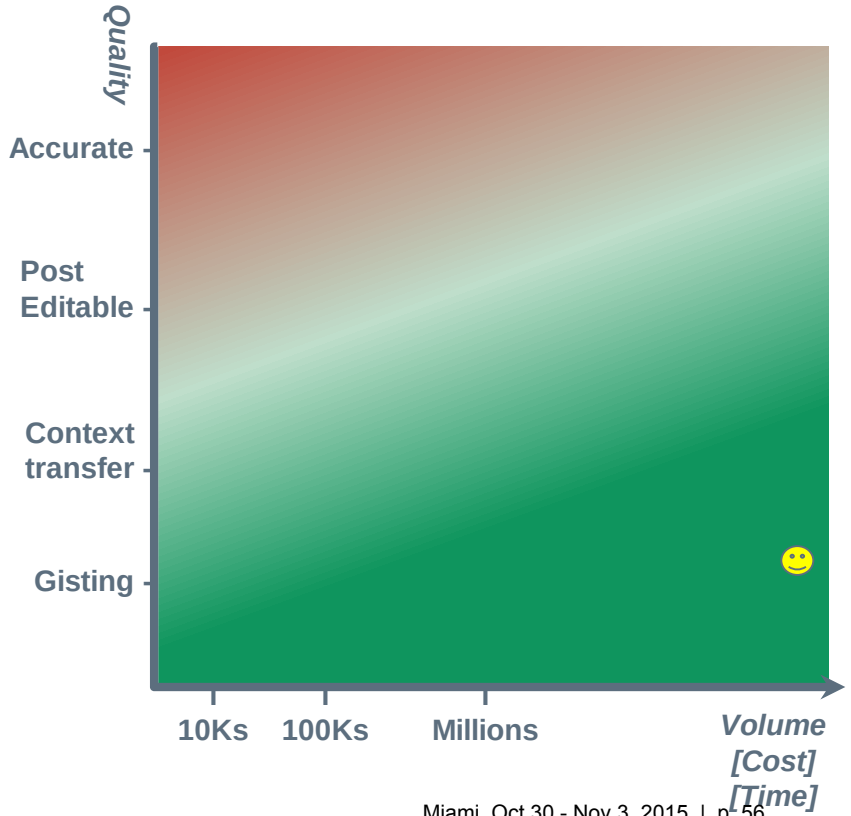
SDL

Customer D

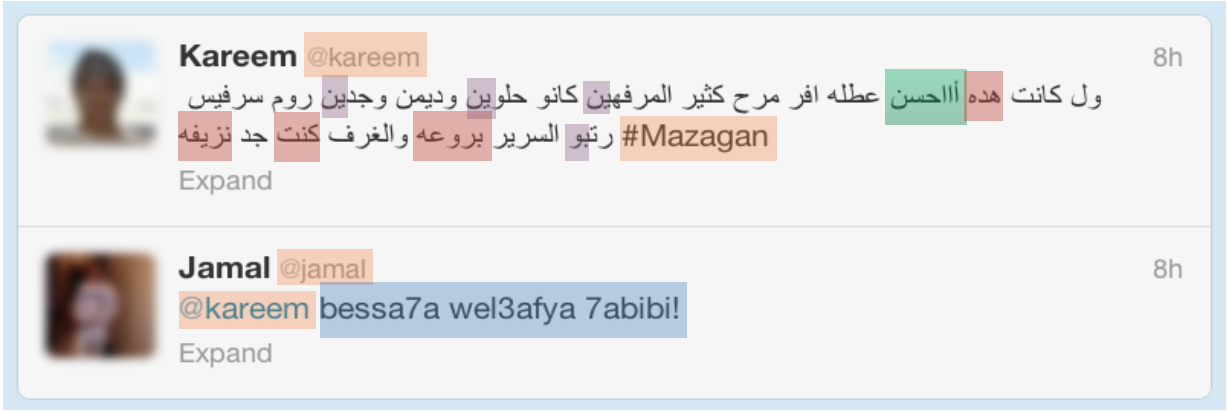
Domain: Social Media

Use-Case: Sentiment analysis

MT Solution: High volume of user-generated informal text through customized MT engine (using domain specific data)



Example of Informal Text

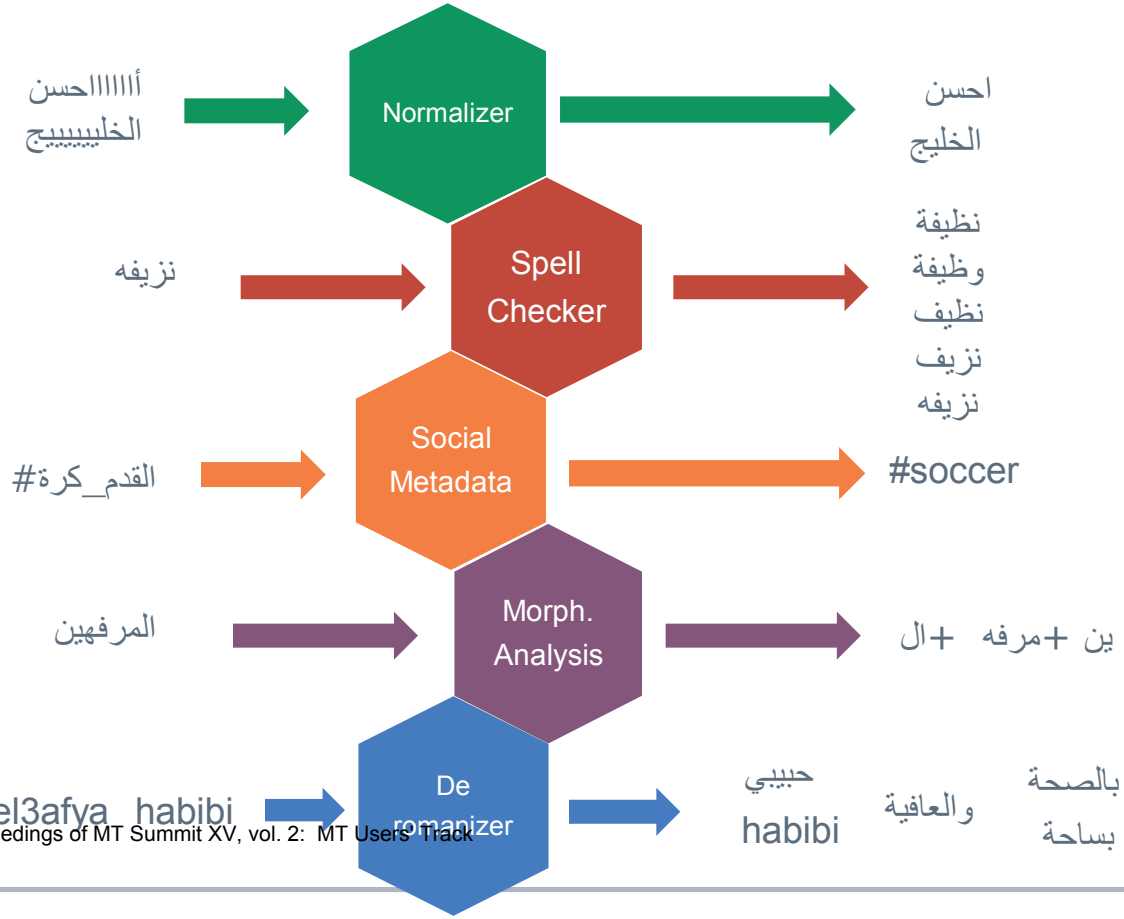


Challenge for MT: user-generated text found on social media are informal and short-handed writings.

Contains issues for MT such as

- Character Repetition
- Spelling Errors
- Morphology
- Metadata
- Romanization

Solution with Flexible MT Platform



Example of Romanized Arabic Translation

Source

- la2a hia katir fi lakhbar.
- ma 3ajbanish kida. Lazim t3'iyyer l3ounouane
- Enty habla ?
- Kalemni lama t3raf ezay tebatal teshtemni
- 3andy soda3 fi rassi... 5oshy namy badal chat. a7san lik Ah sa7

Existing MT

- La2a hia katir Fi lakhbar.
- Ma 3ajbanish kida. lazim T3 (iyyer L3ounouane
- enty habla?
- kalemni Lama T3RAF ezay tebatal teshtemni
- 3Andy soda3 Fi rassi ... 5oshy namy badal Chat. A7San lik Ah SA7

Solution with Flexible MT Platform

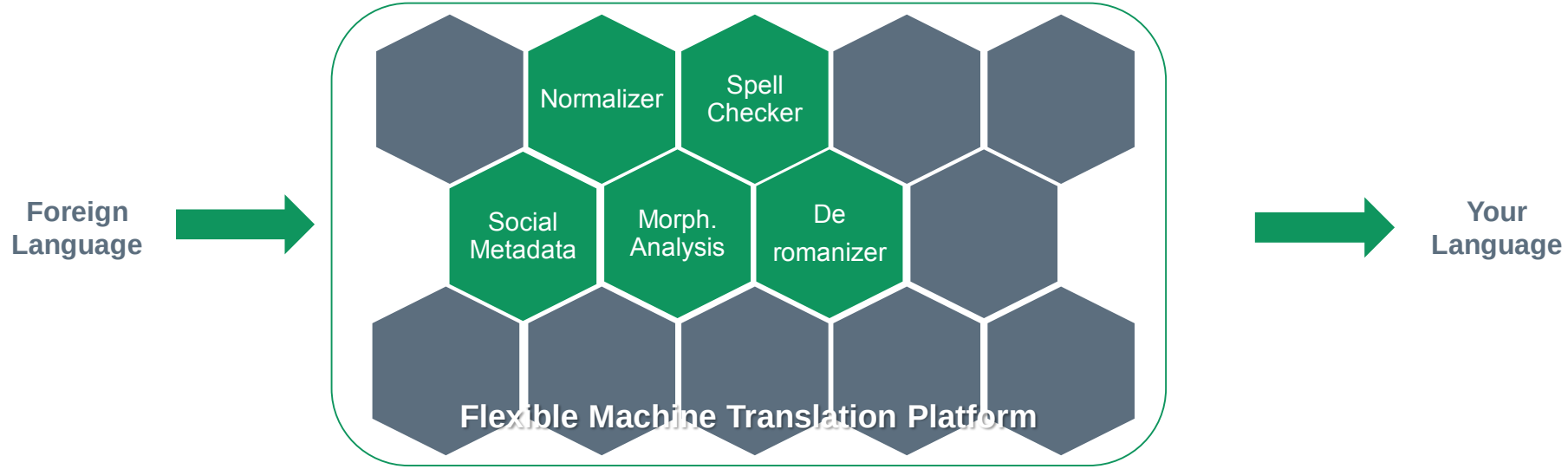
Source

- la2a hia katir fi lakhbar.
- ma 3ajbanish kida. Lazim t3'iyyer l3ounouane
- Enty habla ?
- Kalemni lama t3raf ezay tebatal teshtemni
- 3andy soda3 fi rassi... 5oshy namy badal chat. a7san lik Ah sa7

Informal MT

- No, it is very much in the news.
- I don't like this. We must change the title
- Are you an idiot?
- Talk to me when you know how to stop insulting me
- I have a headache in my head. Go to sleep, instead of chat. It is better for you, Yes, sa7

Solution with Flexible MT Platform



Normalizer, Spell Checker, Social Metadata and Morphology Analysis modules:

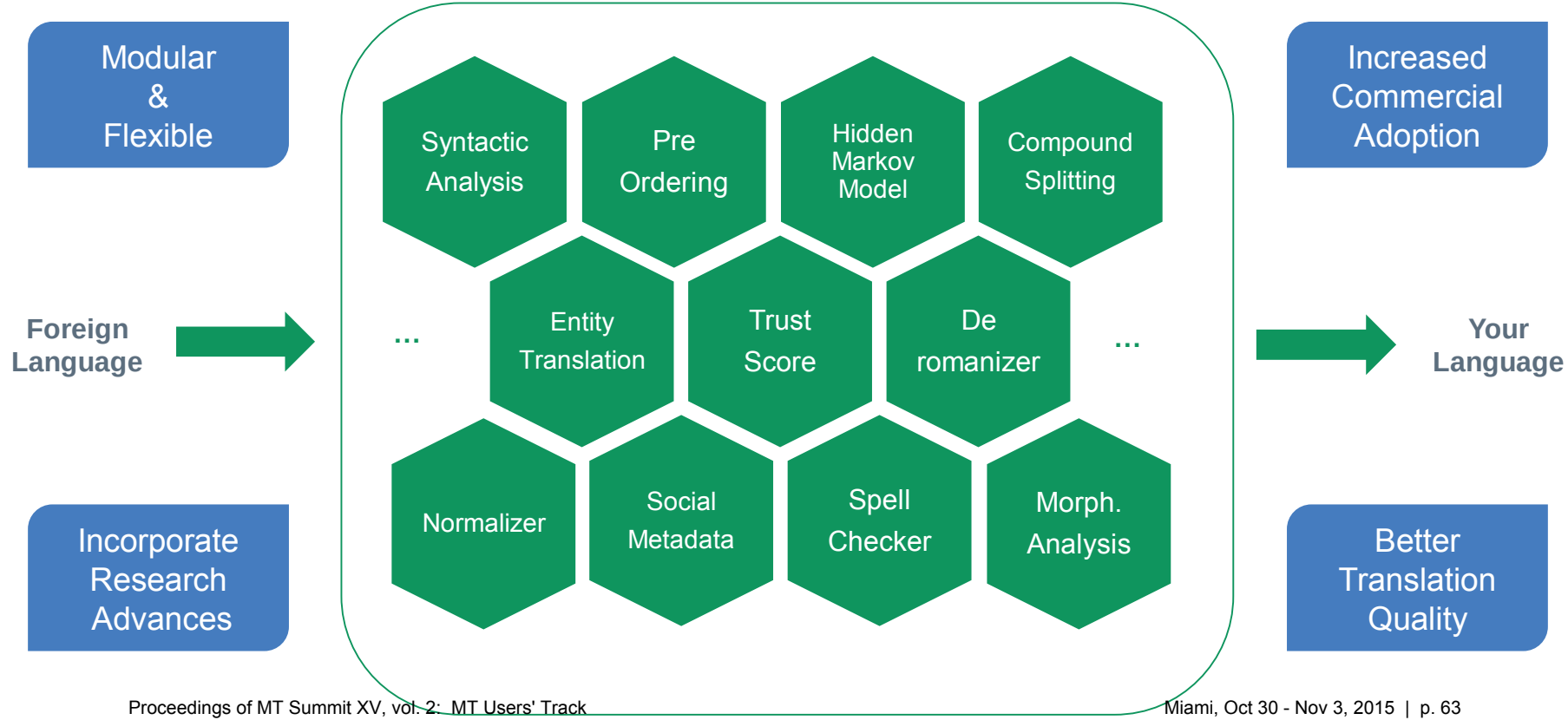
- **72% improvement in translated text**

Deromanizer module:

- **No translated text to mostly translated text**

Flexible MT at a Glance

Flexible Translation Architecture



Global Customer Experience Management

Quality Evaluation of Four Translations of a Kidney Document: focus on reliability

Abstract

This paper describes the Kidney project, which began as an experiment to determine whether human translation and fully post-edited machine translation are interchangeable and if so which is more efficient. In the experiment, an English-language patent dealing with kidney cells was translated by a professional human translator and by a commercial machine translation system. The raw machine-translation output was then fully post-edited by three other translators. Thus, four translations of the Kidney patent were available. When the four translations were evaluated by professional human translators, it was found that the evaluation results were not sufficiently consistent with each other. That is, the evaluation process was not sufficiently reliable. The focus of the Kidney project then turned to increasing reliability by analyzing evaluations linguistically to decide how to develop a revised evaluation instruments. As of September 2015 the analysis is in progress. When the revised metric is available, translators not previously involved in the project will be trained and will apply the metric to the same four translations to determine whether reliability has increased or decreased. The Kidney project is being conducted within the MQM framework (http://qt21.eu/mqm-definition), which was developed under the leadership of DFKI (http://www.dfki.de/lt/).

1. Credits

The Kidney project is a collaborative effort of the Translation Research Group at Brigham Young University (Provo, USA), and the Tradumàtica Group at Universitat Autònoma de Barcelona (Bellaterra, Spain). The main participants are Daryl Hague, Pilar Sanchez-Gijon, Kekoa Riggin, Carla Ortiz, and Alan Melby. We thank DFKI for use of MQM.

2. Some Background on the Kidney Project

This paper is an interim report on an on-going project whose focus is to increase the reliability of translation quality evaluation in a particular environment, namely, patent translation for the purpose of filing with a patent office in another country. The project described in this paper is called the Kidney project because it is based on a medical industry patent about kidney cells. However, it is hoped that the results of this project will be applicable to other translation environments, after appropriate adaptation to particular requirements.

Logically, any project involving evaluation of translation quality would begin by defining translation quality, although this is seldom done in practice. The quality of a translation, regardless of how it is produced, can be defined as the degree to which it meets agreed on specifications, so long as those specifications take into account the needs of the intended end users. Of course, some would challenge this definition. Various perspectives on translation quality are presented in issue 12 (December 2014) of the journal of the Tradumàtica group (http://revistes.uab.cat/tradumatica/issue/view/5).

The Kidney project is based on the MQM framework, which has adopted a specifications-based definition of translation quality compatible with the one in the previous paragraph.

The MQM framework is being developed at DFKI (http://www.dfki.de/lt/). See http://qt21.eu/mqm-definition for the official definition of MQM and note that MQM has accepted ASTM International standard F2575-14, Section 8, for defining structured translation specifications (see www.astm.org and search for F2575 to obtain a copy of this standard). For readers familiar with TAUS DQF (see https://evaluate.taus.net), it is relevant that in parallel with the Kidney project, MQM and DQF have been harmonized, under the QT21 project (see http://www.qt21.eu/). Thus, the next stage of the Kidney project will be both MQM and DQF compatible, and when MQM is mentioned, it should be understood as the MQM-DQF approach.

MQM has a broad scope of application. One way to divide up types of translation to be evaluated is by how a translation is produced: classic human translation at one end, raw machine translation at the other end, and post-edited machine translation in the middle. MQM is intended to apply to all three types. The QT21 project emphasizes evaluation of raw machine translation, within a larger context of developing new methods for machine translation. The Kidney project involves human translation and post-edited translation. Thus, the MQM aspect of the Kidney project and the MQM aspect of the QT21 project are complementary.

At this point, it is important to note that the MQM approach to translation quality evaluation contrasts with typical translation quality evaluation methods that use one or more reference translations and an automatic metric such as BLEU. MQM metrics do not use a reference translation but do require the involvement of a professional human evaluator. The homepage of the QT21 project (http://www.qt21.eu/) indicates that along with developing new techniques for machine translation, an important QT21 objective is "improved evaluation and continuous learning from mistakes, guided by a systematic analysis of quality barriers, informed by human translators". It is recognized in the QT21 project and elsewhere, based on widely accepted principles of assessment theory and practice, that reliability is always important and can be difficult to achieve when human evaluation is used.

Those readers familiar with BLEU and other automatic evaluation methods might ask why go back to human evaluation, after it was rejected years ago as too costly and unreliable. (See, for example, "Evaluation of Machine Translation and its Evaluation", Joseph P. Turian, Luke Shen, and I. Dan Melamed, New York University, 2006, Accession Number ADA453509). The motivation in both the Kidney project and the QT21 project for putting humans in the loop is the same: "In order to improve quality, reliable and informative quality measures are required" (QT21 project proposal). The QT21 project proposal goes on as follows: "Although very efficient for quick development of systems and for estimating overall quality, metrics such as BLEU ... are not able to work at different levels of granularity, distinguish between different types of quality problems and give any details about the nature of errors." That is, they are not informative about exactly what to do to improve the system.

The Kidney project team is not claiming that MQM-style evaluation will replace BLEU-style evaluation of raw machine translation. However, we do predict that MQM will become an important factor in evaluation of various types of translation when an informative evaluated is needed, if questions of reliability can be addressed in a satisfactory manner. Thus, the focus of the Kidney project is reliability.

3. Project Description

This section indicates where we are with the Kidney project as of early September 2015. An update will be provided at the MT Summit in late October.

Given a set of translation specifications, the MQM framework can be used to develop a customized translation quality metric. This is exactly what was done in the Kidney project. The Kidney metric is tailored according to the specifications, including the purpose of the translation, which is submission to a patent office in Latin America.

In December 2014, an experiment was conducted to investigate the use of post-edited MT to efficiently produce acceptable translations of patent applications. The experiment aimed to answer the following research question within the larger investigation: Using particular instruments, including a customized MQM metric and specialized training material, can human translators produce a reliable evaluation of the quality of human translation and fully post-edited machine translation? A related research question was whether, in this case, human and post-edited machine translations are indistinguishable on the basis of translation quality evaluation. Any solid conclusions regarding this second question require reliable evaluation and thus an answer to the first question. These two research questions are relevant to a determination of whether post-editing results in acceptable patent translations. Questions of efficiency, while important, are beyond the scope of the current investigation.

In the December 2014 experiment, an English-language patent dealing with kidney cells was translated by a professional human translator and by a commercial machine translation system. The raw machine-translation output was then fully post-edited by three other translators. Thus, four translations of the Kidney patent were available. When the four translations were evaluated by professional human translators who had not been previously involved, it was found that the evaluation results were not sufficiently consistent with each other. That is, the evaluation process was not sufficiently reliable. The focus of the Kidney project then turned to increasing reliability by analyzing the evaluations linguistically and developing a revised metric and associated training material for human evaluators. The question of testing the competence of the evaluators must also be addressed. As of September 2015 the analysis is in progress. When the revised metric is available, translators not previously involved in the project will be trained and will apply the metric to the same four translations, so that we can determine whether reliability has increased or decreased.

In total, seven human translators took part of the December 2014 phase of the project: while four of them participated by translating and post-editing the patent respectively, the other three participated by evaluating the human translation and the fully post-edited machine translation.

4. Discussion of the Results

We are currently in the linguistic analysis phase. The Kidney project team is looking at the first 300 translation units. We have three evaluations of the human translation, and we are examining the differences in how the evaluators annotated each translation unit. In the majority of the translation units, the three evaluators completely agreed. That is, they either all three indicated that there were no errors or all three indicated that there was at least one error and agreed on what the error was.

We are now examining in detail the translation units where there was disagreement among the evaluators. For example, the phrase "prepared from a human kidney-derived cell" appears several times in the patent. There is some debate about the relationships among the constituents and how they affect a translation into Spanish. Is the cell derived from a human kidney or is it a human cell derived from a kidney? Does it make a difference to a patent examiner? Another example is how the linguistic expression "such as" is translated into Spanish in various contexts. Is there any agreement between this expression and other

elements of a sentence? A third example is how the word "removed" should be translated into patents in various contexts.

5. Further Work

The Kidney project is far from over. Once we have completed our analysis of the disagreements among the evaluators of both the human and post-edited translations, we will revise the translation quality metric, taking into account the recently completed MQM-DQF harmonization, and improve the training and screening material for evaluators. For one thing, we will give the evaluators access to terminology database. We will probably also develop a tool to help the evaluators deal more efficiently and consistently with multiple instances of the same error. Then we will run the evaluation portion of the December 2014 experiment again, this time with a new set of evaluators who have not yet been involved in the project.

Hopefully, an analysis of the second evaluation of the same four translations will reveal more reliable results and the techniques we use to increase reliability in the Kidney project will apply to other environments where translation quality evaluation needs to be informative or where there not reference translation is available.

A Survey of Usage Environment of Machine Translation by Professional Translators

Tomoki Nagase — Fujitsu Laboratories
Tatsuhiro Kudoh — SunFlare
Katsunori Kotani — Kansai Gaidai University
Wenjun ye — Crosslanguage
Takeshi Mori — NTT Media Intelligence Laboratories
Yoshiyuki Sakamoto — Tokyo University of Information Sciences
Nobutoshi Hatanaka — Intergroup
Takamitsu Takeda — Universal Content
Shu Hirata — Nagoya University
Hiromi Nakaiwa

About Asia–Pacific Association for Machine Translation (AAMT)

AAMT is the organization that aims at development of machine translation(MT) in the Asia-Pacific. The members of AAMT include researchers, manufacturers, and users of MT. The organization executes evaluation, enlightenment, promotion, and standardization of machine translation system.

Activities

- Holding seminars for MT users
- Executing actual condition survey for MT developers
- Improving test sets for quality evaluation
- Working out specifications of descriptive forms of user dictionary and implementing its standardization
- Publishing “AAMT Journal” which is an in-house magazine
- Holding international conference: “MT Summit”

About AAMT's Surveys on MT Usage

<Method>

Surveying actual conditions of MT usage in business translation(*)
Making announcements of their results in academic conferences [1) 2)] and symposiums

<Objective>

- To MT vendors:
  Facilitating development of services and products that are more convenient to business translators
- To Business Translators:
  Providing tips to use MT more efficiently

(*)*Business Translation*
The work of persons who devote themselves to translations in translation vendors, enterprises, government, and municipal offices

About AAMT's Surveys on MT Usage

<Achievements of past surveys>

- Carrying out online questionnaires via Internet concerning usage of machine translation (annually executed until 2011)
- Asking participants to fill in questionnaires on the spot at show exhibitions (2012 onward)

This report introduced the analysis result of questionnaires for Business Translators which were carried out in 2013 and 2014. These questionnaires were executed by AAMT assignment committee for seeking future direction of MT.

Outlines of this Survey

<Method>

Handing out questionnaires to translators and asking them to fill out the sheets on the spot

Questionnaire form : A sheet of paper, printed in both sides, and anonymous. It is distributed and withdrawn on the exhibit sections.

<Date>

First time: 27 Nov. 2013
Second time: 26 Nov. 2014

<Location>

At the event of Translation Festival ("Honyakusai") which was held by Japanese Translation Federation (JTF), an industry organization of business translators.

<Response Rate>

About eight percent of the total (presumptively eight hundred persons in a year)

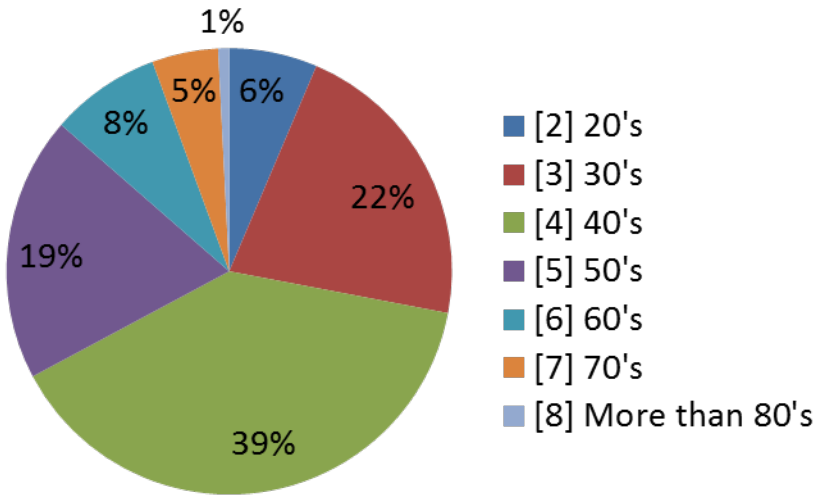
Types of Respondents

Types of Respondents	Number of respondents (people)	
	FY 2013	FY 2014
(1) Freelance Translators	17	18
(2) In-house translators in translation vendors	7	6
(3) Translators of companies (except translation vendors) , governments , etc.	11	11
(4) Translation project managers or coordinators of the translation vendors.	15	15
(5) Translation project managers or coordinators of the companies (except translation vendors) , governments , etc.	6	4
(6) Others/ No answer	8	8
Total	64	62

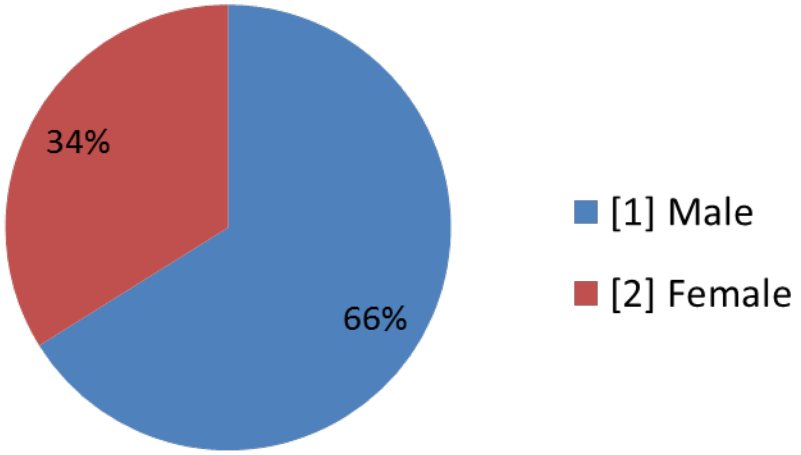
Among 126 people of the respondents (FY2013-2014) :
①Translators：70 people（55%）
②Translation project managers：41 people（33%）
→In this report, answers of ①Translators are tabulated/analyzed.

Properties of Survey Respondents(Translators)

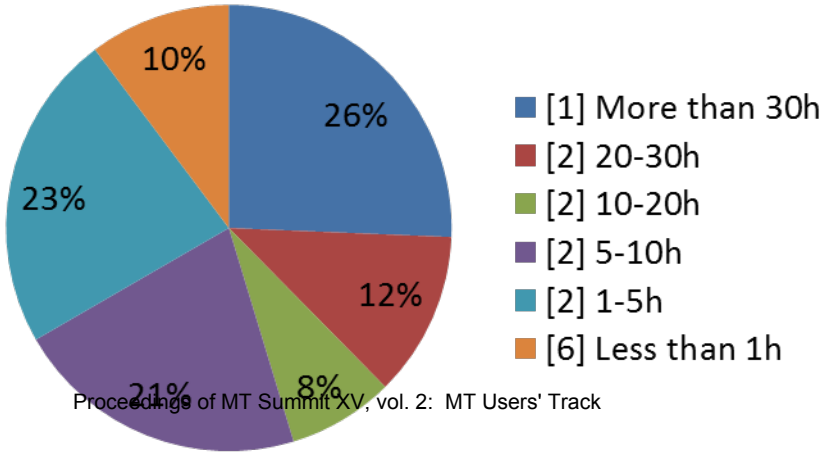
Age



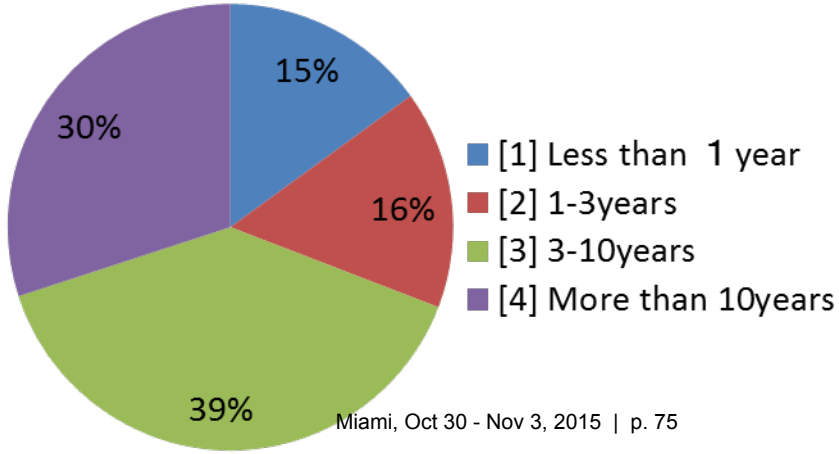
Sex



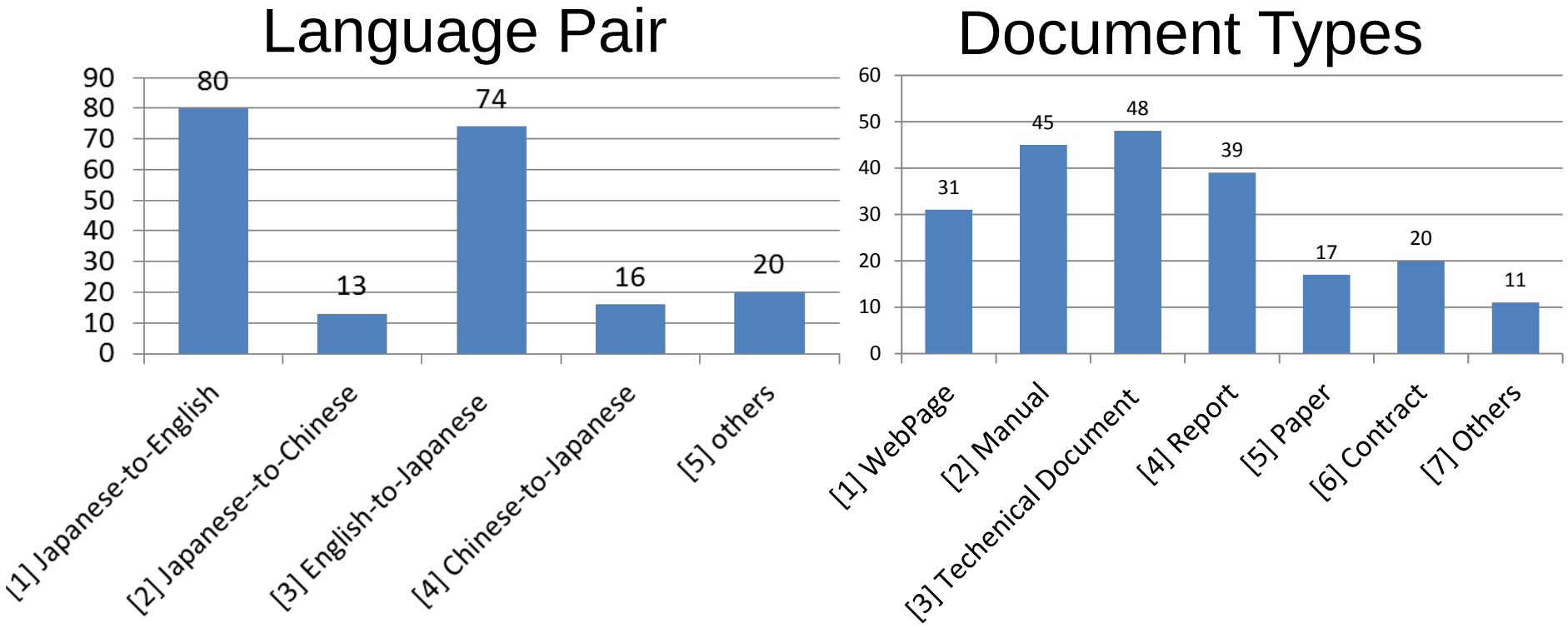
Time for translation work for a week



Work experience

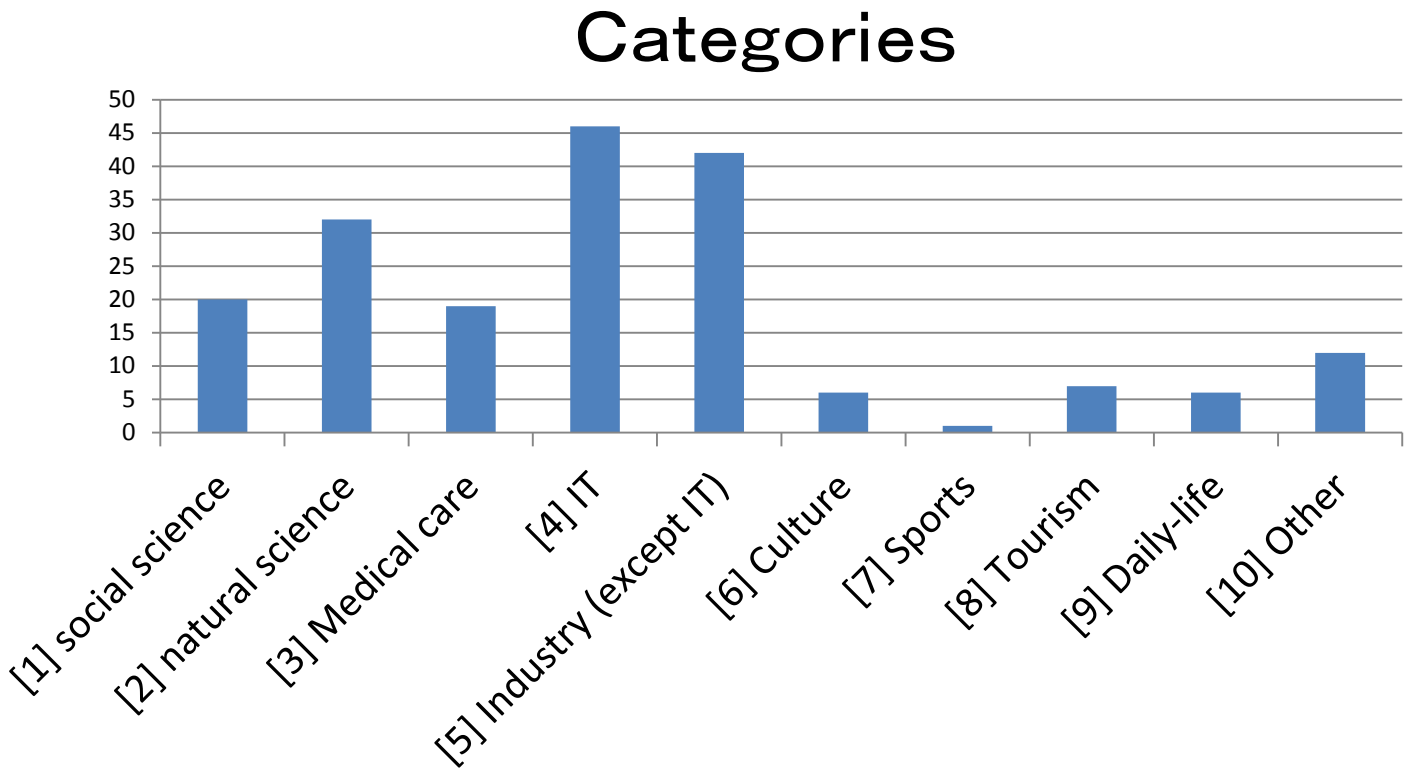


Translation Tasks



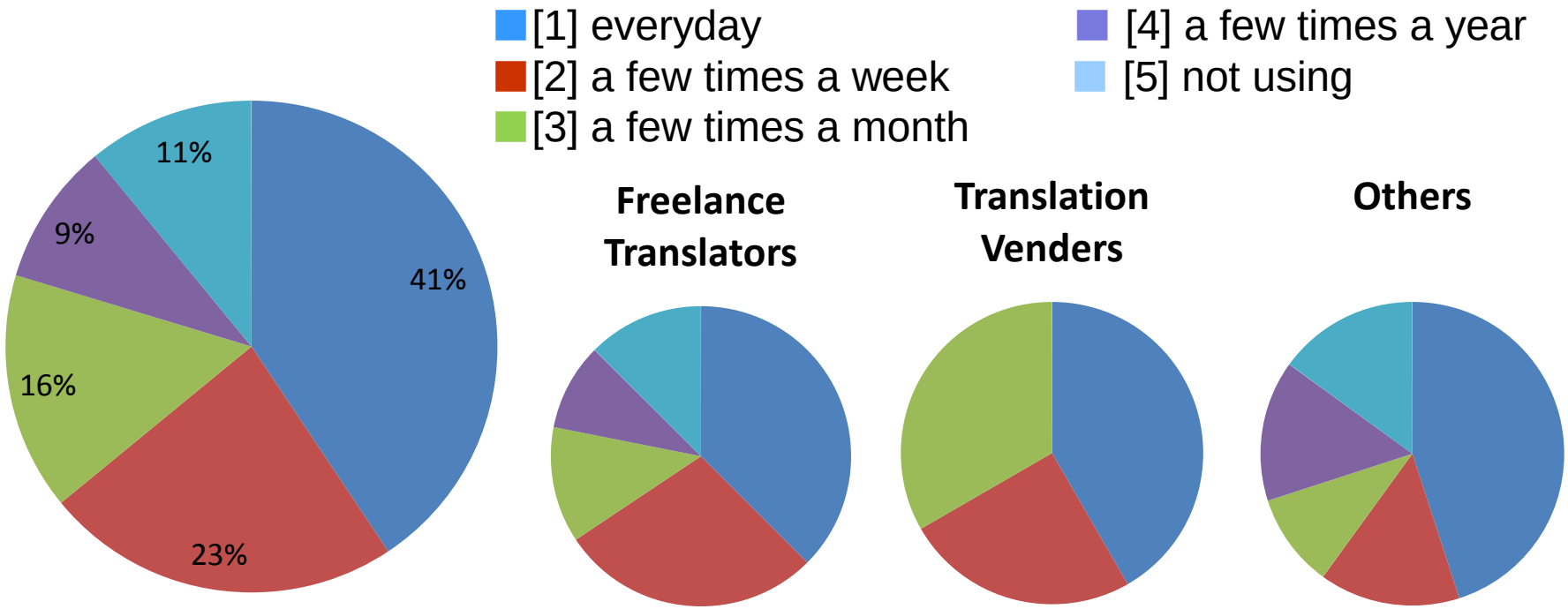
- ✓ High demanded translation directions in Japan are between Japanese and English.
- ✓ J2E, E2J, J2C and C2J occupy 90% of the whole work.
- ✓ As types of text, high demanded document types are technical documents and manuals.

Translation Tasks



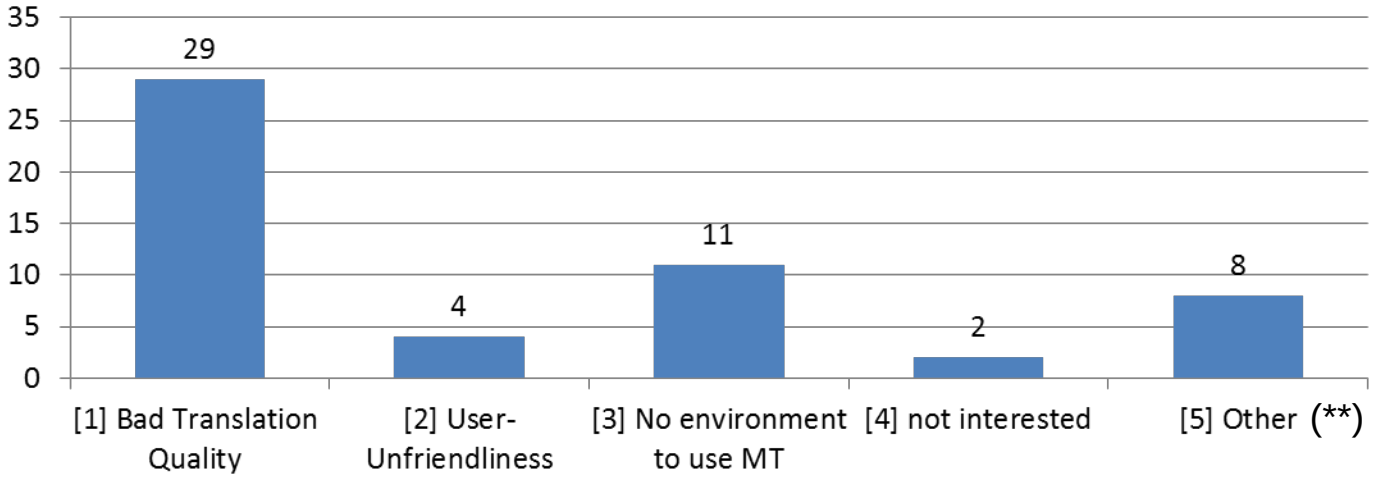
- ✓ As Categories, high demanded fields are Industry and IT.

Frequency of MT Use



- ✓ More than 40% of the respondents(26 people) answered they use MTs almost every day.
- ✓ Only 11%(7 people) answered they do not use MTs.

Why they do not use MTs

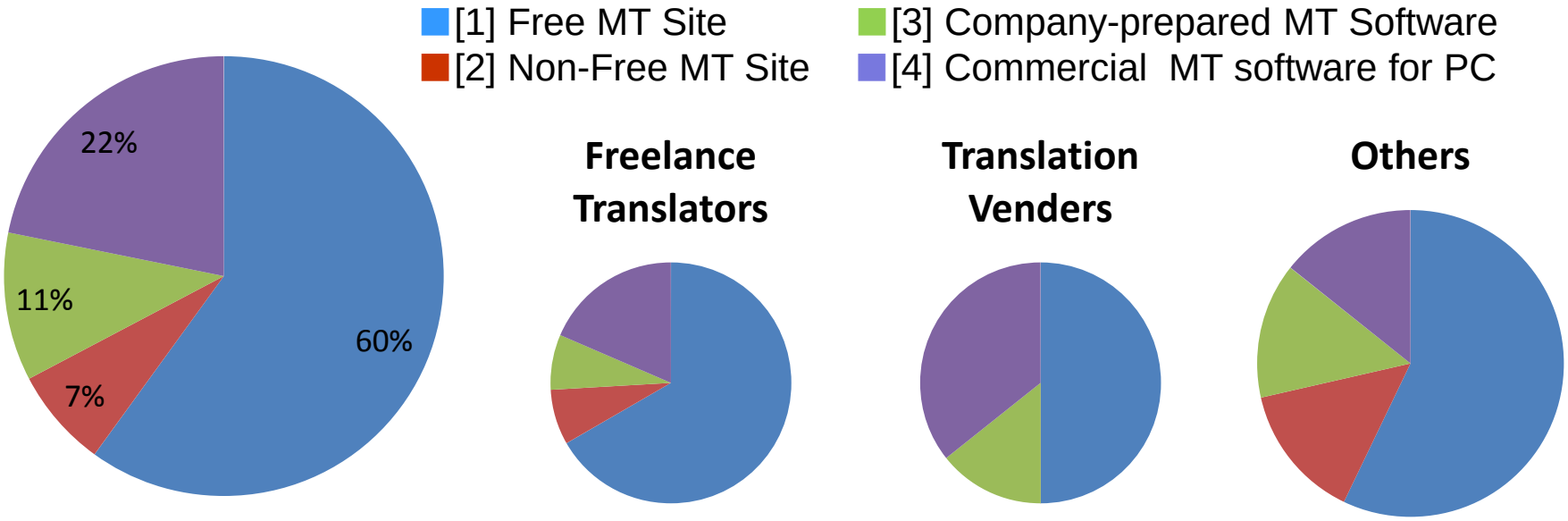


- ✓ The first reason is bad translation quality.
- ✓ The second reason is that they have no choice to use MTs because their company do not have MTs.

(**)The other reasons are:

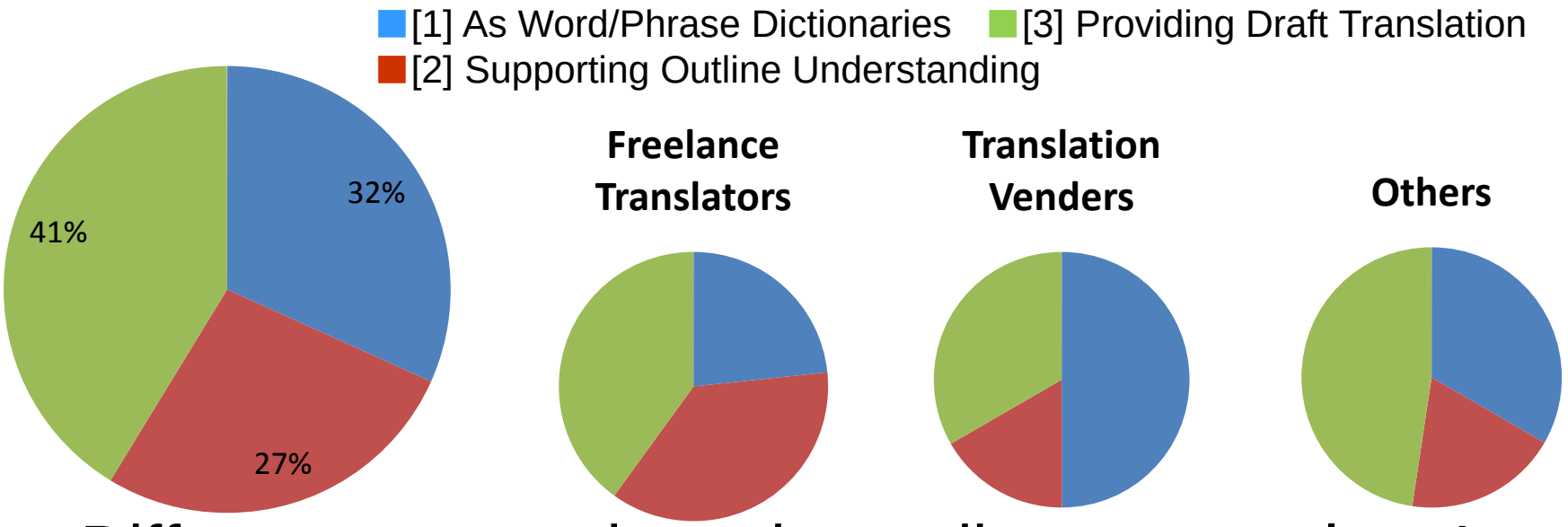
- now under construction or under consideration. (translation company)
- their departments prohibit MT use.

Types of MTs Used by Translators



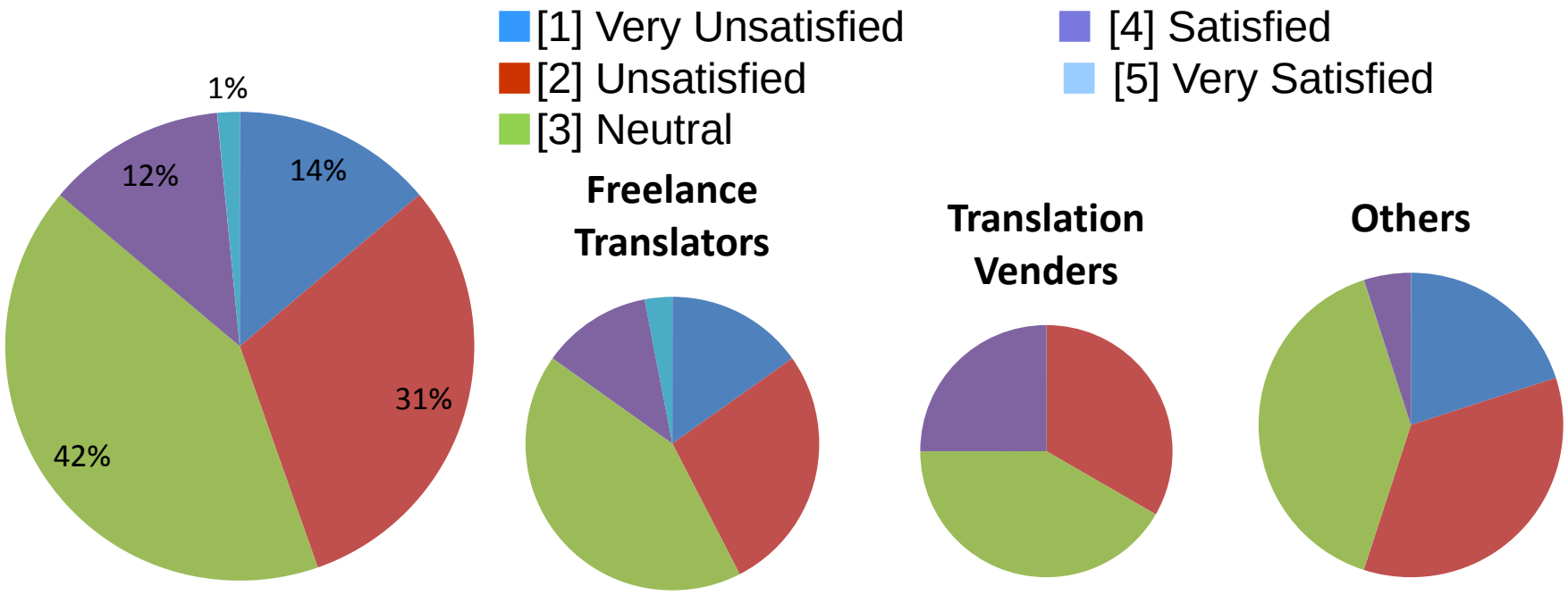
- ✓ Free MT site is the most presently available
- ✓ About 70% of freelance use Free MT sites
- ✓ As for Commercial MT software, the proportion in translation company is twice of others

Expectations for MTs



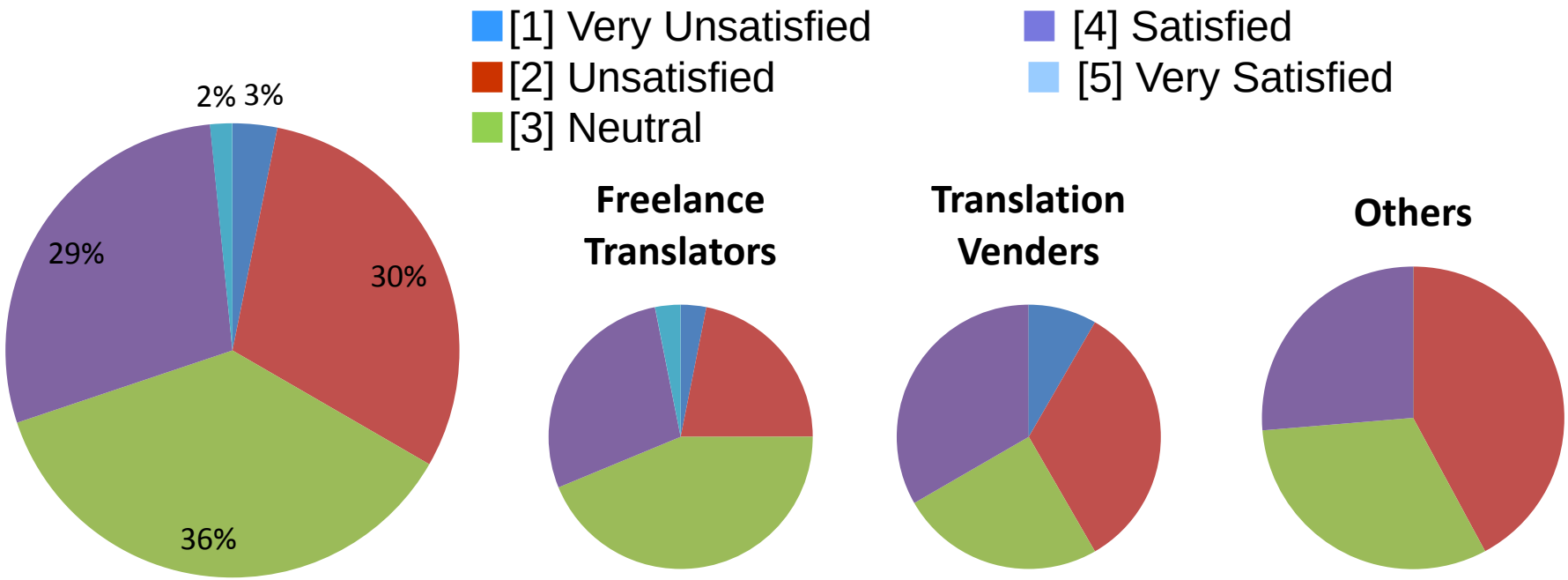
- ✓ Different expectations depending on translator's type
- ✓ Providing draft translation is most expected in Freelance and others
- ✓ Translation venders expect using as dictionaries

Satisfaction of MT's Quality



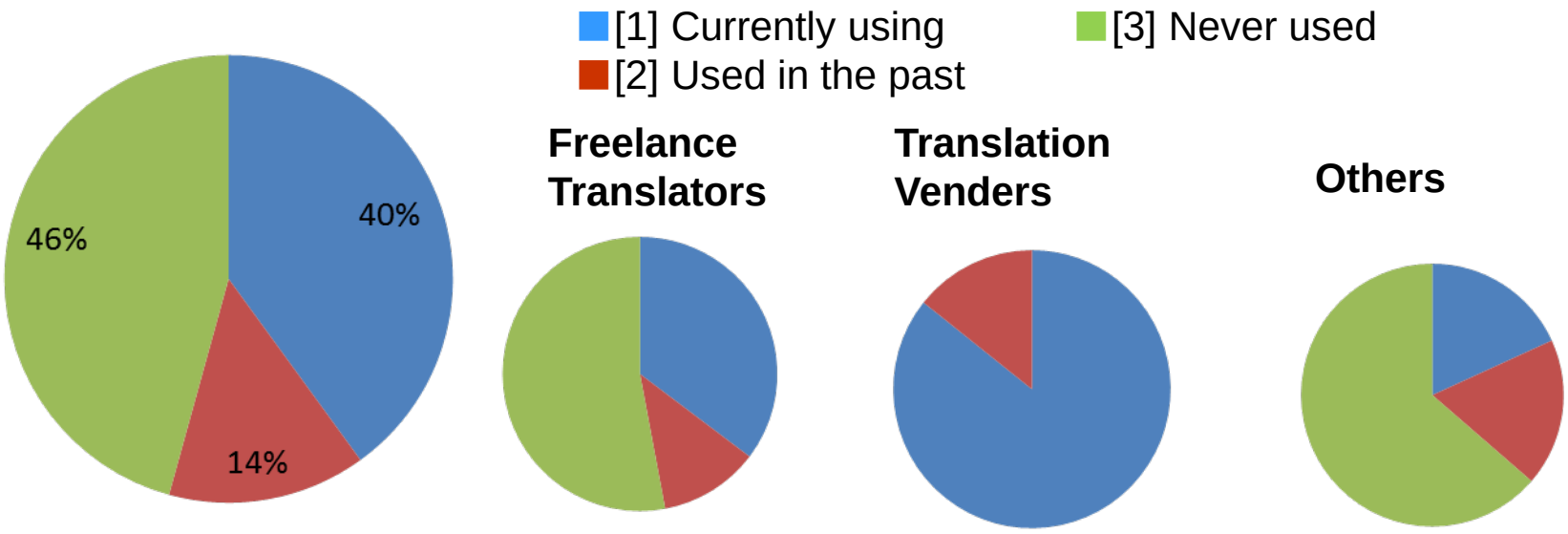
- ✓ Almost half of respondents feel "unsatisfied"
- ✓ "Satisfied" rate in translation venders is higher than freelance translators and others

Satisfaction of MT's Function



- ✓ MT's function is more satisfied than its quality by translators
- ✓ On average the rate of "satisfied" is almost same as that of "unsatisfied"

Usage of Translation Memories (TM)



- ✓ 46%(16 persons) of respondents answered “never used” but 40% answered “currently using”
- ✓ All respondents in translation venders (except one) are currently using TMs

Rate of Using TM on each Document Type

	[1] Webpage	[2] Manual	[3] Technical Document	[4] Report	[5] Paper	[6] Contract
(1) Currently using	7	16	12	5	3	2
(2) Used in the past	2	2	4	2	2	1
(3) Never used	5	3	7	5	2	2
Rate of using TM	50%	76%	52%	42%	43%	40%

- ✓ The rate of using TM on translating manuals is notably high

Relations between MT Type and TM usage

	Currently using TM	Used TM in the past	Never used TM
(1) Free MT Site	9	3	19
(2) Non-Free MT Site	1	1	2
(3) Company-prepared MT Software	2	0	1
(4) Commercial MT software for PC	10	2	0
(5) Not using MT on business	5	3	5

- ✓ 70% of translators answering “never used TM” are using free MT site
- ✓ 83% of translators using commercial MT software are also using TM
- ✓ Translators using only TM account for 13%

Finding in the International Survey[3)]

Use of TMs

- ✓ 90% of the respondents own TMs
- ✓ 70% use TMs that the clients own

Use of MTs

- ✓ 50% use MTs
- ✓ 10% own MTs
- ✓ 40% use MT available for free

Other question: “Have you ever accepted projects in which you were given a raw machine translation output to revise?”

Yes: 25%

No, but I would take them: 15%

No, and I would not take them: 65%

[3)]IAPTI (International Association of Professional Translators and Interpreters) report in 2011

Finding in the International Survey[4)]

Why use MTs?

- ✓ Typing aid
- ✓ Source of inspiration for alternative translations available in TMs
- ✓ Quick draft for improvement
- ✓ Help for heavy workloads

Why do not use MTs?

- ✓ Need to know more about MTs
- ✓ Poor quality
- ✓ Severe mistakes

[4)]paper published in Languages and Translation, vol. 6 (Fontes 2013)

Summary（About MTs）

- More than 40% of translators use MTs (mainly on free translation sites) every day

  ⇒ The same trend is shown in an international survey

- Only 10% of translators are satisfied with MT's quality

  ⇒ MT's quality should be improved to meet translator's needs

- Only 11% of translators don't use MTs but about 30% of users expect MTs not for quick draft but for searching dictionaries

  ⇒ Knowing more how MTs work is important to spread the use of MTs

Summary（About TMs）

- Only 40% of translators are using TMs in Japan
  - ⇒ In the international survey, 90% of them own TMs
  - ⇒ Japan lags behind Western countries in TM usage
- Most of translators using commercial MT software in translation companies are users of TMs
  - ⇒ Further investigations of combined usage of TMs and MTs are required

References

1) Sakamoto,Y. and Moriguchi, M. 1999. Report on Machine Translation Market in Japan. *Proc. of MT Summit Ⅶ, pp92-99.*

2) Nagase, T. et al. 2014. Report on Machine Translation Usage in Business Translation. *Proc. of Annual Conference of The Association for Natural Language Processing (in Japanese)*

3) Piroth, A. 2011. Translation automation survey among translators. *IAPTI (International Association of Professional Translators and Interpreters) report*

4) Fontes, H. et al. 2013. Evaluating Machine Translation: preliminary findings from the first DGT-wide translators. *paper published in Languages and Translation, vol. 6*

Machine Translation for enterprise technical communications – a journey of discovery

Morgan O'Brien mobrien@mcafee.com
Ana Duarte ana_duarte@mcAfee.com
Global Language Services, Intel Security, Cork, Ireland

Abstract

There has always been a high quality requirement from large corporations regarding machine translation due to perceptions and common beliefs which in turn makes it difficult to break into the area without well-maintained engines and processes. This project details from the combination of various internal efforts towards automation in translation of technical communications. Namely working with external language providers/partners, testing the offerings, understanding the nature of our source content and the inclusion of machine translation technologies for rolling out Post-Editing Machine Translation (PEMT).

1. Credits

This paper is derived from research on tools, technology and processes since 2012 towards the implementation of Machine Translation (MT) within the Localization workflow of Intel Security. It takes input from machine translation service providers, translation vendors and post-editors, language quality team, localization professionals within the company and the various departments where machine translation helps their productivity and allows them to reach their target customers and a wider audience.

2. Introduction

If one was to listen to sales pitches from various MT providers, one could learn a lot. It is noticeable that many claim their system is better in some unique way when compared to other systems or the general knowledge in the area. Too much focus on the individual selling points will distract the receiver from possibly more important pitfalls of the rollout process. To separate the jargon from the relevant, one needs to take a step back and look at content from its conception to its consumption and analyze the supply and demand of it.

Quite often, the place to start isn't on the MT, but the internal content, the tools and the people: the three main pillars we need to "shape" in order to get acceptance for MT and effect whatever systems are needed. We selected 5 different content types; Product Documentation, Knowledge Base, Community generated content, Global Definitions database and Product UI. Our main focus was on Knowledge Base articles and Product Documentation as these were two areas where large corpus existed for MT training and structured authoring teams were in place. Below we will follow the discovery process on our initial testing of content to MT. We will detail some of the tests and the systems that need to be altered and provides some recommendations based on the lessons learned. The main output from this is to be our MT strategy to a PEMT rollout.

3. Source, Target and Speed

The first time people start learning about MT there can be a lot to digest. Confidence scores are quoted as unique selling points, BLEU scores are proudly displayed as a metric of quality. MT providers offer pre-ordering, automatic post editing and domain adaptation which they say increases their quality, and there can be decisions to be made around Statistical Machine Translation (SMT), Rules Based Machine Translation (RBMT) or Hybrid methods which are a combination of systems. It can't be denied that all these things play a part as there are studies that prove this. But how much impact do they really have effecting the ability to roll out high quality MT? Where does attention need to be paid, and what are the priorities? It can be hard to tell at the beginning of an MT program.

Having gone through the required ramp up of technical knowledge with the many great online resources and taken advantage of talking to peers in the industry, it comes to a point when actions need to be taken and one must choose a method to proceed with. The focus for our tests was on PEMT and 3 areas stood out for measurement; Productivity, Target Language Quality and Source Language Quality. These were largely driven by internal requirements for Cost, Speed and Quality.

It was important we looked at Productivity as demonstrated in Post-Editing studies like Plitt, M., & Masselot, F. (2010) where post-editing substantially increased their productivity when compared to Human Translation (HT). This paper the authors talk about productivity measured in Time, and this seemed a reasonable starting point. We then found a number of tools available to do this such as iOmegaT where timing data is measured in a desktop translation tool. One of the advantages of this tool was the ability to track time per segment, but also revisits to segments, which give a clear picture of how much effort was given to each Post-Edited segment. It would be considered normal practice in Translation and Post-Editing for the translator to revisit segments once context became clearer while translating similar segments. Also as iOmegaT is a desktop Computer Aided Translation (CAT) tool, it is closer to the native working environment for translators reducing that variable from the tests. We also mixed the post-editing task by creating a project TM with segments to be fully translated (no MT) and also inserted some previous TM matches for segments to be leveraged/reviewed. The final measurement we wanted from this was how many words were Post-Edited in 1 day relative to how many were translated using the same environment and project. Again, "words per day" is a standard metric in our business for forecasting translation time in projects so it was important to leverage something that is already generally understood. We displayed the throughput relative to the 4 final engines we were testing (4 languages). The final throughputs (Fig. 1) were recorded where the minimum quality bar was met (80% pass mark for LQA).

	#1	#2	#3	#4
Doc	4982	7292	8826	5455
KB	4743	18262	4461	5176

Figure 1: Throughput per day (words) for Doc and Knowledge Base content

We needed to measure quality of target language and for this we already had Language Quality Analysis steps in place. Our LQA score is another measurement that exists in our business day to day, so again it's something that people in the company already understand. This quality measurement is based on the LISA LQA model and the results are in the form of a chart with an overall score out of 100. While the LISA LQA model will do for general quality assessment we also needed something more specific and repeatable. For this we complimented LQA with an Edit Distance measurement on every segment, giving us some drill-

down data when investigating problematic segments later on. We displayed this on a graph to demonstrate where most of the effort was for Post-Editing.

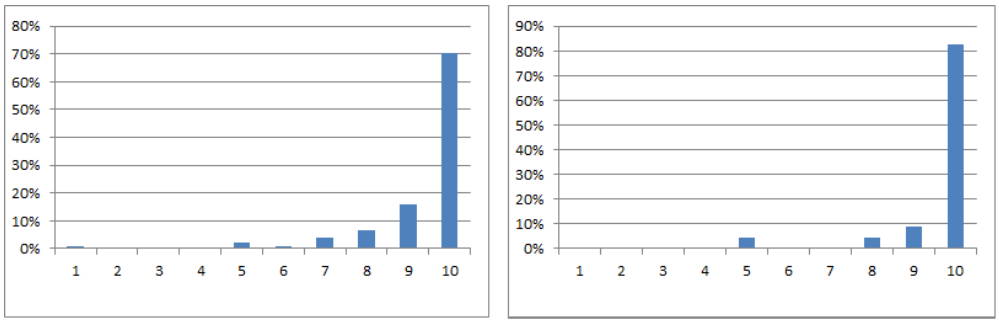


Figure 2: Edit distance split for Documentation and Knowledge Base content

Finally for an end user confidence we did some usability studies on samples of MT where users scored segments on a scale of 1 to 4 where 1 was bad and 4 was good. During our studies we did notice that automated metrics such as BLEU and Edit Distance correlated somewhat with our human usability tests on individual segments but lesser so with a usability test done by a trained linguistic reviewer when looking at the overall project. We put this down to individual strings such as long strings (in excess of 15 words) which caused issues due to writing style.

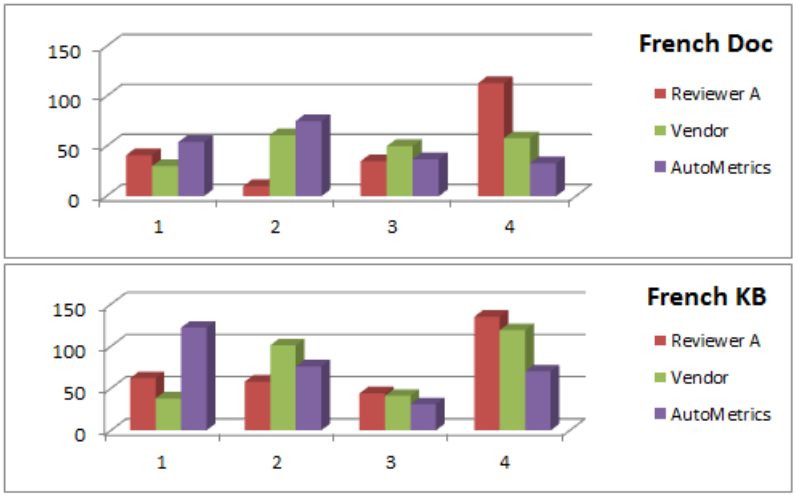


Figure 3: Usability results for Documentation and Knowledge Base content

We have some ability to control standards in our Technical Authoring process, so studies such as Roturier, J. (2004) on Controlled Language (CL) rules effect on MT systems also gave us inspiration for measuring source appropriateness. The idea is that if you have good controlled source authoring style and terminology, then Machine Translation will work better. To understand the nature of this within the company we undertook the task of rewriting some source content to be in a controlled language style. We used this in addition to our normal

source content to be Machine Translated for benchmarking in the discovery stages of project. The CL rules we used were based on a limited standardized terminology set and some other basic rules such as sentence length. The two types of content we put into a Controlled Language were standard Technical Documentation; Software Doc/Help and Knowledge Base Articles. It should be noted that by creating new source content and style we must expect some impact on the MT statistics as the non-Controlled Language style is used previously to populate the Translation Memories which in turn are used in training the MT engines.

Using the Edit-Distance, we counted how many segments did not need editing or needed only a low amount of editing and we could see that there was a higher percentage of 100% Match or Fuzzy Match segments that did not need to be Post-Edited with the content rewritten for Controlled Language. This already showed a clear difference between normal authoring and Controlled Language Authoring with regards to the effectiveness of the MT system when displayed across 4 different MT systems. Fig. 4 pertains to the final 4 engines being tested relative to the throughputs recorded in Fig. 1.

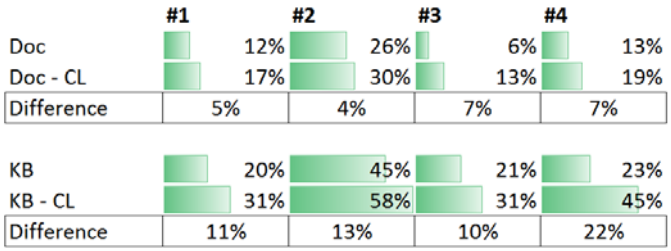


Figure 4: Percentage of 100% matches for Content v's Controlled Language Content

It could be said in hindsight that some of our testing was unnecessary. We used a number of other tools that were readily available such as Reading Ease metrics (Flesch-Kincaid and others), measuring the segments and words, average words per segment etc. We wanted full visibility on anything we could measure that may affect output and studied these source metrics on various content types. Despite these measurements not being immediately necessary, they were easy to do, and the lessons learned during this phase do help in the future such as in the ability to notice a high level problem in the authoring process if the numbers move greatly on a particular topic within the Content Management System (CMS).

Corpus Analysis

Source Corpus	Technical Documentation/Help	Knowledge Base Articles
Flesch-Kincaid Reading Ease	51.5	48.2
Grade Levels		
Flesch-Kincaid Grade Level	9.8	9.3
Gunning-Fog Score	11	9.6
Coleman-Liau Index	13.5	13.5
SMOG Index	9.5	8.8
Automated Readability Index	9.6	7.5
Average Grade Level	10.7	9.7
Text Statistics		
Character Count	11,529	11,685
Syllable Count	3,835	4,089
Word Count	2,319	2,345
Sentence Count	153	214
Characters per Word	5	5
Syllables per Word	1.7	1.7
Words per Sentence	15.2	11

Figure 5: Readability Metrics for Documentation/Help and KB Articles

3.1. Make your Enterprise change

A large company does not change processes at the speed of light, it changes slowly and that occasionally makes you actually wonder if it is changing at all. To be the one who tries to turn the enterprise ship can be a daunting task. At the start of discovery in MT it is important to act as an Influencer. This role is to point out areas that could change and the improvements that could be made, find reasons where MT could help and see where people react. Build up buzz around the topic, prove some results and educate your colleagues. These approaches help show what life might be like in the future with a new practice and may demonstrate the potential value of change. Through this movement, followers join your cause and MT will start to go from a topic of conversation through to being involved in projects. People who believe in you and in what you are doing are allies you need to gather in order to make MT a reality. Every conversation potentially helps the cause as many pre-conceptions can exist due to peoples personal experiences with Google Translate and such.

The Enterprise requires due diligence, so every step towards rolling out Post-Editing for productivity should be layered with tests, discussions and some time for people who are not living in the MT or academic world to consume and understand the results. To help in this we have used standard metrics for our organization and we continued this with use of the "Trados Grid" which has an industry understood breakdown of matches in a TM. In Figure 6 we used a GNU license application called KNIME which has some ability for custom workflows of text analytics.

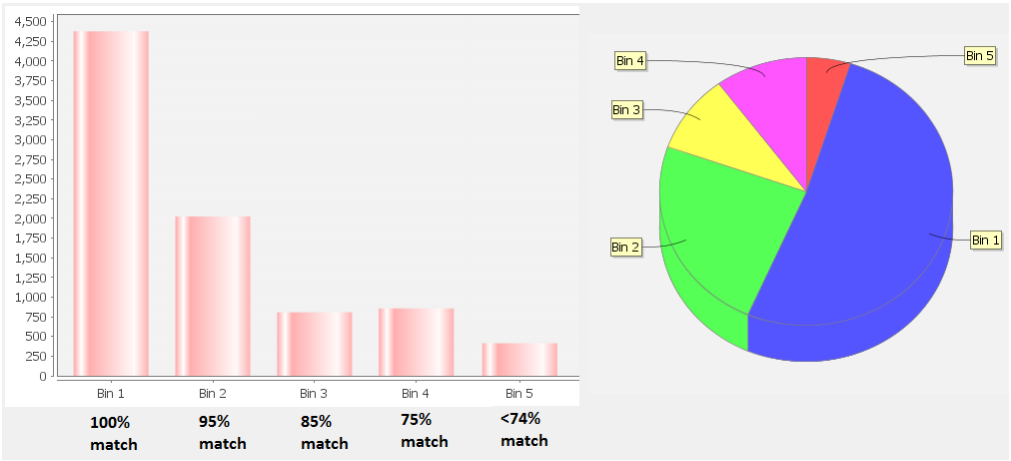


Figure 6: MT Distance matches shown in standard Trados breakdowns using KNIME

It can feel like an uphill battle sometimes to get organizations to change. There is so much to prove to ensure the business case. Luckily there are some very useful resources which can help you prove the theories you preach. The MT suppliers often provide excellent information that can be reused and you should ask them for this if in doubt. Research and white papers can be very useful and many of the MT users meet, collaborate and share their experiences at conferences and online. A newcomer to this area could do well to make friends and ask questions as it can be through these connections that you may build confidence in your own ability to make the best steps forward. Not every approach works for every person or company. Compare, focus and learn the appropriate subjects that you need and ultimately help guide your company towards the right path.

There has to be a need or a gap that can be filled by using MT technology. The first thing that should be done is to identify what the specific business needs are. Having more than one business need gives you a platform to build a proposition to show value and Return on Investment (ROI) in this area.

Business needs for MT in localization are born out of content and publishing. The content gets created and needs to reach an audience. MT can boost the efficiency of that effort through a number of different ways and these are the high level unique selling points for your internal customers that you need to find and understand. Some basic business needs are:

- Increase productivity of translation (Plitt, M. and Masselot, F. 2010).
- Allow on-demand translation for content that normally does not get translated.
- Enable internal users to have access to a larger set of content in their language (Burgett, W., Chang, J., Martin, R. and Yamakawa, Y. 2012).
- To speed up a process of collating sentiment analysis from content.
- Help understand the “gist” of text not available in your own language.
- Enable early versions of localized Documentation or Software.

For the purpose of focus in this paper, we are looking at the productivity of translation as it is an area that can show immediate financial savings. Through increasing productivity many lessons will also be learned and skills gathered necessary for many of the other areas of the value proposition while aiming to save money and time.

3.2. Is MT all that is needed?

This is not only about the MT system as already mentioned earlier in the paper. In some regards it is not even about the MT system itself as the investment in quality of commercial MT systems has been good over recent years, and they continue to get better. The internal workflow is probably the area where most change must happen. This affects a set of items from content writing and curation through how you manage your bilingual content and right down to the end result of publishing and the feedback loop back to the content creation and MT maintenance.

Mentioned previously, Controlled Language Source is probably the most important area to start making changes as it can have a massive impact (Roturier, J. 2006) (Doherty, S. 2012) if it is done right. If you are thinking about rolling out MT in your company, you should start here. Who is writing content that will eventually end up being machine-translated? Is the content good for translation? Does anyone need to change their work practice? After all, in a Globalized company, the source content is most likely only a small percentage of the content distributed to customers around the world. If we can control our language, all the target languages will benefit.

Some basic rules on the content creation side can have a great impact, and consequently, effect on throughput and accuracy in the future. We did some after the fact analysis on Distance per segment and noticed some patterns. The basics that seem to make a big difference are:

- Managed and maintained terminology for authoring – reduction of synonyms
- Basic style rules – keeping all authoring similar
- Reuse of repetitions and phrases in writing
- Source content profiling your authoring into groups (Domains) for MT systems.

Translation process is the next area that needs attention. At the end of the day, the translators are your direct link to your market, and to ensure the best language quality possible and the most accurate message possible, they need to be included in your plans. On paper you may have MT systems with high BLEU scores, but does this become good PEMT in the end? The most important factor towards good quality of PEMT is the translator. In our initial PEMT tests we identified a wide discrepancy in results of translator productivity and quality.

SPANISH	Vendor A	Vendor B	Vendor C
Doc/Help Throughput / day	10688	15280	21360
KB Throughput / day	7464	14088	28392
Quality (LQA Score)	93%	35%	66%
Years experience with Intel Security	5	0	3
Years experience as Translator	7	3	3
Years experience as Post-Editor	1	3	3

Figure 7: Throughput PEMT Spanish with basic Translator Profile info.

After analysis it seemed that the one variable in the process was the individual doing the post-editing. We could not effectively baseline results from one group of post-editors to another with this variable so we needed to reduce or eliminate it and started looking at the concept of Translator Profiling. We would like all translators to translate at the same rate and produce the same quality. This just isn't the case. So there are several parts of a profile that can vary the results, such as individual motivation, or when using freelance or crowdsource

models where profiling isn't possible. But as quality is the main requirement for our PEMT there were a number of factors that stood out as a requirement for Translator Profiling for us:

- Experience as a translator is important.
- Post-editing experience is less important (but needs to have some. 2 years is good)
- Age is not necessarily a factor, but (technical) ability to leverage tools might be
- Understanding the content subjects is the most important aspect to reach quality.

What this basically means is that PEMT resources are needed who have spent a good amount of time working on your content so that they understand both your content and your quality expectations. This is evident from Vendor A who has experience on our content and scored high on quality, but lower on throughput. But Vendor B and C did not meet our quality expectations (despite Vendor C having 3 years with our content). The number of years' experience in post-editing is important to throughput with Vendor B and C retaining very high productivity but Vendor A was slower.

The workflow is also an important area to look at. Translation Management Systems (TMS) exist with MT plugged in through APIs. There are other decisions you need to make for your workflow though. Can you trust your translators to not make mistakes such as missing a file for translation? Do you review to ensure no mistakes? Do you allow PEMT segments back into the Translation Memories (TMs) of your main products? Does your TMS edit content before going to the MT system (such as protecting tags or internals)? Do you apply a match penalty on your MT segments and by how much? There is no quick answer for these questions. It is crucial then to understand the nature of the content you want to MT (source) and the nature of the market where you want to publish it (target and quality). Basic localization decisions from normal workflows may need to be rethought when you include MT into the translation strategy.

3.3. Testing MT

Ultimately you will need to test the MT output. Whether you create the MT systems yourself or use a service, or even outsource completely, it is imperative that you run a test. What you want to achieve from this test is a confidence that the standard of quality is high enough on the output for post-editing to happen with extra efficiency.

- Productivity
- Quality (Automatic Estimation)
- Quality (Human Evaluation)

For our tests on productivity we decided that time data was the most important. There are other ways to conduct a test, but at the end of the day if a translator takes less time to post-edit than if they were to translate from scratch, then you are on the right path. Time data is difficult to track, but thankfully over the last few years Computer-Assisted Translation (CAT) tools have evolved to start measuring this (Moran, Lewis & Saam 2014). For our tests we used iOmegaT, which is an adaptation of OmegaT, to gain access to the instrumentation and telemetry on the translator's activities while they post-edit. There are some privacy concerns with this initially; however we found that all translators were happy to be involved once this is part of a test and they had some control over when the feature could be turned off on their console when not in test.

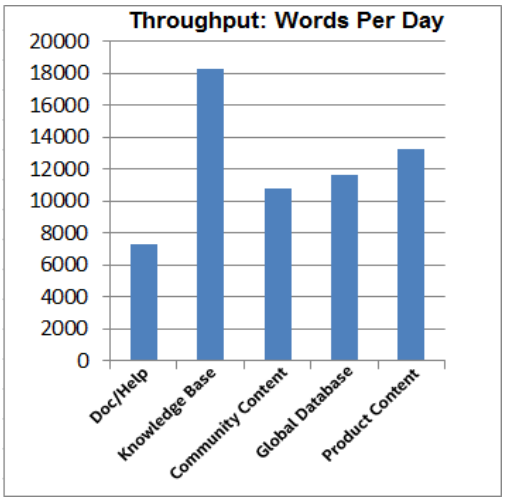


Figure 8: Throughput per type of content showing efficiency compared to others

When we say time data, we basically mean the time it takes to post-edit a segment. It should also include subsequent visits to a segment (not just the first attempt) as it can be a common practice for translators to revisit segments after getting a feel of the document they are translating. Some segments will show erroneous measurements, which is explained when a translator takes a break while having a segment open. We allowed an adequate amount of time for any research the translator may have to do, but we did apply a cutoff to reduce the inclusion of these segments in the test. Timing data can be measured in actual time (ms), but what we used is "words per day" throughput as this is something that most people in our company will understand quickly and easily.

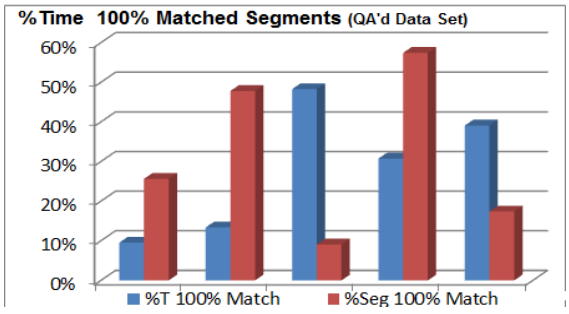


Figure 9: Time spent on 100% match segments from MT

From time data we can learn a lot if we further refine the application of the time data to other metrics such as the brackets for length of segment, the number of repetitions, the number of segments that do not need editing or the number of segments that need a lot of editing (long time segments). We noted that for some content types 100% matches required excessive time to complete when compared to the time spent in other content types. Upon questioning some of these results, the respondents claimed that some 100% match segments require more time to read and understand before they agree that the MT segment is of correct meaning and

language quality. These results helped build up a profile of the content for "post-edit ability" and in turn make some recommendations back to the writing teams.

In our discovery tests we sampled 5 different content categories that all loosely communicated within the same domain. While the writing styles and lexical complexities may diverge in their own subdomain, the core subjects are the same. Nevertheless this process was worthwhile as we learned as much about our own internal content as we did about the ability for MT to work with it. You could say that this practice taught us a lesson about Content Profiling before pushing a content type through an MT workflow.

Moving on to quality, the world of Six Sigma says that "Quality is what your customer wants". So before we enter into a linguistic quality test we should keep that in mind. It's not often that a customer will come to you and tell you what they want, so we use trained domain linguists to conduct a linguistic quality analysis as appropriate and added some segment usability scores and automatic/automated algorithms such as General Text Matching (GTM) or Levenshtein distance to the test matrix.

Language Quality Assurance (LQA) in its traditional form proved to be sufficient in this case for Quality Evaluation (QE), but more advanced error topologies could also help such as TAUS Dynamic Quality Framework. But with so many factors affecting each segment, you must consider these with a soft focus on the overall quality output as some aspects may need to be prioritized or weighted as having an increased effect on the overall output. What that basically means is that LQA parameters need to be aligned with the quality expectation, and this is hard to manage if you are to baseline quality evaluations against something that may be subjective. In the end, you need a number to go by, but you may actually be more interested in the details of the test (accuracy errors, terminology errors, priority or severity of errors etc.) than the overall score achieved by the text.

To balance the linguistic assessment of your MT text with opinions from would be consumers, usability studies can be carried out. We did a number of tests in our company with various native French, Spanish, Chinese and Italian speakers. The test involved going through 100+ segments and scoring them out of 4 (1 for bad and 4 for great). To get a better idea of the diversity in the scoring, we then applied the same scale to LQA and GTM scores (breaking down the percentage brackets into 1 to 4). From the graph below (Figure 3.) we can see the 186 segments in this test more or less correlate to the same sentiment across the 3 types of measurements applied. The results in this case show that the MT for this language is mostly good and there are a minimal amount of truly bad quality segments.

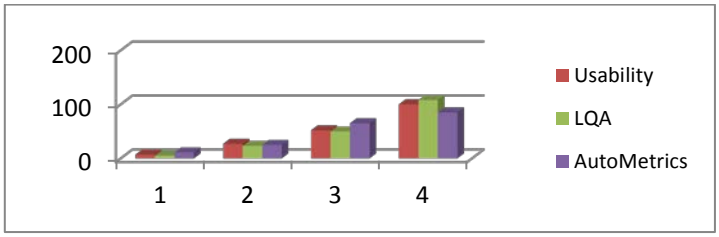


Figure 10: Correlation between 3 quality evaluations (Content Type 1: Documentation)

There is one more thing that we learned while testing our MT and content types. Some content is more prone to error than others when using a static set of Statistical Machine Translation training data. This seems like a reflection on the domain appropriateness, the quality of the MT training and the content writing governance. So we looked at the results and applied

"error probabilities" as something to track for the future. The error probability is almost like a predictive metric, but it is tracked at the end of the process so you can learn lessons for the next time. There seems to be a correlation between error probability and "posteditability", which is the intangible measurement of how difficult or easy it will be for a translator to post-edit a segment and ultimately achieve higher productivity while not sacrificing quality. This was seen when a content type and the time it took to post-edit that content type were taken into account while looking at the number of errors. Ultimately this is like a basic type data that one could use when content profiling.

Type	Segments	Error Segments	Error probability ratio
Content 1	152	102	0.671052632
Content 2	23	12	0.52173913
Content 3	11	10	0.909090909
Content 4	94	31	0.329787234
Content 5	23	19	0.826086957

Figure 11: Error Probability rates per content type

3.4. The workflow, the whole workflow, and nothing but the workflow

In localization we may be guilty of looking at the workflow as being the point where we push our content for translation into a TMS. This may be the old way of doing things, but plugging in MT now means we need to know a little bit more about both ends of the workflow. People often talk about "moving upstream", and this means being more involved with the individuals who write content in your company. They are also part of your company's workflow even if you don't control their part. If they can understand the value of making changes to their process, then they will be able to do that better for you. Similarly you need to deliver the multilingual content to your audience, but the audiences are not going to move upstream and tell you what they want, so it's up to the localizers of the world to better understand product promotion, business compliance, marketing and sales.

Moving upstream as a localizer means that you need to get the message across to anyone who writes content, that if they did it in a slightly different way, the rewards to the company could be a great advantage. The basics of content optimization from authors are Terminology, Style, Reuse and Governance. If many authors in the same company can write in a similar way and reuse the same terminology and phrasing, MT will work well for their content in addition to other benefits.

Working with translators can be very useful and they are not that far downstream as they work with the localization teams every day. Their importance comes from relying on them for quality translations but also because they are possibly the first people to read your documentation outside the company. So, if the translator sees something wrong with the content and you have an ability to track their comments, you have a system to create continuous improvement loops on both content accuracy and style of writing. Further downstream again you have the deployed content, and by tracking who uses your product and what they are clicking on, you can further make strategic decisions on the usage of MT and Post-Editing.

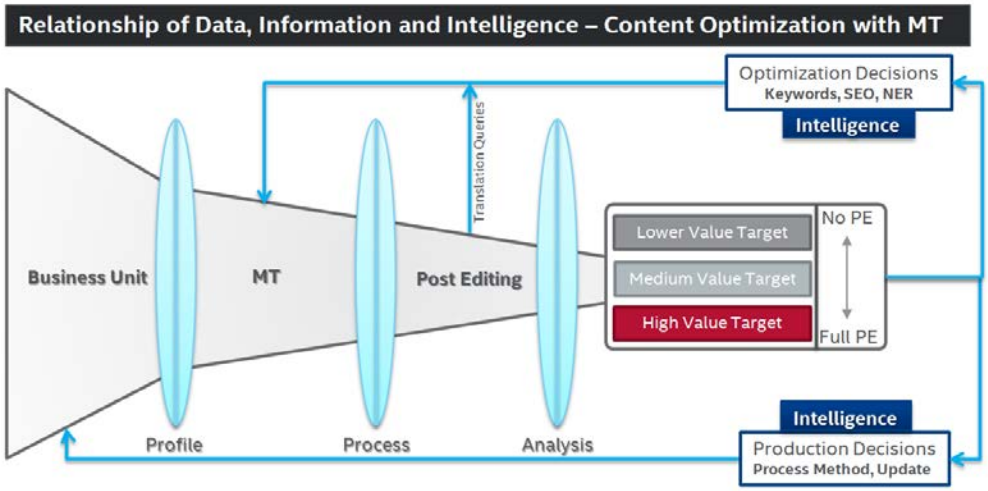


Figure 12: Relationship between upstream and downstream to help MT

So the modern translation workflow, when you include MT, can be improved by working on your source. And the output and consumption can give you great insight into what is working and not working, and needs to be brought back in so you can create a continuous improvement loop.

4. Conclusions

We used methods in this work that draw from 2 basic principles that Controlled Language will help MT output and PEMT can increase throughput in the localization workflow. Our results show that there were significant advantages in using Post-Editing in this case and due to our situation where we could influence the writing standards with our Technical Publications teams these were good starting points for us. Furthermore we understood that some source corpus would be more prone to error when compared to others. While we haven't fully investigated why these differences are yet, we at least have confidence in the texts that do MT well and a basic understanding of the importance of Content Profiling.

The technology is evolving in this area in both the back end of the MT systems and also the new front end Post-Editing Environments being made available. The access to systems such as iOmegaT gave us confidence in the measurements and results when compared to CAT tool agnostic systems such as TAUS DQF for MT QE alone and this real data allowed us to digest a lot of the sales jargon from various suppliers of both MT and Post-Editing services. Having said that TAUS DQF and systems alike do have a place in the process for more sophisticated error topology.

From a higher level in a large corporation, MT can only grow with help from others. We learned that one must spend a lot of time working with the problems of the internal customers while offering the MT solution. ROI must be taken into account a lot at the start so particular focus must be spent on Productivity and Quality Evaluation methods. If the MT project doesn't save money, it's hard to make it grow.

Working with more than one MT supplier can help broaden knowledge quickly: the free or cheap or trial services are ideal to gain insight, learn and build knowledge.

Developers may be needed, either internally or as part of an outsourced partner as there are very few out-of-the-box-solutions and none that will fit all scenarios. But be wary of customizations to TMS and other systems as they can have a costly life length during upgrades.

And finally, once PEMT starts to work, we have the opportunity to look for other ways to use MT in the company. PEMT is a perfect launch pad for the MT program in your company as it is a way to save money and show ROI.

5. Acknowledgements

I would like to thank the translation providers who were part of this project, the Program Managers in the company who helped make these tests possible with their content, the Technical Publications group and the Engineering team. I would like to also thank the various people who helped with multilingual usability studies across the Intel Security site in Cork. Finally I would like to thank Dublin City University and the Center for Next Generation Localization for their talks, expertise and advice.

Glossary

- Machine Translation (MT)
- Statistical Machine Translation (SMT)
- Rules Based Machine Translation (RBMT)
- Post-Editing Machine Translation (PEMT)
- Human Translation (HT)
- Controlled Language (CL)
- Return on Investment (ROI)
- Gist - The substance or general meaning of a speech or text
- Bilingual Evaluation Understudy (BLEU)
- Translation Management Systems (TMS)
- Translation Memories (TMs)
- Computer-Assisted Translation (CAT)
- General Text Matching (GTM)
- Language Quality Assurance (LQA)
- Quality Evaluation (QE)

References

Doherty, S. (2012). Investigating the Effects of Controlled Language on the Reading and Comprehension of Machine Translated Texts: A Mixed-Methods Approach

Roturier, J. (2006). An investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness, and Acceptability of Machine Translated Technical Documentation for French and German users.

Burgett, W., Chang, J., Martin, R. and Yamakawa, Y. (2012), Enabling Multilingual Collaboration through Machine Translation. URL http://www.intel.ie/content/dam/www/public/us/en/documents/white-papers/enabling-multilingual-collaboration-through-machine-translation.pdf

Plitt, M. and Masselot, F. (2010). A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context.

Schmidtke, D. (2005). Microsoft office localization: use of language and translation technology. URL http://www.tm-europe.org/files/resources/TM-Europe2008-Dag-Schmidtke-Microsoft.pdf

O´Brien, S. (2005). Methodologies for measuring the correlations between post-editing effort and machine translatability. Machine Translation,

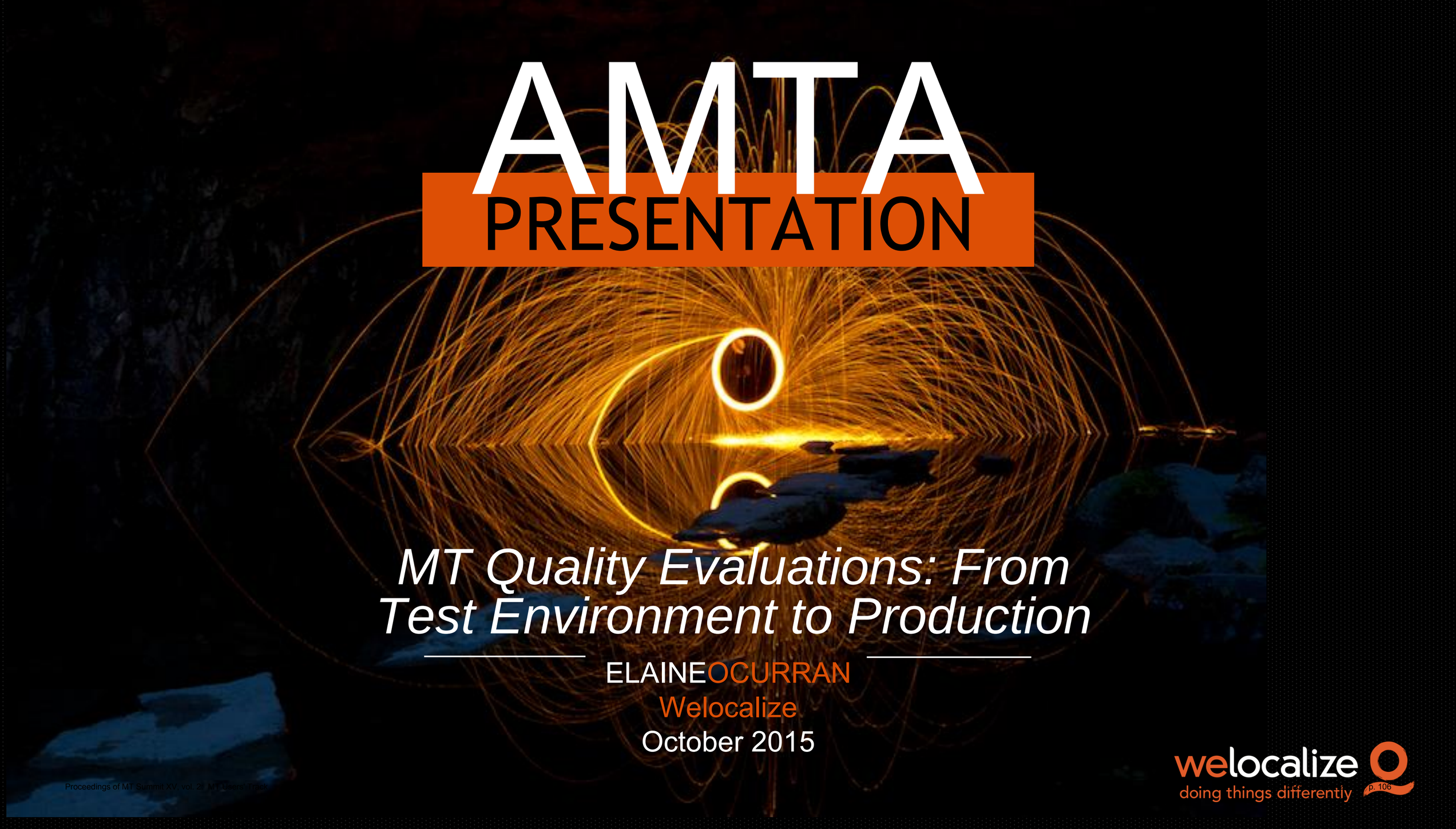
Claesen, J. Quality Metrics for Machine Translation Output. http://www.yamagata-europe.com/en-gb/blog/item/909/quality-metrics-for-machine-translation-outp

Moran, J. Lewis, D and Saam, C. (2014). Towards desktop-based CAT tool instrumentation.

Plitt, M., and Masselot, F. (2010). A Productivity Test of Statistical Machine Translation Post-Editing in a Typical Localisation Context. The Prague Bulletin of Mathematical Linguistics, 93(-1). http://doi.org/10.2478/v10108-010-0010-x

Roturier, J. (2004). Assessing a set of Controlled Language rules: Can they improve the performance of commercial Machine Translation systems. In Proceedings of the international conference translating and the computer (Vol. 26). Retrieved from http://www.mt-archive.info/Aslib-2004-Roturier.pdf

Thiel, K., and Berthold, M. (2012). The KNIME Text Processing Feature. Retrieved from http://www.knime.org/files/knime_text_processing_introduction_technical_report_120515.pdf



AMTA PRESENTATION

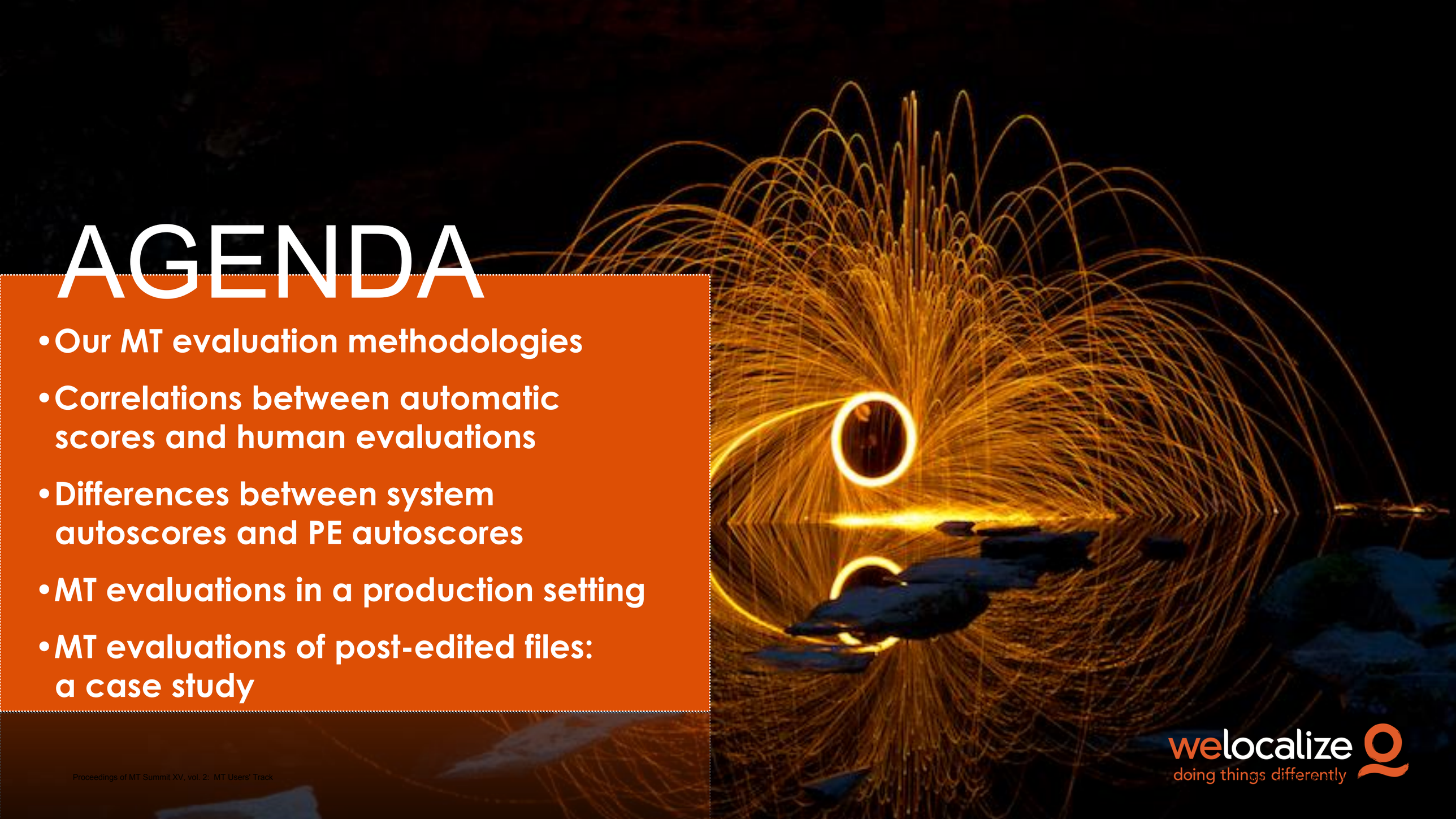
MT Quality Evaluations: From Test Environment to Production

ELAINEOCURRAN
Welocalize

October 2015

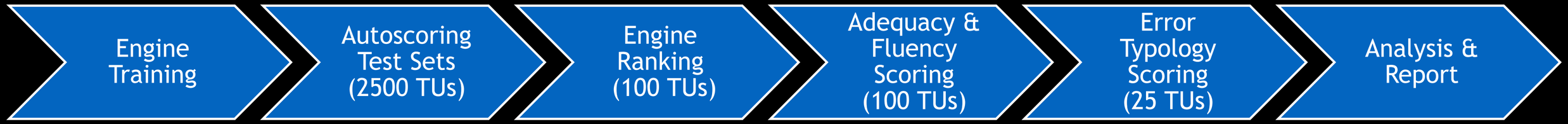
AGENDA

- Our MT evaluation methodologies
- Correlations between automatic scores and human evaluations
- Differences between system autoscores and PE autoscores
- MT evaluations in a production setting
- MT evaluations of post-edited files: a case study



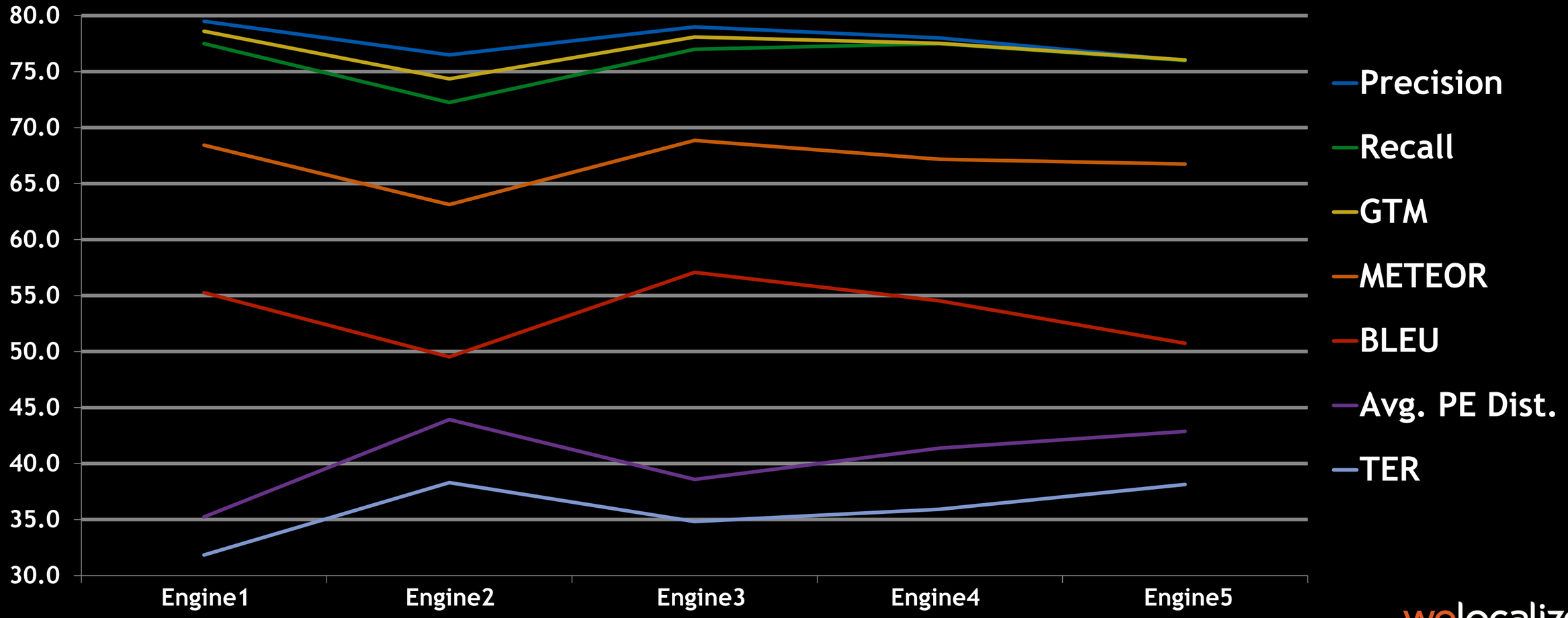
OUR EVALUATION METHODS

A TYPICAL EVALUATION PROCESS PER LOCALE AND PER ENGINE



OUR EVALUATION METHODS

AUTOMATIC SCORES GENERATED BY WESCORE



welocalize
doing things differently

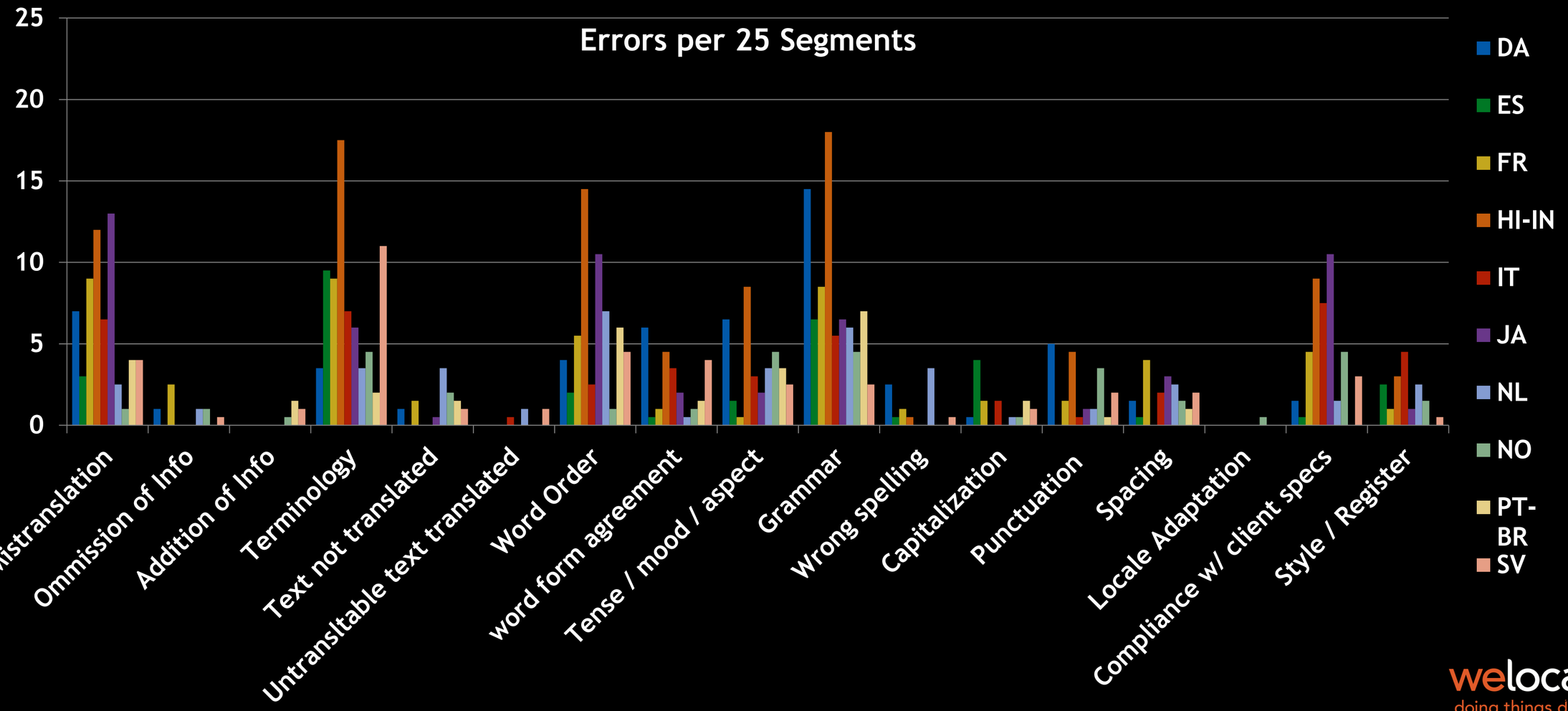
OUR EVALUATION METHODS

HUMAN EVALUATIONS: ADEQUACY AND FLUENCY SCORING

SCORE	ACCURACY	FLUENCY
5	All meaning expressed in the source fragment appears in the translation fragment.	Native language fluency. No grammar errors, good word choice and syntactic structure. No PE required.
4	Most of the source fragment meaning is expressed in the translation fragment.	Near native fluency. Few terminology or grammar errors which don't impact the overall understanding of the meaning. Little PE required.
3	Much of the source fragment meaning is expressed in the translation fragment.	Not very fluent. About half of translation contains errors and requires PE.
2	Little of the source fragment meaning is expressed in the translation fragment.	Little fluency. Wrong word choice, poor grammar and syntactic structure. A lot of PE required.
1	None of the meaning expressed in the source fragment is expressed in the translation fragment.	No fluency. Absolutely ungrammatical and for the most part doesn't make any sense. Translation has to be re-written from scratch.

OUR EVALUATION METHODS

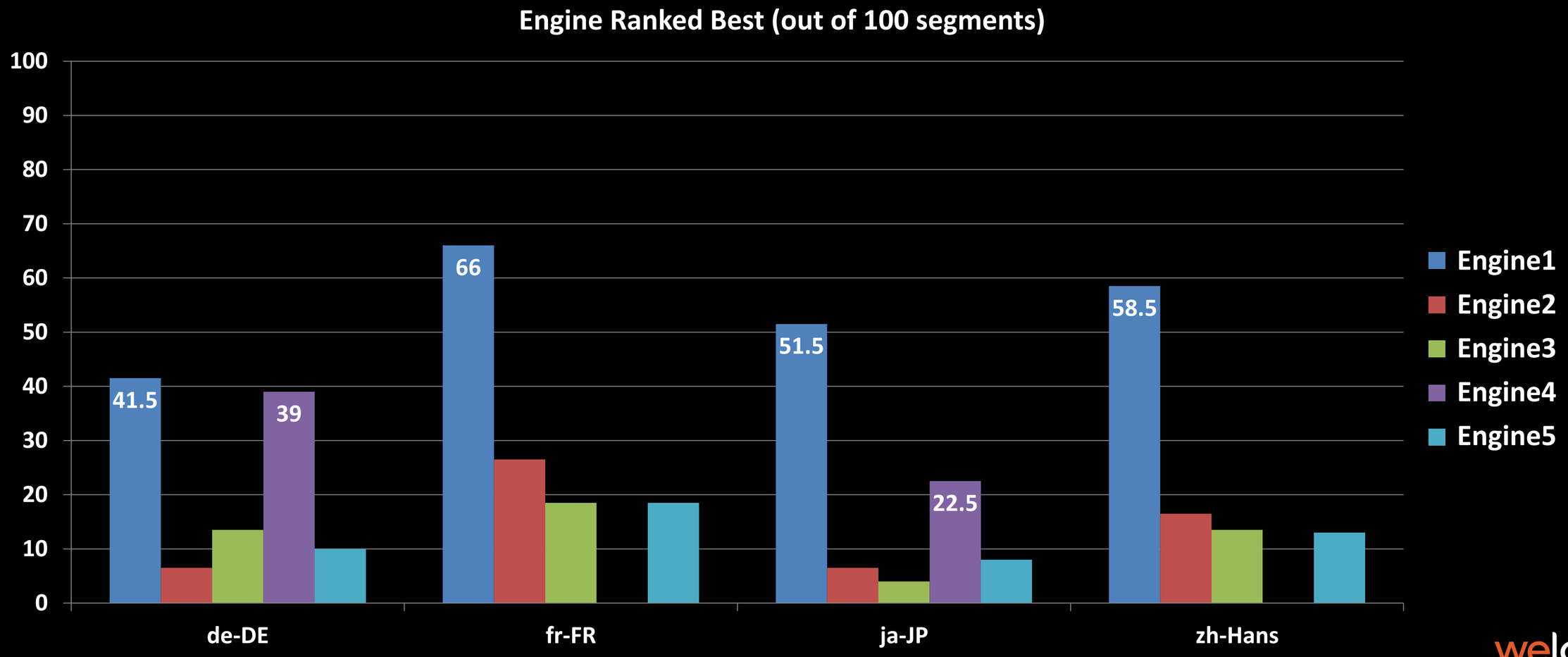
HUMAN EVALUATION: ERROR TYPOLOGY



welocalize
doing things differently

OUR EVALUATION METHODS

HUMAN EVALUATION: ENGINE RANKING



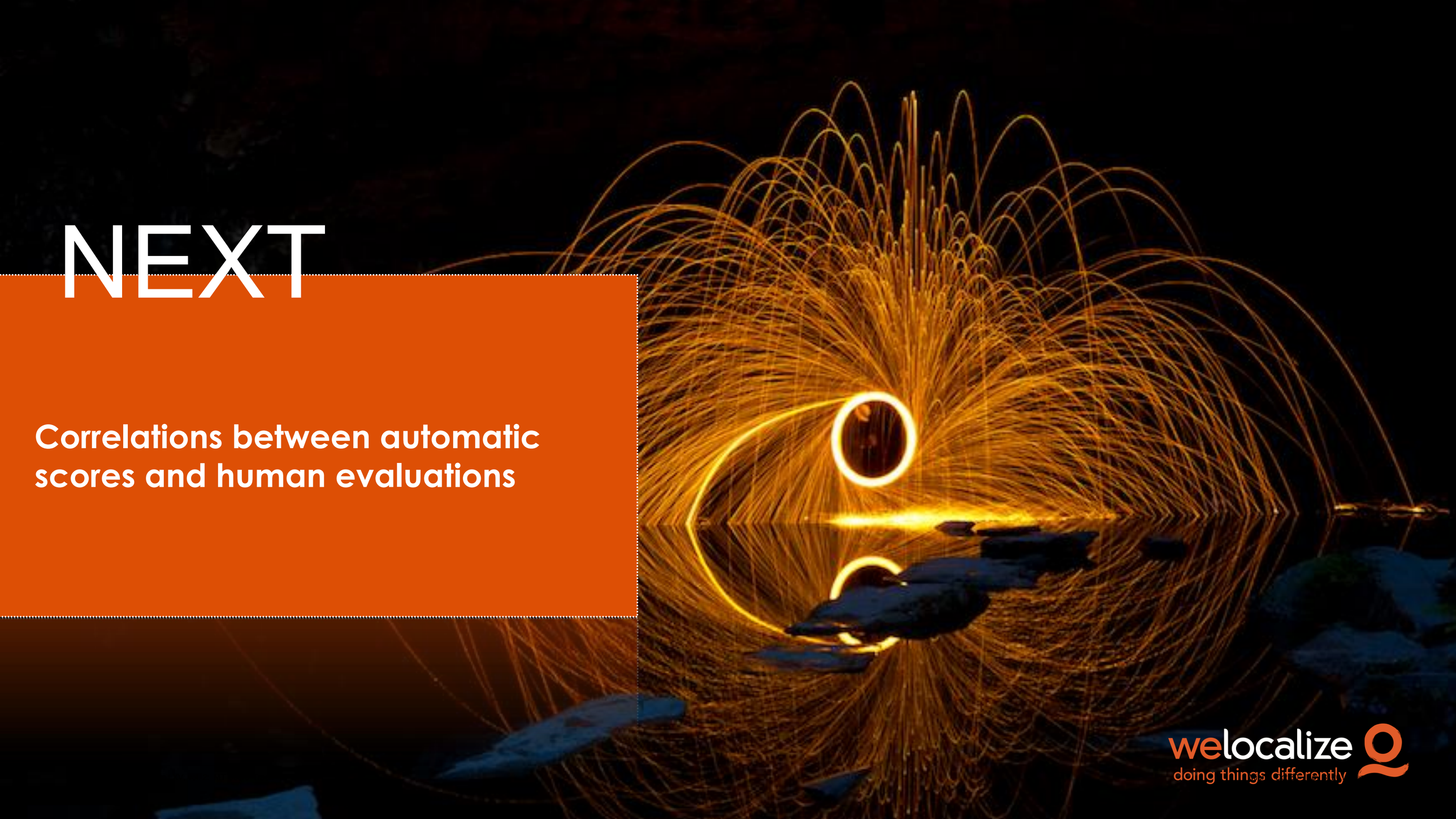
welocalize
doing things differently

LESSONS LEARNED

- We always perform autoscoring PLUS human scoring for all our MT evaluations. We have internal thresholds that qualify an engine ready for deployment and it's level of maturity.
- For bake-offs between several engines, we always include engine ranking in addition to our standard scores.
- Productivity tests are valuable during the initial phase of an MT program to build up productivity data for future reference across languages, domains and MT systems.
- Our MT program is now mature and we are able to perform most of our evaluations based on autoscoring PLUS human scoring, and by referencing the productivity data we have collected over a number of years.

NEXT

Correlations between automatic scores and human evaluations



welocalize
doing things differently

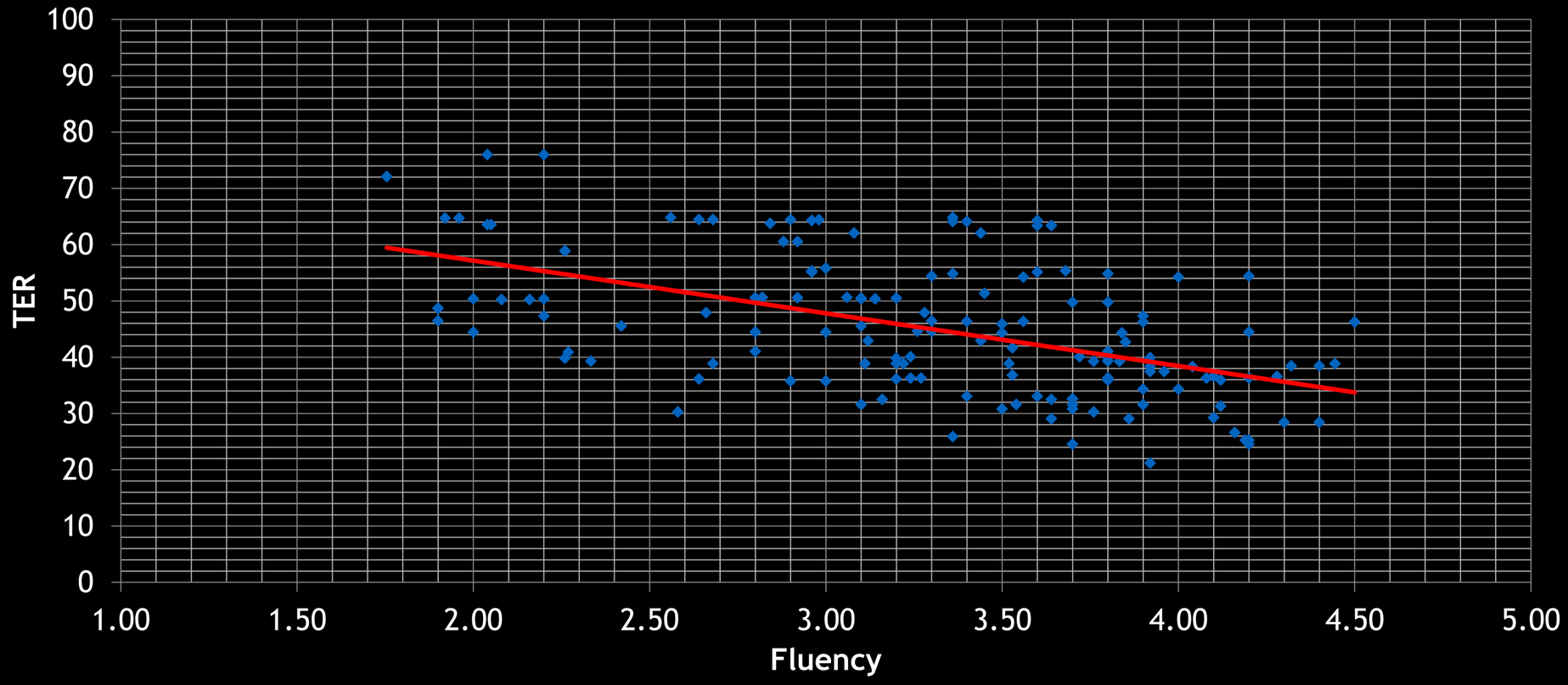
CORRELATIONS

CORRELATIONS BETWEEN AUTOMATIC SCORES AND HUMAN EVALUATIONS

Pearson's r	Variables	Strength of Correlation	Tests (N)	Locales
0.50576955	Fluency & METEOR	Strong positive relationship	150	11
0.50070425	Fluency & BLEU	Strong positive relationship	150	11
0.49816365	Fluency & Recall	Strong positive relationship	150	11
0.49724893	Fluency & NIST	Strong positive relationship	150	11
0.49195687	Fluency & GTM	Strong positive relationship	150	11
0.47064566	Fluency & Precision	Strong negative relationship	150	11
0.38293518	Adequacy & NIST	Moderate negative relationship	150	11
0.31354314	Adequacy & METEOR	Moderate negative relationship	150	11
0.2940756	Adequacy & Recall	Weak positive relationship	150	11
0.28586852	Adequacy & GTM	Weak positive relationship	150	11
0.28386332	Adequacy & BLEU	Weak positive relationship	150	11
0.26685854	Adequacy & Precision	Weak positive relationship	150	11
-0.40270902	Adequacy & TER	Strong negative relationship	150	11
-0.4788575	Fluency & PE Distance	Strong negative relationship	150	11
-0.5385275	Adequacy & PE Distance	Strong negative relationship	150	11
-0.5421933	Fluency & TER	Strong negative relationship	150	11

CORRELATIONS

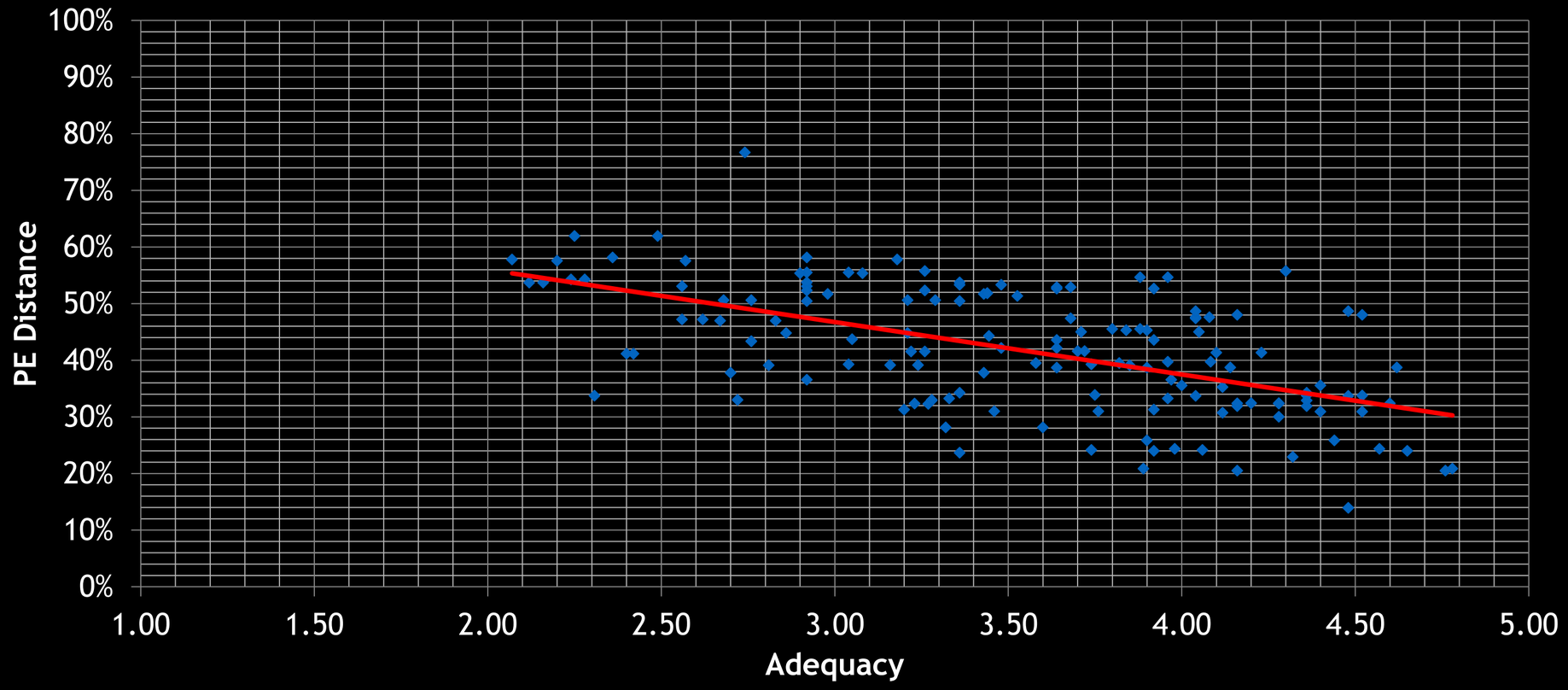
THE STRONGEST CORRELATION WAS FOUND BETWEEN FLUENCY AND TER



welocalize
doing things differently

CORRELATIONS

THE 2ND STRONGEST CORRELATION WAS FOUND BETWEEN ADEQUACY AND PE DISTANCE



LESSONS LEARNED

- It seems that we cannot rely solely on autoscores as long as the correlation with human judgment is not stronger than the data suggests
- TER and PE Distance show the strongest correlation to both Fluency and Adequacy, and therefor seem closer to human judgment than the other scores.
- Fluency correlates stronger with system autoscores than Adequacy overall.
- PE Distance is the only metric that correlates stronger with Adequacy than Fluency. PE Distance is also the only character-based metric.

NEXT

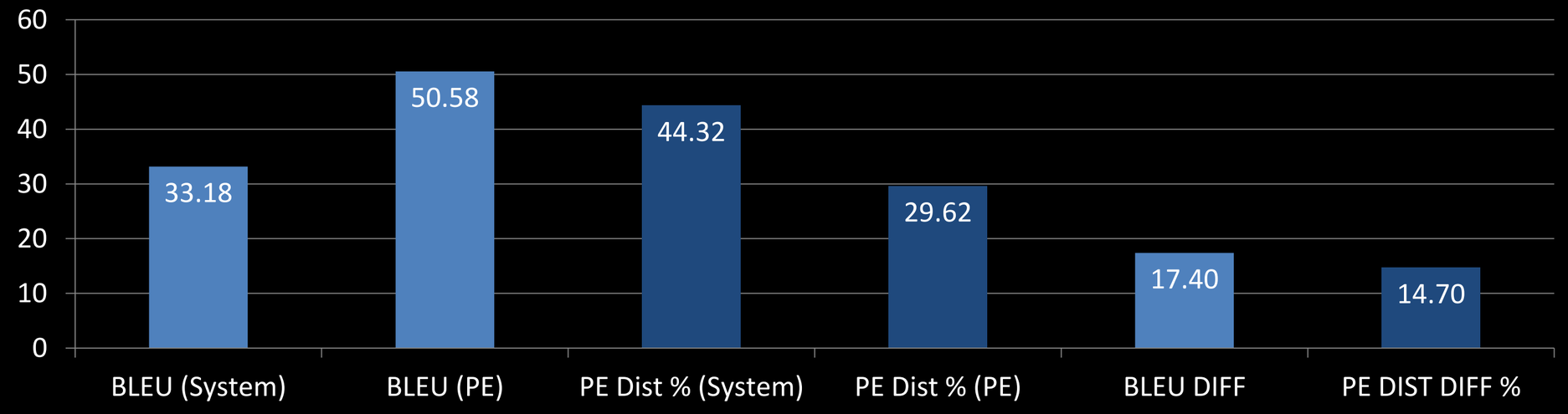
Differences between system autoscores and post-editing autoscores

welocalize
doing things differently

SYSTEM VS PE AUTOSCORES

ON AVERAGE, THE POST-EDITING SCORE IS 15 AND 17 POINT HIGHER FOR PE DISTANCE AND BLEU RESPECTIVELY

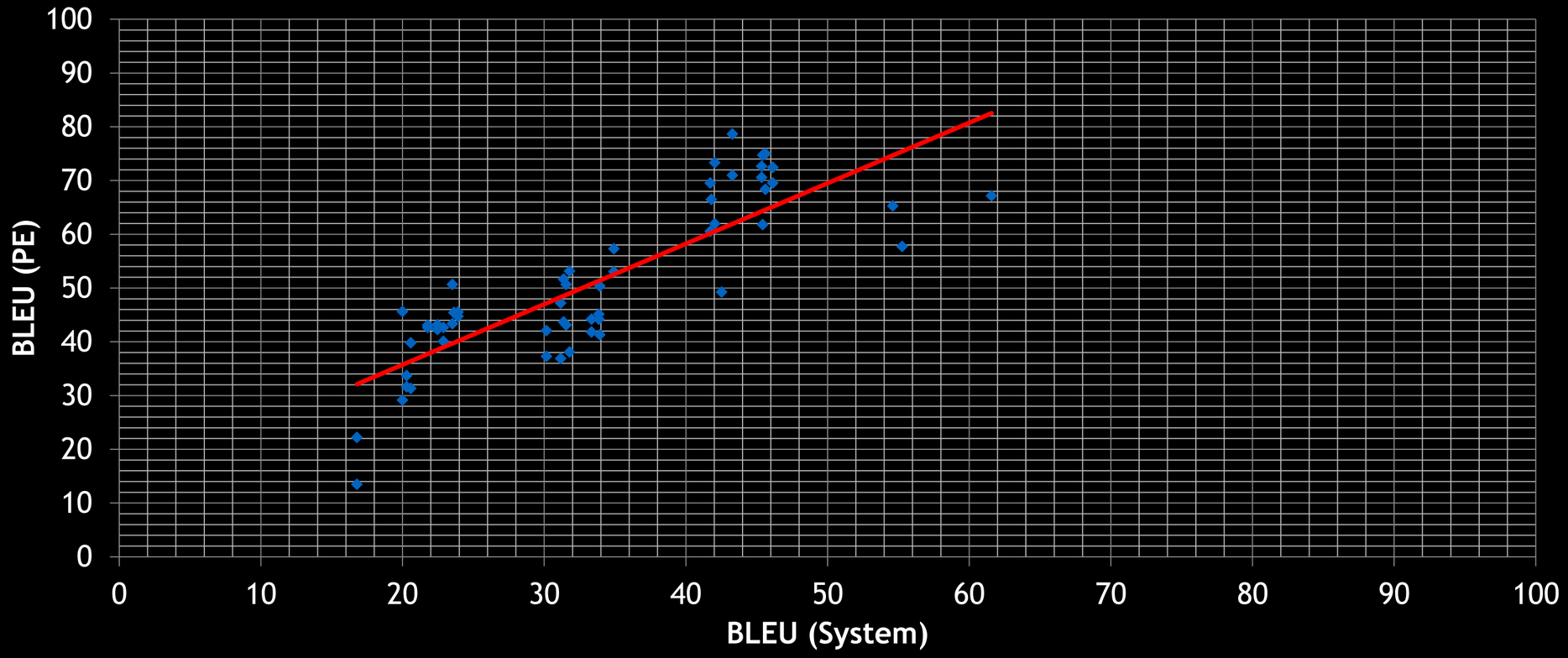
Pearson's r	*Variables*	*Strength of Correlation*	*Tests (N)*	*Locales*
0.832226688	*BLEU (System) & BLEU (PE)*	*Very strong positive relationship*	*57*	*9*
0.832218909	*PE Distance (System) & PE Distance (PE)*	*Very strong positive relationship*	*57*	*9*



welocalize
doing things differently

SYSTEM VS PE AUTOSCORES

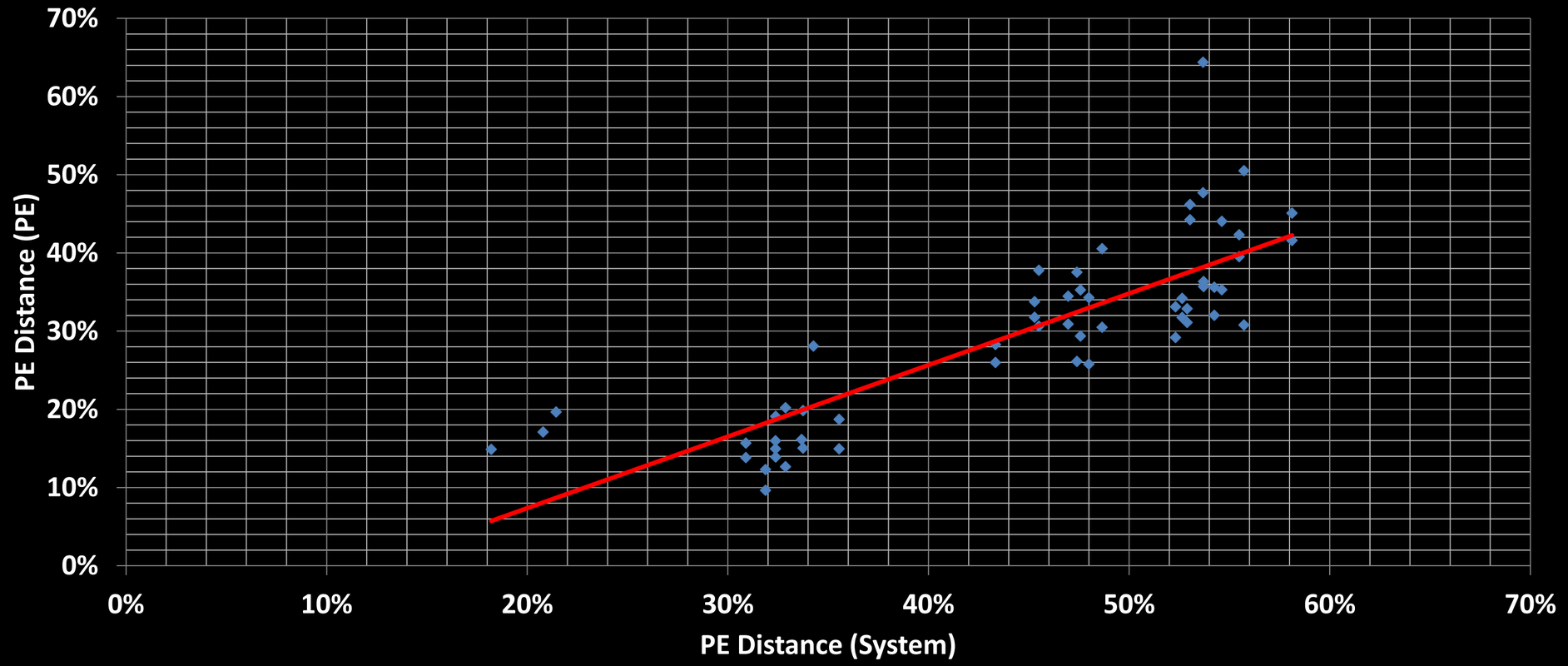
CORRELATIONS BETWEEN SYSTEM BLEU AND POST-EDITING BLEU



welocalize
doing things differently

SYSTEM VS PE AUTOSCORES

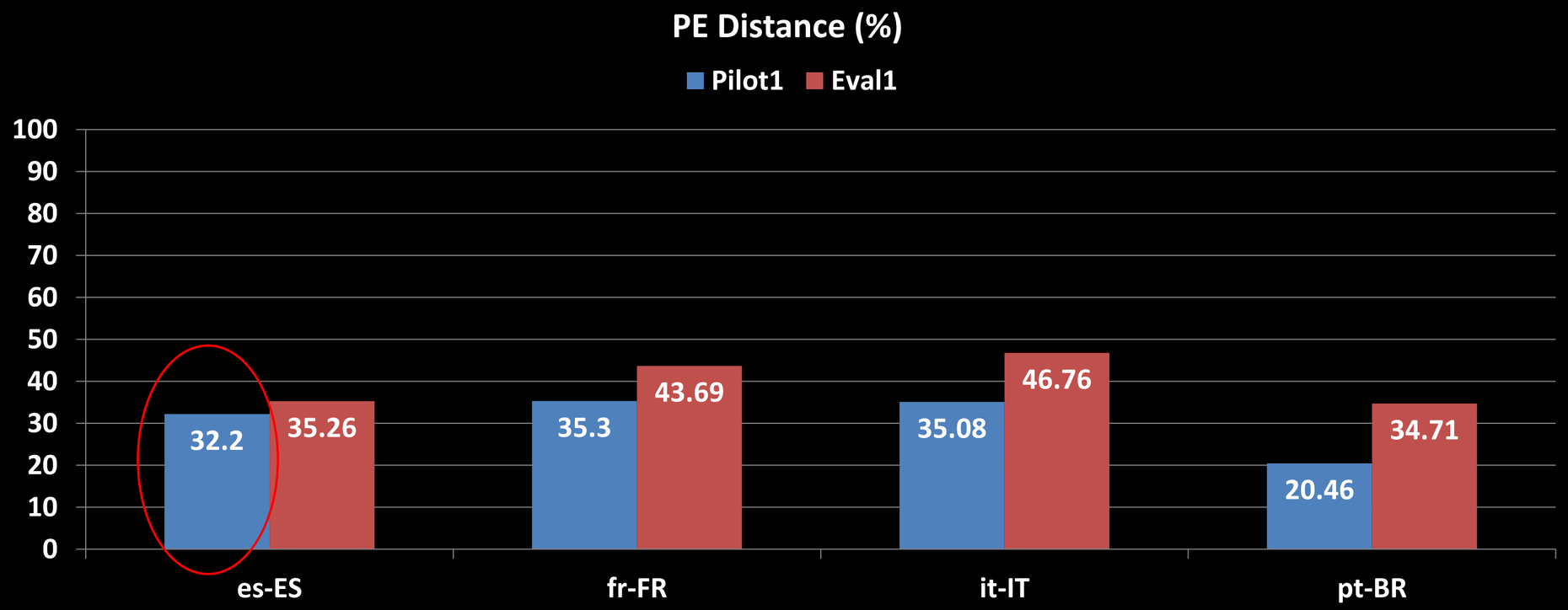
CORRELATIONS BETWEEN SYSTEM PE DISTANCE AND POST-EDITING PE DISTANCE



welocalize
doing things differently

SYSTEM VS PE AUTOSCORES

REAL DATA WHERE WE COMPARE EVALUATION SCORES WITH SCORES FROM A 3-MONTH PILOT



LOOK FOR CONSISTENCY AND BEWARE OF OUTLIERS

welocalize
doing things differently

LESSONS LEARNED

- There is a very high correlation between the MT system autoscores generated during the evaluation phase and the autoscores generated from production using the same engines.
- However, the post-editing autoscores are considerably better than the MT system autoscores by around15%.
- We now differentiate the autoscores in our database as ‘System’ and ‘PE’.


NEXT
MT evaluations in a production setting
welocalize
doing things differently

PRODUCTION SETTING

HOW TO MEASURE POST-EDITING EFFORT

- It is important to monitor the performance of MT and post-editors, especially during the initial launch of a new program
- The use of autoscoring to analyze post-project files is a valuable and cost-effective method to measure the post-editing effort
- They support rate negotiations and can help us to identify over- or under-editing by post-editors
- TER and PE Distance are useful metrics, with different underlying algorithms

PRODUCTION SETTING

HOW TO MEASURE POST-EDITING EFFORT

PE Distance - lower is better!

- Measures the number of insertions, deletions, substitutions required to transform MT output to the required quality level
- PE Distance values are derived by comparing the post-edited segments with the corresponding machine translation segments
- In our analysis the PE distance applies the Levenshtein algorithm and is **character-based**. This captures morphological post-edits, such as fixing word forms.

PRODUCTION SETTING

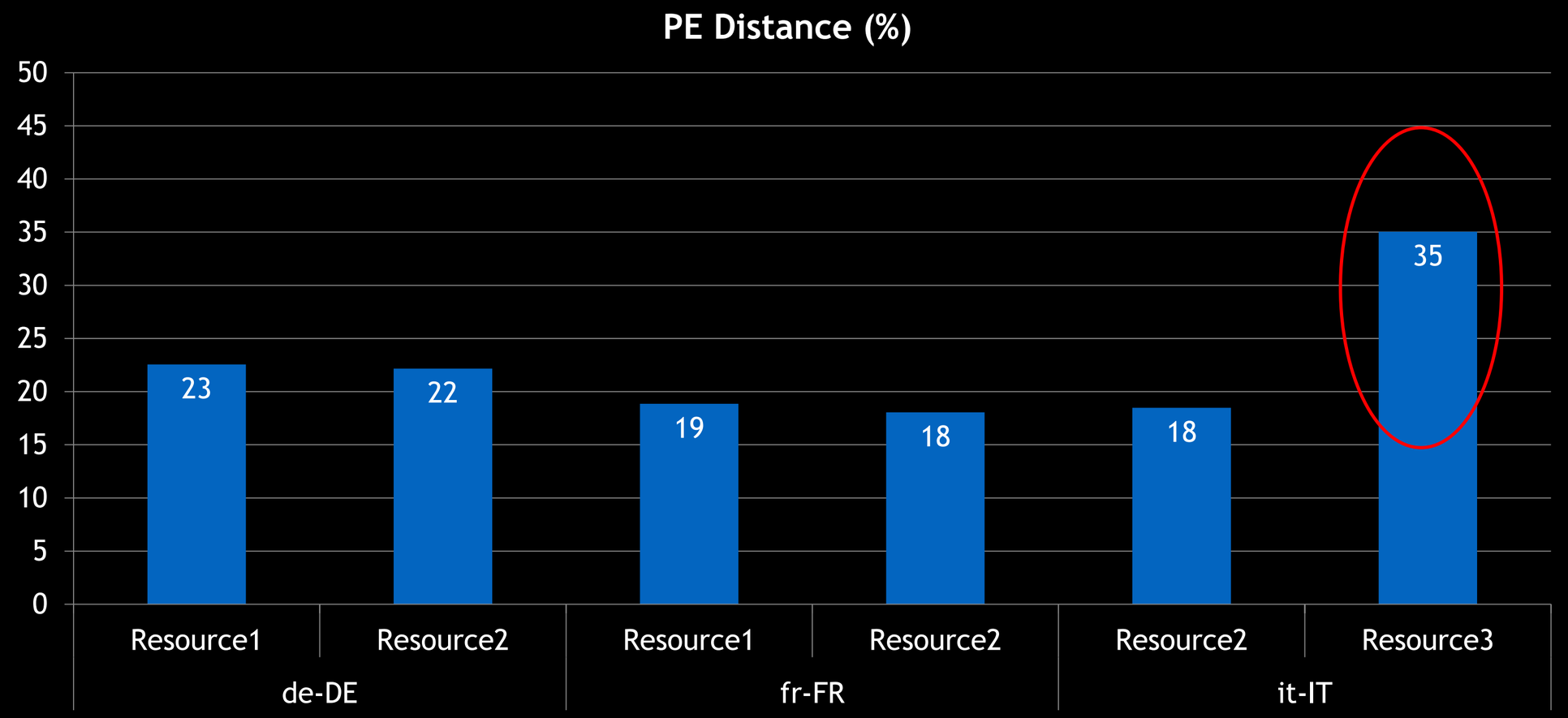
HOW TO MEASURE POST-EDITING EFFORT

TER - lower is better!

- TER stands for Translation Edit Rate
- It is an error metric for machine translation that measures the number of edits required to change a system output into the post-edited version
- Possible edits include the insertion, deletion, and substitution of single words as well as shifts of word sequences.
- Unlike PE Distance, TER is a **word-based** error metric and therefor does not capture morphological changes during post-editing.

PRODUCTION SETTING

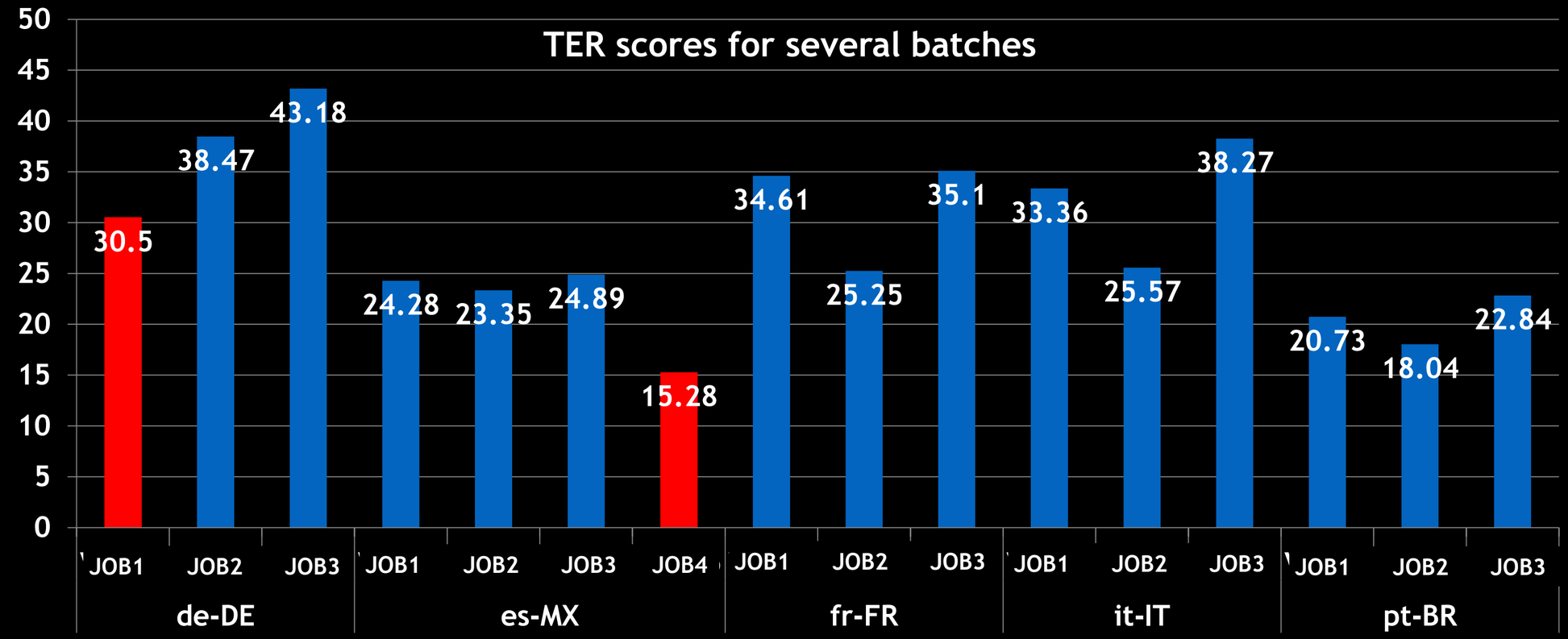
LOOK FOR CONSISTENCY AND BEWARE OF OUTLIERS



welocalize
doing things differently

PRODUCTION SETTING

LOOK FOR CONSISTENCY AND BEWARE OF OUTLIERS: POST-PROJECT AUTOSCORES INDICATE UNDEREDITING



welocalize
doing things differently

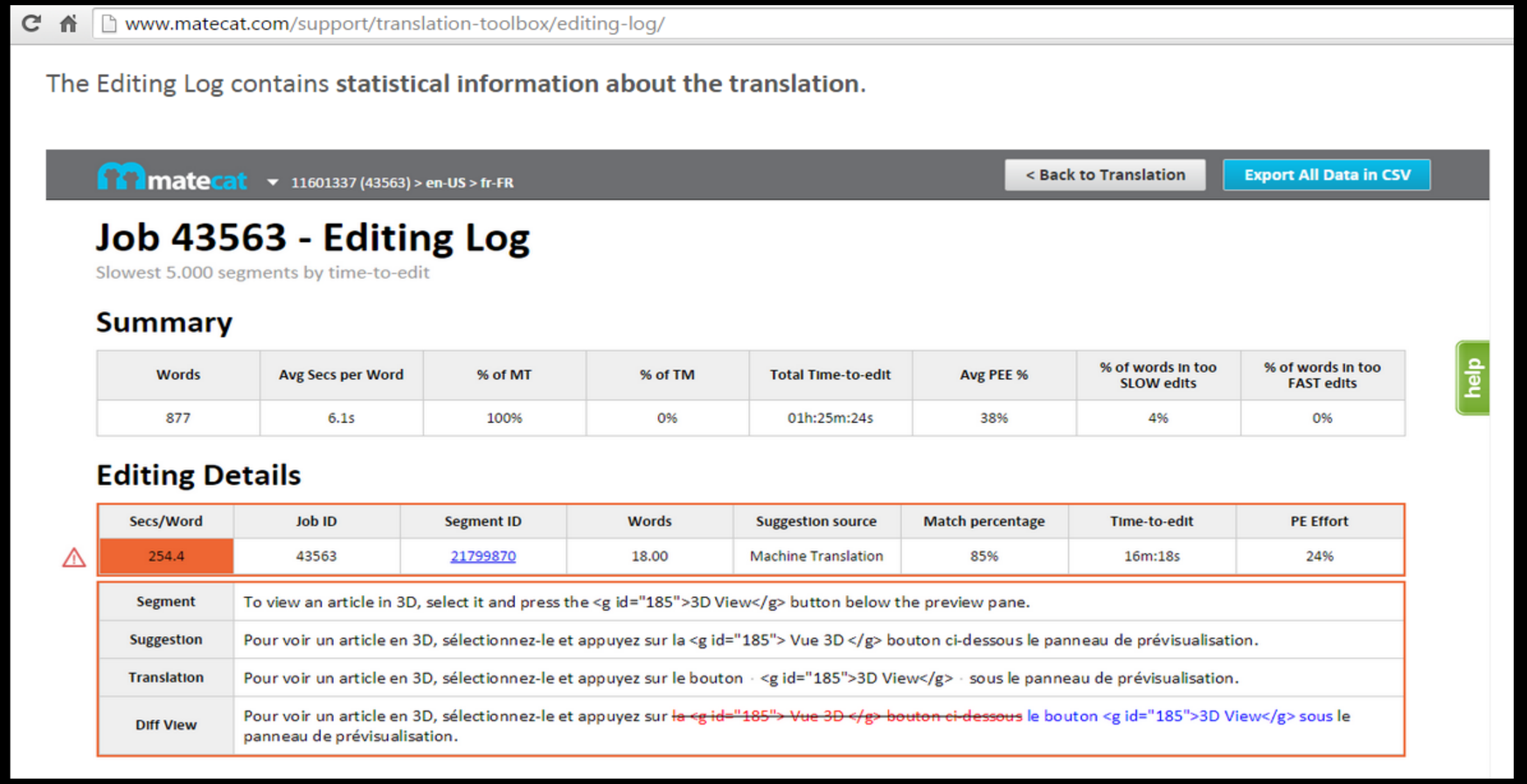
PRODUCTION SETTING

TOOLS TO MEASURE POST-EDITING EFFORT

TOOL	INPUT FILES	OUTPUT REPORT	PROS	CONS
iOmegaT	xliff & more	xml	Includes productivity data	Generated in the CAT tool during translation, requires post-editor buy-in
MateCat	xliff	Excel	Includes productivity data as a built in feature	Generated in the CAT tool during translation, requires post-editor buy-in
Okapi	xliff	html	Allows us to measure PE distance post-project	Requires access to pre- and post-edited file sets
Post-Edit Compare	sdlxliff	html	Allows us to measure PE distance post-project	Requires access to pre- and post-edited file sets
Qualitivity	sdlxliff	Excel	Includes productivity data	Generated in the CAT tool during translation, requires post-editor buy-in
wescore	tmx	Excel	Allows us to measure PE distance post-project	Proprietary tool, Requires access to pre- and post-edited file sets

PRODUCTION SETTING

MATECAT IS A FREE ONLINE CAT TOOL WITH EDITING LOG



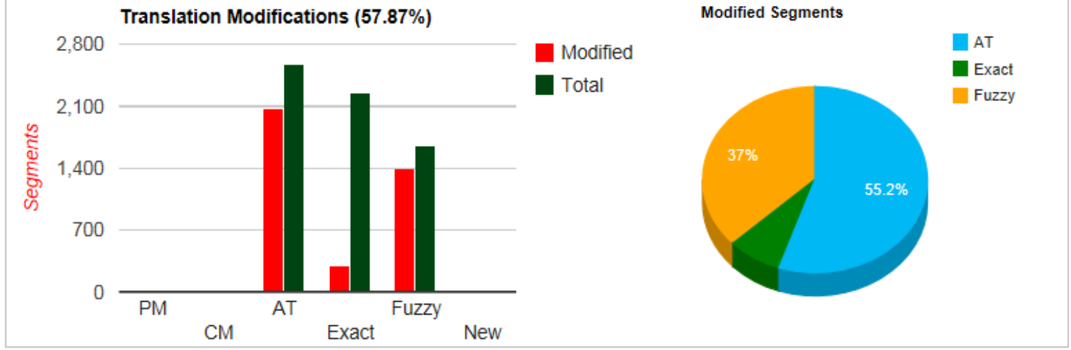
http://www.matecat.com/support/translation-toolbox/editing-log/

welocalize
doing things differently

PRODUCTION SETTING

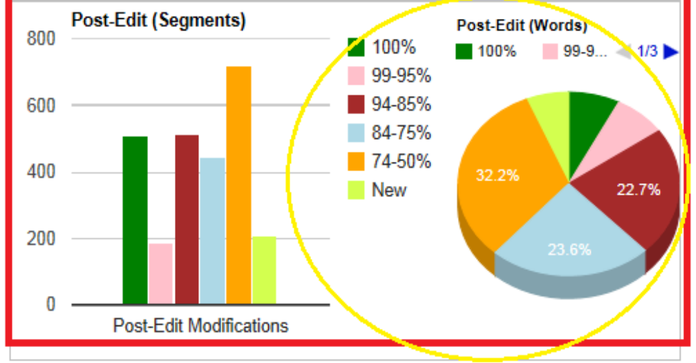
USE POST-EDIT COMPARE TO ANALYSE SDLXLIFF FILES

Translation Modifications	Segments			Words		Characters		Tags/Placeables	
	Total	*Modified*	*%*	*Added*	*Removed*	*Added*	*Removed*	*Added*	*Removed*
Perfect Match	0	0	*0%*	0	0	0	0	0	0
Context Match	0	0	*0%*	0	0	0	0	0	0
Exact Match	2251	294	*13%*	339	322	1963	2282	47	47
Automated Translation	2582	2074	*80%*	9128	9057	53240	50679	1337	1337
Fuzzy Match	1661	1391	*84%*	3049	2709	18279	15052	671	30
New	1	0	*0%*	0	0	0	0	0	0
Sub-Total				**12516**	**12088**	**73482**	**68013**	**2055**	**1414**
Total	**6496**	**3759**	*58%*	**24604**		**141495**		**3469**	



Post-Edit Modifications Analysis

Type	Segments	Words	Characters	Percent	Tags	Total
100%	509	2296	22581	7.50%	0	0 ()
95% - 99%	188	2361	12352	7.71%	54	0 ()
85% - 94%	514	6953	36004	22.71%	276	0 ()
75% - 84%	443	7217	38309	23.57%	428	0 ()
50% - 74%	721	9868	52569	32.23%	1286	0 ()
New	208	1926	11008	6.29%	582	0 ()
-						
Total	**2583**	**30621**	**172823**	**100%**	**2626**	**0 ()**



http://www.translationzone.com/openexchange/app/post-editcompare-495.html

welocalize
doing things differently

PRODUCTION SETTING

OKAPI FRAMEWORK TRANSLATION COMPARISON STEP

Summary

Repartition for Trans1 to Trans2:

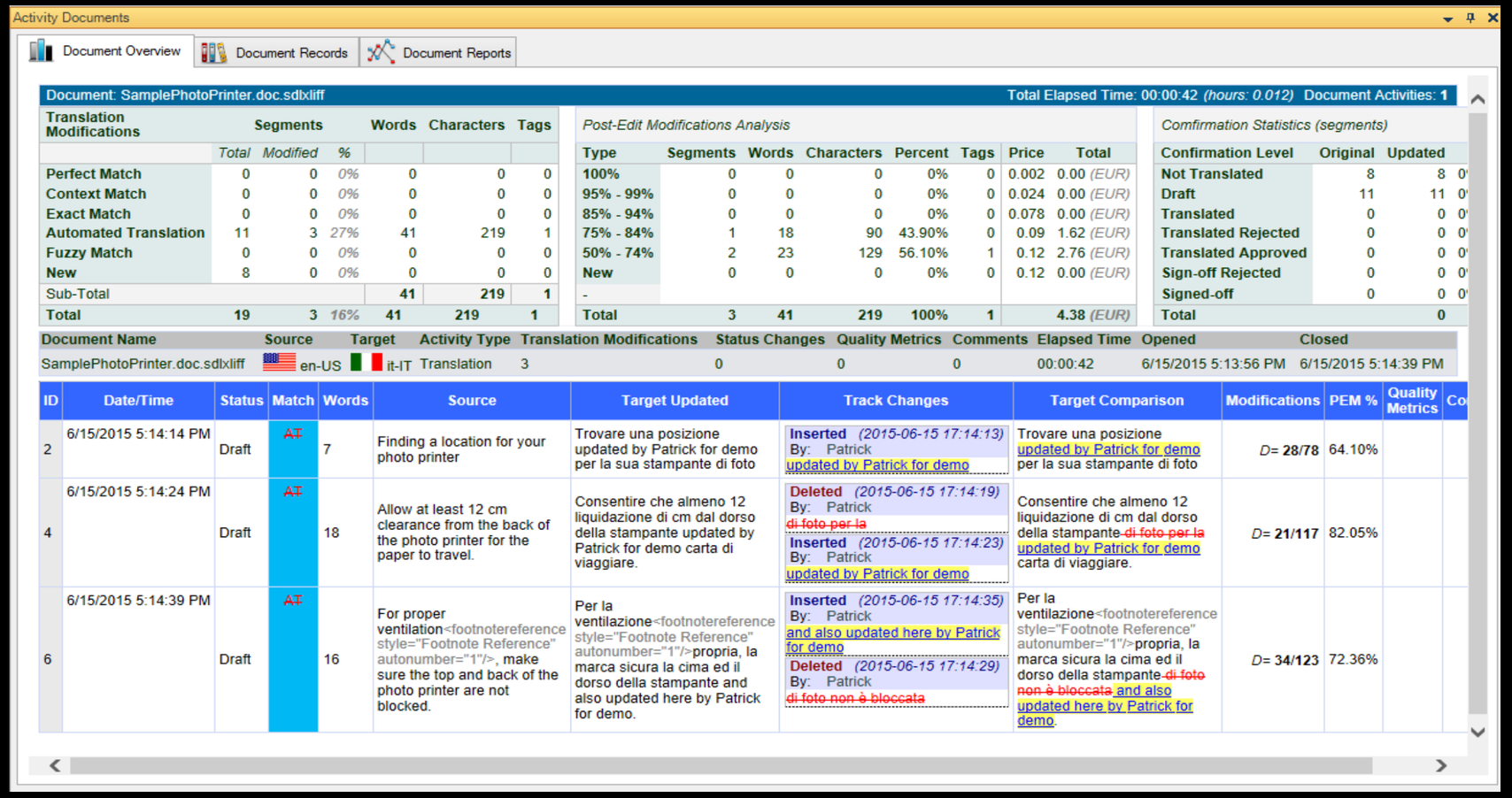
Scores	ED-Scores				FM-Scores			
	Segments	%	Words	%	Segments	%	Words	%
100	139	3	1414	3	176	4	1802	4
90 - 99	350	8	3954	8	346	8	3864	8
80 - 89	862	20	9850	20	674	16	7659	15
70 - 79	971	22	11137	23	804	19	9191	19
60 - 69	1078	25	12423	25	805	19	9332	19
50 - 69	598	14	6794	14	655	15	7500	15
40 - 59	197	5	2215	4	392	9	4479	9
30 - 39	33	1	359	1	240	6	2775	6
20 - 29	2	0	22	0	102	2	1159	2
10 - 19	1	0	4	0	36	1	398	1
0 - 9	104	2	1258	3	105	2	1271	3
Total	4335	100%	49430	100%	4335	100%	49430	100%

```
Total Number of Segments:       4335
Total Number of Words:          49430
Average word count per segment: 11.40
Average ED-Score (by segment):  Trans1 to Trans2 = 69.95
Average FM-Score (by segment):  Trans1 to Trans2 = 65.48
Average ED-Score (by word):     Trans1 to Trans2 = 69.76
Average FM-Score (by word):     Trans1 to Trans2 = 65.18
Edit Effort Score:              32.53
```

http://www.opentag.com/okapi/wiki/index.php?title=Translation_Comparison_Step

PRODUCTION SETTING

QUALITIVITY PLUGIN FOR SDL TRADOS STUDIO



http://www.translationzone.com/openexchange/app/qualitivity-788.html

welocalize
doing things differently

LESSONS LEARNED

- The use of autoscoring to analyze post-project files is a valuable and cost-effective method to measure the post-editing effort.
- A productivity test requires upfront organization and buy-in from translators.
- It is important to find a tool that works with the given file format and workflow.
- Access to pre- and post-edit versions of projects is required. This is a challenge on some accounts.
- Identification and separation of MT segments from fuzzy segments may be required for some tools.
- Look for consistency across languages and resources. Unusually high or low scores can be a sign of over-editing or under-editing.


NEXT
MT evaluations of post-edited files: a case study
welocalize
doing things differently

CASE STUDY

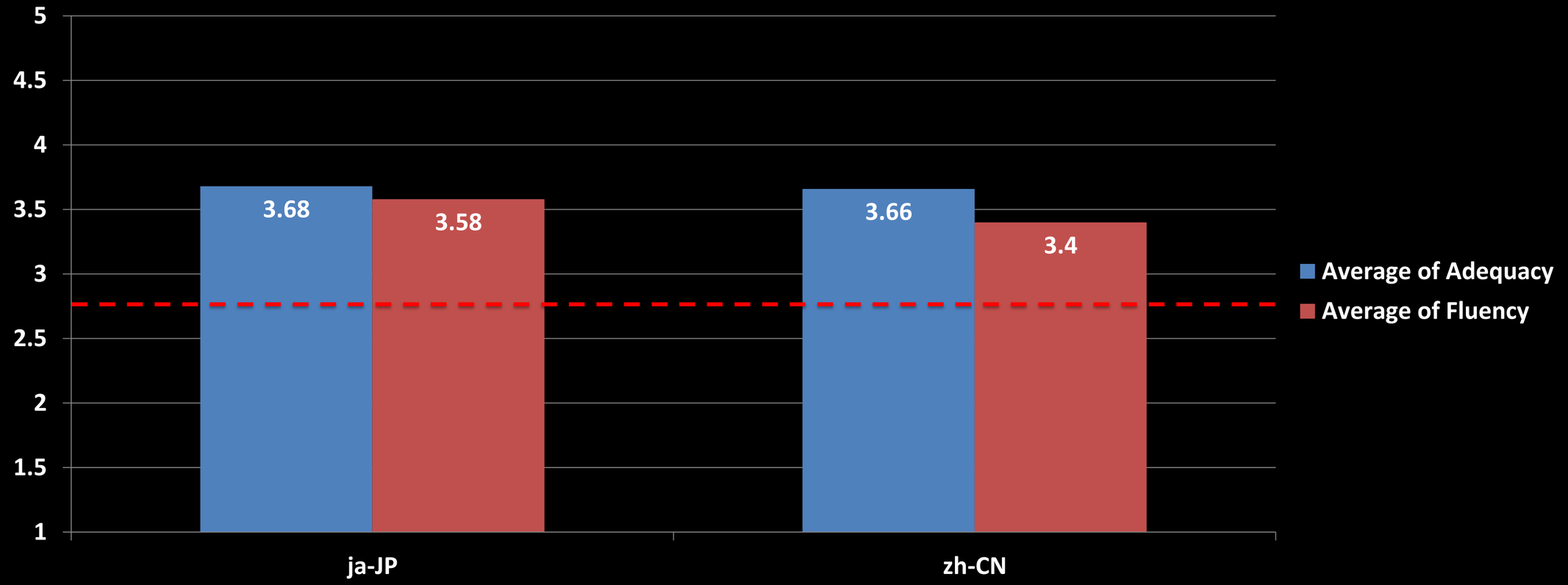
TEST PILOT FOR LIGHT AND FULL POST-EDITING

- Languages: Chinese (Simplified) and Japanese
- The resources are regular translators for this client
- In order to have comparable data, the same resource performed both light and full post-editing tasks of 438 segments

welocalize
doing things differently

CASE STUDY: HUMAN EVALS

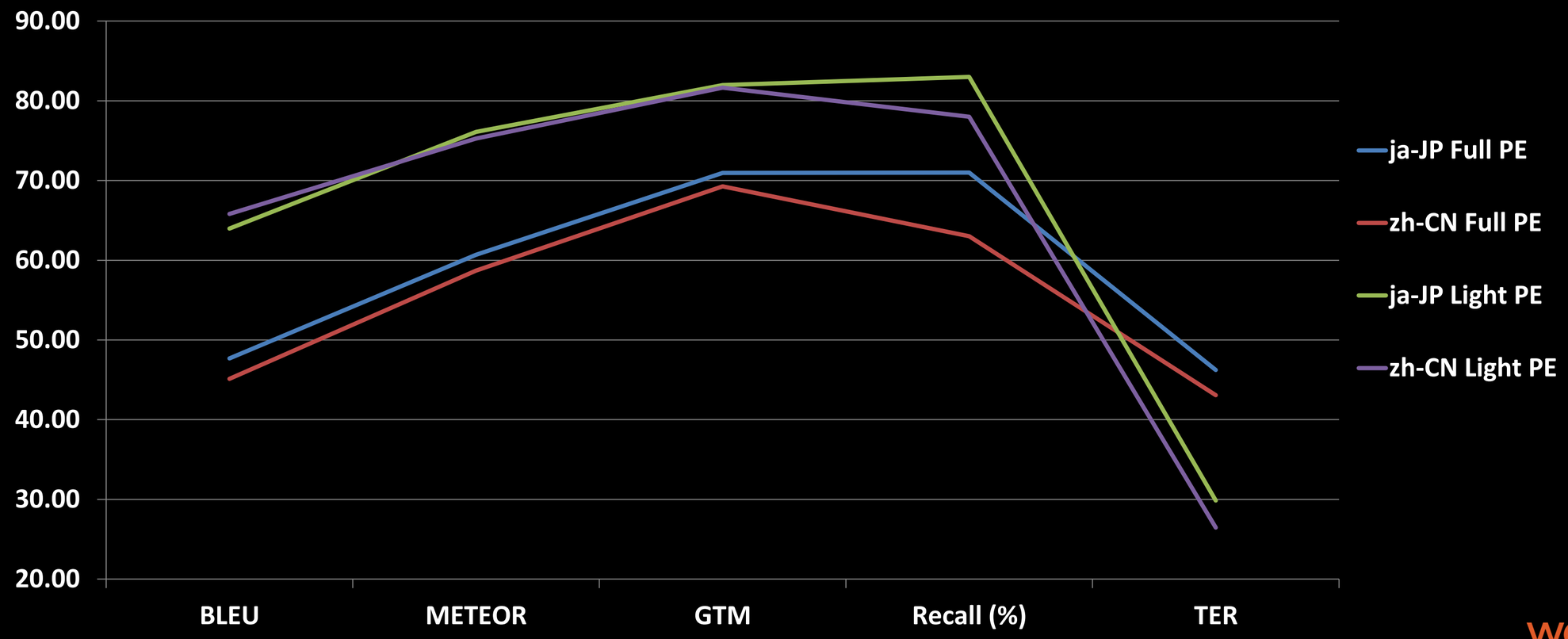
ADEQUACY AND FLUENCY SCORES



welocalize
doing things differently

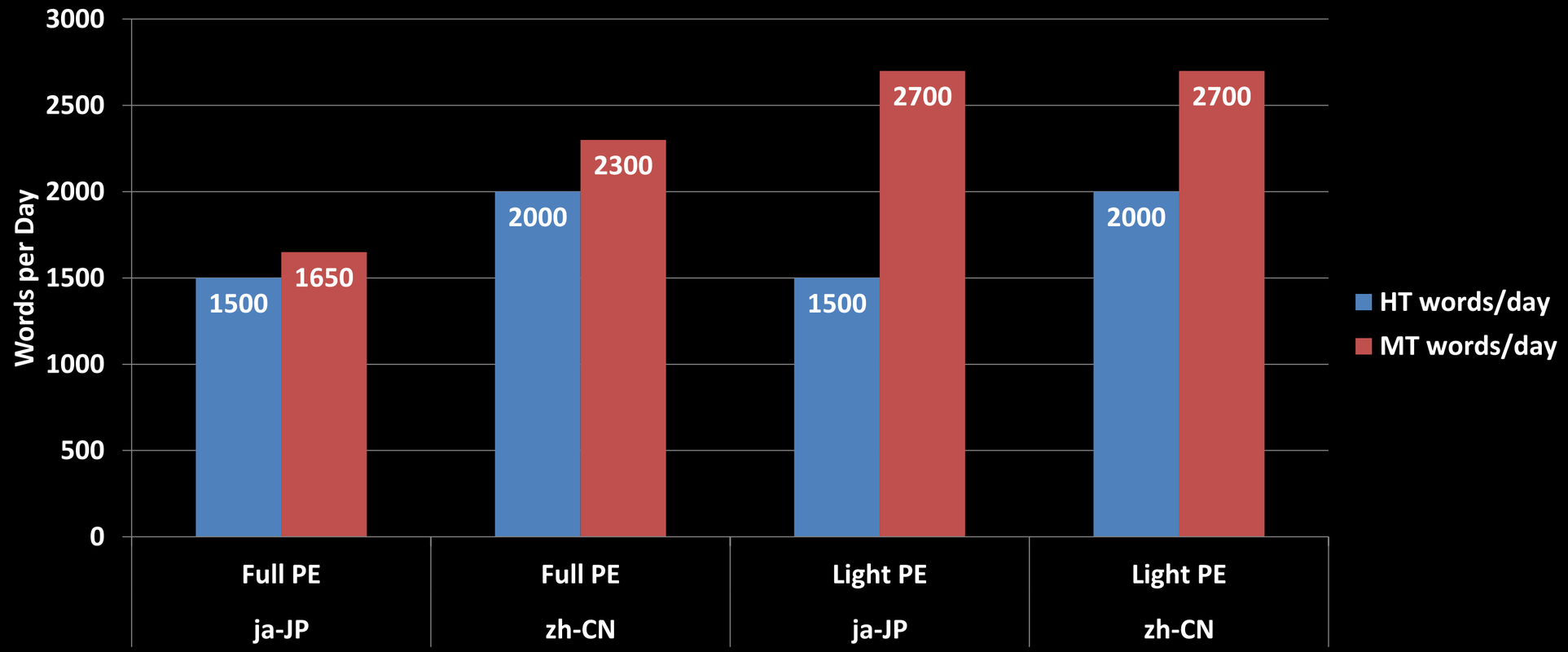
CASE STUDY: AUTOSCORES

AUTOSCORES FOR LIGHT AND FULL POST-EDITING



CASE STUDY: PRODUCTIVITY

PRODUCTIVITY FOR LIGHT AND FULL POST-EDITING



welocalize
doing things differently

CASE STUDY: LESSONS

LESSONS LEARNED

- Using autoscores on post-edited translations can indicate the level of post-editing effort involved for a specific content and MT engine
- The autoscores also illustrate the difference in effort between Light and Full Post-editing, approximately 20 point delta for BLEU and 15 point delta for TER
- The autoscores confirm that the resources have indeed managed to perform two distinct post-editing levels

ELAINEOCURRAN
Welocalize
October 2015

Enterprise Application of MT: Progress and Challenges

Craig Plesco
Nestor Rychtyckyj
Ford Motor Company
Dearborn, MI 48121

cplesco@ford.com
nrychtyc@ford.com

Abstract

At the AMTA conference in 2012 we first reported on the deployment of a global enterprise machine translation service at Ford Motor Company. As a multi-national company doing business around the world Ford has many requirements for translation. These translation requirements come in many different varieties and often lead to different solutions. As we have deployed machine translation for use throughout Ford, we are usually the first point of contact for new requests. In this presentation we will discuss the progress and challenges that we have faced over the last several years.

Our internal translation service uses the translation software from Systran Software Inc. (through the Systran Enterprise Server (SES) and Apptek Inc. We have developed over 90 Ford and industry-specific dictionaries that can be used to customize a translation request. The translation requests can be processed through a text box or with various supported file formats. In addition, we have developed a system interface for translations that allows applications to programmatically access the translation systems with background or real-time requests. These translation servers are hosted internally and provide a secure environment for our users. In order to address translation requests from new users, we have developed a process and template for users to describe their needs and expectations for translation.

We currently support 19 languages and process more than 200,000 translation requests daily. Batch translation requests are scheduled on a staggered timeline to enable reliable throughput. Our users include manufacturing, quality systems, plant floor systems, dealer concerns, warranty claims and other applications. As word of our service has spread throughout the company we are frequently approached to support other users and applications. These requests fall into a number of different categories.

In some cases we determine that MT is not the right solution. These include translations of legal or corporate documentation that would need significant customization and post-editing to deliver usable results. If we determine that MT is a viable solution we then perform text analytics on the source text to extract the terminology that will need translation. This extraction is done using natural language processing and ontologies with the goal of identifying “not found” terms and their frequency of usage. These “not found terms” are then translated manually and added into the translation glossaries.

Another common request is need for new languages which are not commonly available. In these cases we often work with the users and the vendor to apply statistical approaches to develop a language model more quickly than formal language modeling and with acceptable business accuracy. This was the approach that we used with Systran to build an English-Thai translation system. We have also developed and deployed a translation system with Systran for Romanian where the source text is converted to a parse tree prior to translation which results in much more accurate results.

We have found that Machine Translation has many significant advantages and uses in a corporate setting. However, it is critical to understand the user requirements and manage their expectations in regards to translation accuracy.

MACHINE TRANSLATION QUALITY ESTIMATION

A Linguist's Approach

ebay

WHAT IS MT QUALITY ESTIMATION?

Automatically providing a quality indicator for machine translation output without depending on human reference translations.

Our objective:
Estimate quality and post-editing effort for eBay listing titles and descriptions

ONE big CHALLENGE

$$\min_W \sum_{t=1}^{T} ||(W^{(t)}X^{(t)} - Y^{(t)})||2^2 + \lambda_s||S||_1 + \lambda_b||B||_{1,\infty} \text{ subject to: } W = S + B$$

or

"State-of-the-art QE explores different supervised linear or non-linear learning methods for regression or classification such as Support Vector Machines (SVM), different types of Decision Trees, Neural Networks, Elastic-Net, Gaussian Processes, Naive Bayes, among others"

(Machine Translation Quality Estimation Across Domains, de Souza et al, year))

A LINGUIST'S APPROACH

Using linguistic features from 3 dimensions:

COMPLEXITY **ADEQUACY** **FLUENCY**

FEATURES

Complexity:

- Length
- Polysemy

Adequacy:

- QA
  - Terminology
  - Patterns
  - Blacklist
  - Numbers
- Automated Post-Editing
- (POS)
- (NER)

NWT EUC Lot of 4 JCrew Old Navy Skinny Tank Tops Pink Green Grey Women's Med

$7.50 3 bids

Fluency:

- Misspellings
- Grammar errors

New Toshoba C50-A-053 Celeron 1.9Ghz 6GB DDR3 500GB 15.6" DVD HDMI USB 3.0 Si

IMPLEMENTATION

Checkmate+LanguageTool

Reusable Profile

Detailed Report

Score

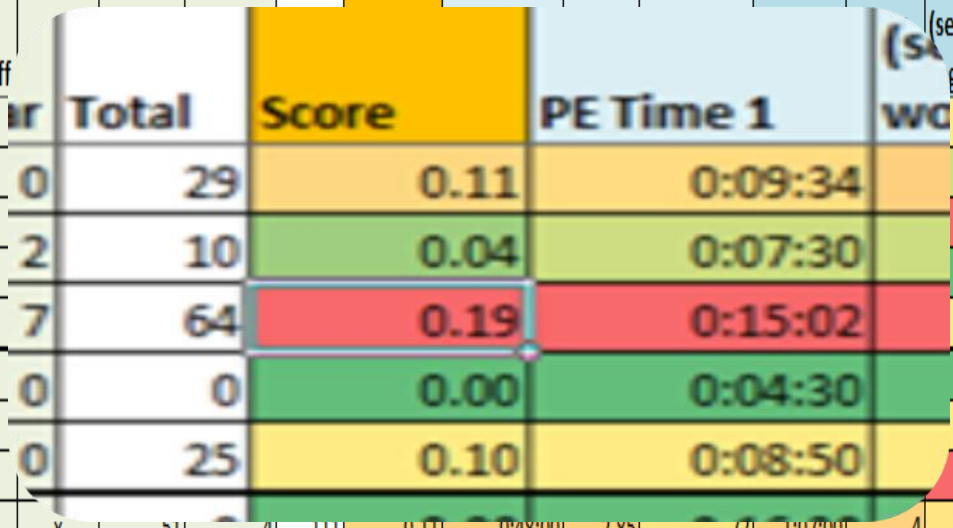
TESTING

- One Language (es-LA)
- Short samples (~300 words)
- Bigger samples (~1000 words)
- Post-Edited files (~50,000 words)
- pt-BR, ru-RU, zh-CN

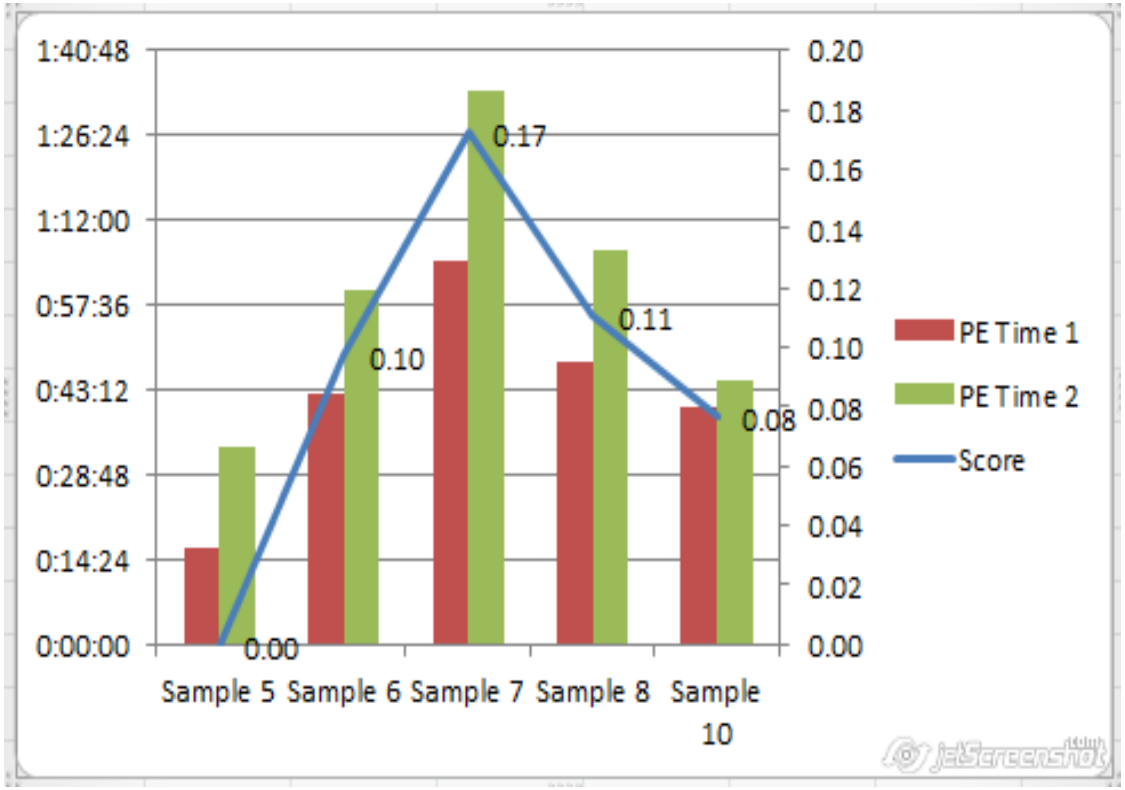
RESULTS

MEASURING RESULTS

	Statistic Data			Complexity		Adequacy				Fluency			Score	PE1			PE2		
Sample	Size (seg)	Size (w)	Aver. Seg (W/S)	Length	Polysemic	QA	Postprocessing Script changes	PoS Diff							PE Time	Ave PE time		PE Time	Ave PE time (secs x gment)
Sample 1 -	10	254	25	0	10	1	0	x											84
Sample 2 -	10	256	25	0	4	0	0	x											78
Sample 3 -	10	330	33	10	10	6	3	x											94.4
Sample 4	10	273	27	0	0	0	0	x											63.1
Sample 9	10	250	25	0	9	1	0	x											84
Sample 5	37	997	27	0	0	0	0	x											54.2
Sample 6	37	984	26	0	14	5	3	x											97.2
Sample 7	34	1033	30	10	31	10	1	x											165
Sample 8	40	1004	25	0	16	6	1	x	x	51	4	111	0.11	0:48:00	2.85	72	1:07:00	4	100.5
Sample 10	38	1015	26	0	20	6	3	x	x	14	4	78	0.08	0:40:30	2.39	63.9	0:45:00	2.66	



SAMPLES - SCORE AND TIME ALIGN



Sample	Score	PE Time 1	PE Time 2
Sample 5	0.00	0:16:30	0:33:27
Sample 6	0.10	0:42:45	1:00:00
Sample 7	0.17	1:05:00	1:34:00
Sample 8	0.11	0:48:00	1:07:00
Sample 10	0.08	0:40:30	0:45:00

FILES - SCORE AND ED ALIGN

XLIFF_66497_file6	Size (w)	Total	Score	Edit Distance
MT output	53777	4559	0.08478	
delivery 1	53777	3503	0.06514	64.06
delivery 2	53777	3484	0.06479	70.76

Average ED (es-LA, descriptions) = 72

MT QE OVER TIME

Sample	Size (w)	Total	Score	Date
mt_desc_items_latam_2014_02_19.400k.final_train_WLA1	47728	5558	0.11645	3/2/2014
mt_desc_items_latam_2014_02_19.400k.final_train_WLA2	46746	6135	0.13124	3/2/2014
mt_desc_items_latam_2014_02_19.400k.final_train_WLA3	47743	5814	0.12178	3/2/2014
mt_desc_items_latam_2014_02_19.400k.final_train_WLA4	46558	6054	0.13003	3/2/2014
66497 MT_descriptions_training_PE_ENESLA_File1	56779	5237	0.09223	4/25/2014
66497 MT_descriptions_training_PE_ENESLA_File2	56470	5475	0.09695	4/25/2014
66497 MT_descriptions_training_PE_ENESLA_File3	56714	5726	0.10096	4/25/2014
66497 MT_descriptions_training_PE_ENESLA_File4	56819	5546	0.09761	4/25/2014
66497 MT_descriptions_training_PE_ENESLA_File5	56286	5253	0.09333	4/25/2014
66497 MT_descriptions_training_PE_ENESLA_File6	53777	4559	0.08478	4/25/2014
68865_en-es-CO-descriptions-mar15-b-translated.xliff_0	50433	3640	0.07217	5/3/2015
1293_MT_eBay_descriptions_en_es_latam_30k_PE_P2-B_0	32209	3120	0.09687	7/2/2015
1294_MT_eBay_descriptions_en_es_latam_70k_PE_P1-B_1	21500	2266	0.10540	7/2/2015

SAMPLES - OTHER LANGUAGES

	Words	Issues	Score	PE Time
BPT				
PE Sample 1	500	21	0.0420	33 min
PE Sample 2	500	11	0.0220	14 min
PE Sample 3	500	7	0.0140	5 min
RU				
Sample 1	500	91	0.1820	58 min
Sample 2	500	75	0.1500	19.5 min
Sample 3	500	62	0.1240	9.3 min
ZHCN				
Sample 1	500	9	0.0180	39 min
Sample 2	500	8	0.0160	21 min
Sample 3	500	6	0.0120	15 min

CHALLENGES

- False positives
- Matching score and post-editing effort
- Same weight for all features

WHAT'S NEXT

- Tracking scores over time
- Adding scores to our post-editing tool
- Adding new languages
- Researching new features

HOW CAN *YOU* USE THIS?

- Tailor the model to your needs
- Estimate quality at the file/segment level
- Target post-editing, discard bad content
- Estimate post-editing effort/time
- Compare MT systems
- Monitor MT system progress

Q&A

THANK YOU!

jrowda@ebay.com

Industry Shared Metrics with the TAUS Dynamic Quality Dashboard and API

www.taus.net

TAUS

What About Translation Quality

Old School: "One size fits all"

Since the 1980's

LISA QA Model, SAE J2450 prescribe today's quality processes:

1. **Static**:
   - One quality fits all purposes, all content, all audiences
2. **Subjective**:
   - Evaluations are often subjective and anecdotal
3. **Costly**:
   - QE causes friction, delays
   - QE can cost up to 25% of total translation costs
4. **Non-transparent**:
   - Necessity without remedy

Dynamic
Quality
Framework

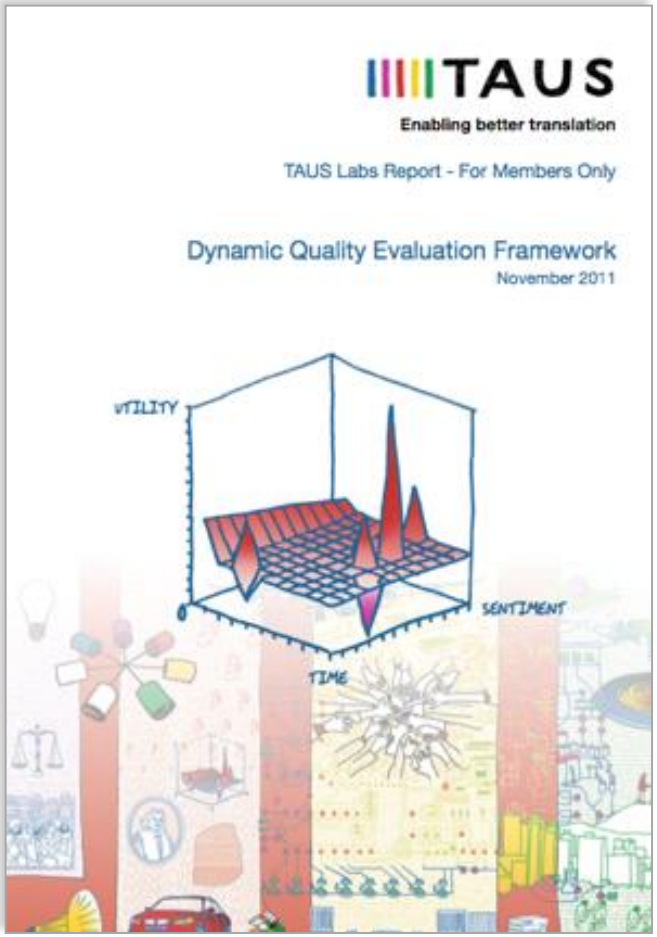
TAUS

Industry Collaborative Program

DQF started in 2011

Participating members

Adobe
Appen
Autodesk
AVB
CA Technologies
Cisco
Crestec
Crosslang
Dell
DFKI
eBay
EMC
Google
Hewlett Packard
Intel
LDS Church
Lingo24
Lionbridge
Medtronic
Microsoft
Moravia
Nikon
Oracle
Pactera
Pangeanic
Paypal
Philips
PTC
Siemens
Spil Games
Systran
VMware
Welocalize
Yahoo!



TAUS

From DQF Tools to Quality Dashboard

DQF Tools

Since January 2014

Tools on TAUS web site:

to measure:

- Productivity
- Adequacy
- Fluency

to review and count:

- Translation errors

to get:

- Stats and reports

Used by 100+ members

Quality Dashboard

Launched June 2015

DQF integrated in:

- CAT Tools
- TMS Systems

Use of DQF plug-in provides:

- Enhanced statistics
- Benchmarking

Open to everyone

TAUS

The Power and Value of the Quality Dashboard

- DQF collects data and generates reports on the Dashboard real-time
- Translators, managers, buyers, developers get their own stats, benchmarks and analytics
- Not only track and benchmark against your own data, but also against industry averages, between translators, customers, projects, technologies

Quality Evaluation  **Business Intelligence**

TAUS

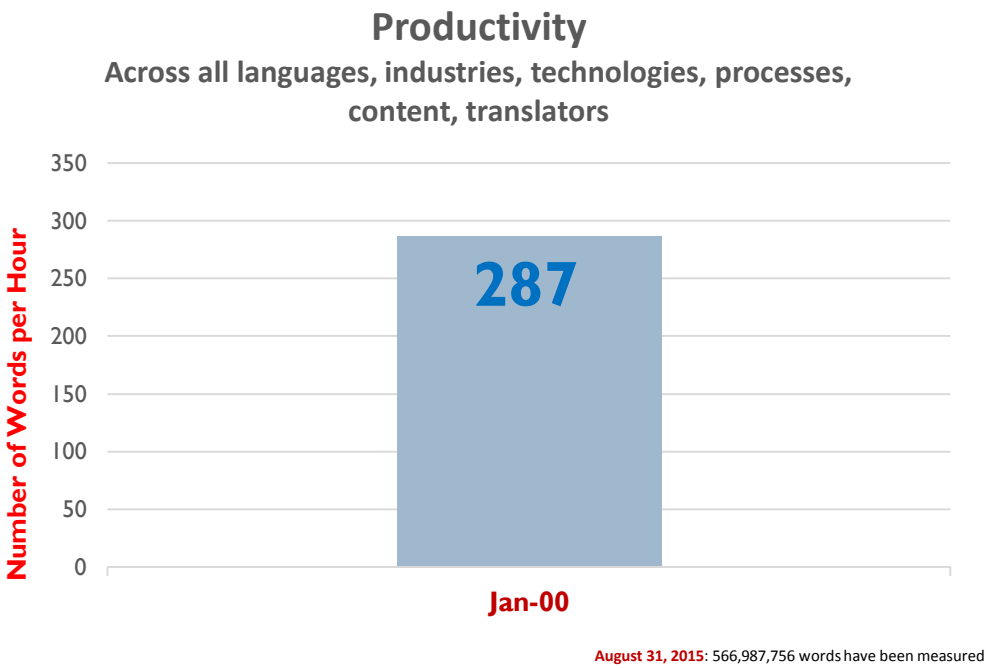
What is the average productivity?

TAUS Enabling Better Translation

Home Services Academy Events Blog Review Membership My TAUS

Quality Dashboard

- **Productivity**
- Efficiency
- Adequacy
- Fluency
  - Language
  - Time
  - Technology
  - Process
  - Content
  - Industry
  - Project
  - Translator/vendor
  - Customer
- Statistics
  - Language
  - Time
  - Technology
  - Process
  - Content
  - Industry
  - Translator/vendor
  - Customer
- Distribution of segments
  - Language
  - Time
  - Content
  - Industry
  - Project



August 31, 2015: 566,987,756 words have been measured

More information

Benchmark

Quality Dashboard
Content profiling
Quality evaluation
Adequacy/Fluency
Error review
MT Ranking
Productivity measurement
DQF API

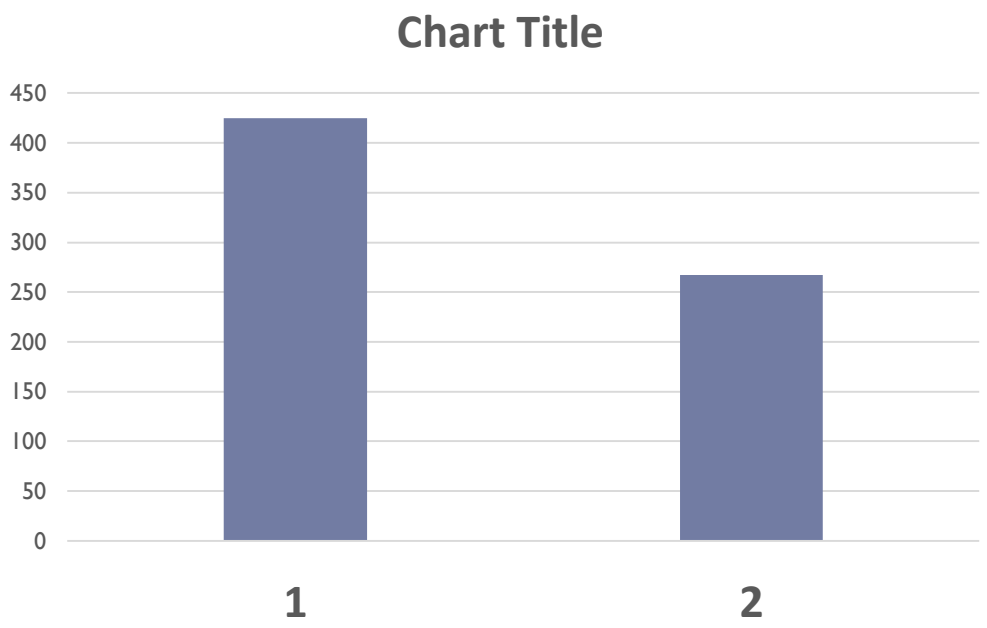
What is the average productivity of MT vs. TM?

TAUS
Enabling Better Translation

Home Services Academy Events Blog Review Membership My TAUS

Quality Dashboard

- **Productivity**
- Efficiency
- Adequacy
- Fluency
  - Language
  - Time
  - Technology
- **Process**
  - Content
  - Industry
  - Project
  - Translator/vendor
  - Customer
- Statistics
  - Language
  - Time
  - Technology
  - Process
  - Content
  - Industry
  - Translator/vendor
  - Customer
- Distribution of segments
  - Language
  - Time
  - Content
  - Industry
  - Project



August 31, 2015: 566,987,756 words have been measured

More information

Benchmark

Quality Dashboard
Content profiling
Quality evaluation
Adequacy/Fluency
Error review
MT Ranking
Productivity measurement
DQF API
API Integrators

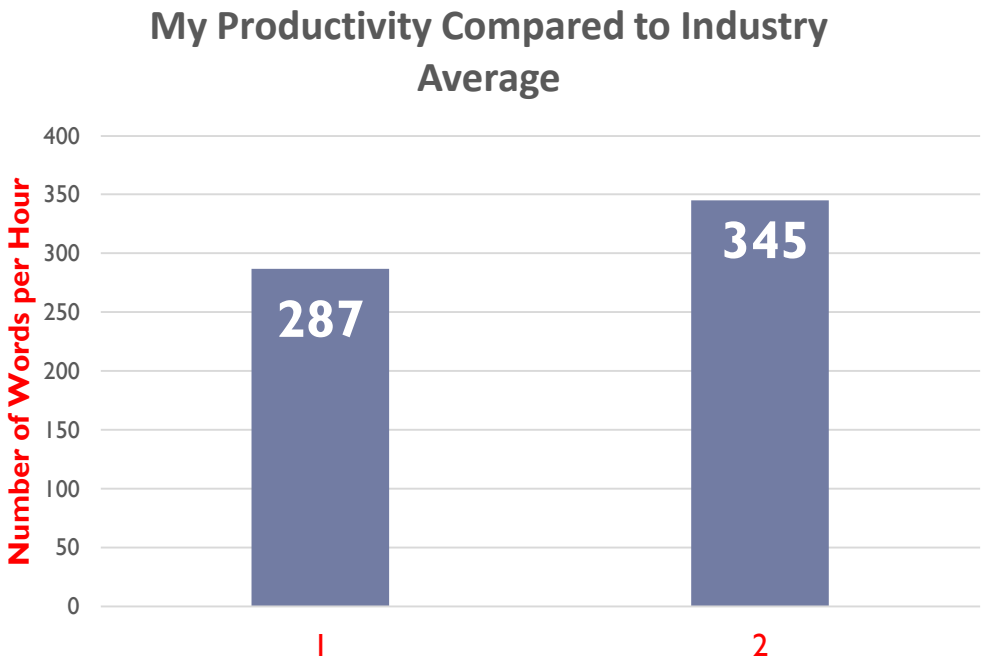
What is my productivity compared to industry?

TAUS
Enabling Better Translation

Home Services Academy Events Blog Review Membership My TAUS

Quality Dashboard

- **Productivity**
- Efficiency
- Adequacy
- Fluency
  - Language
  - Time
  - Technology
  - Process
  - Content
  - Industry
  - Project
  - Translator/vendor
  - Customer
- Statistics
  - Language
  - Time
  - Technology
  - Process
  - Content
  - Industry
  - Translator/vendor
  - Customer
- Distribution of segments
  - Language
  - Time
  - Content
  - Industry
  - Project



August 31, 2015: 566,987,756 words have been measured

More information

Benchmark

Quality Dashboard
Content profiling
Quality evaluation
Adequacy/Fluency
Error review
MT Ranking
Productivity measurement
DQF API

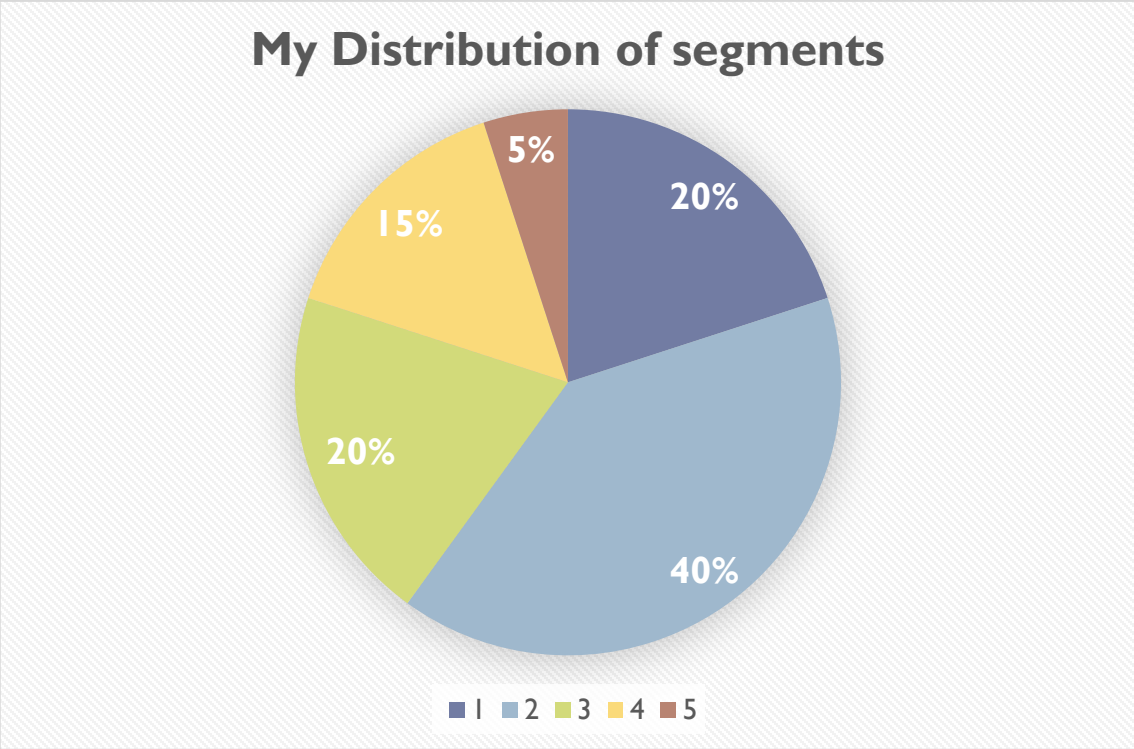
Where do my translations come from?

TAUS
Enabling Better Translation

Home Services Academy Events Blog Review Membership My TAUS

Quality Dashboard

- **Productivity**
- Efficiency
- Adequacy
- Fluency
  - Language
  - Time
  - Technology
  - Process
  - Content
  - Industry
  - Project
  - Translator/vendor
  - Customer
- Statistics
  - Language
  - Time
  - Technology
  - Process
  - Content
  - Industry
  - Translator/vendor
  - Customer
- Distribution of segments
  - Language
  - Time
  - Content
  - Industry
  - Project



August 31, 2015: 566,987,756 words have been measured

More information

Benchmark

Quality Dashboard
Content profiling
Quality evaluation
- Adequacy/Fluency
- Error review

MT Ranking
Productivity measurement
DQF API

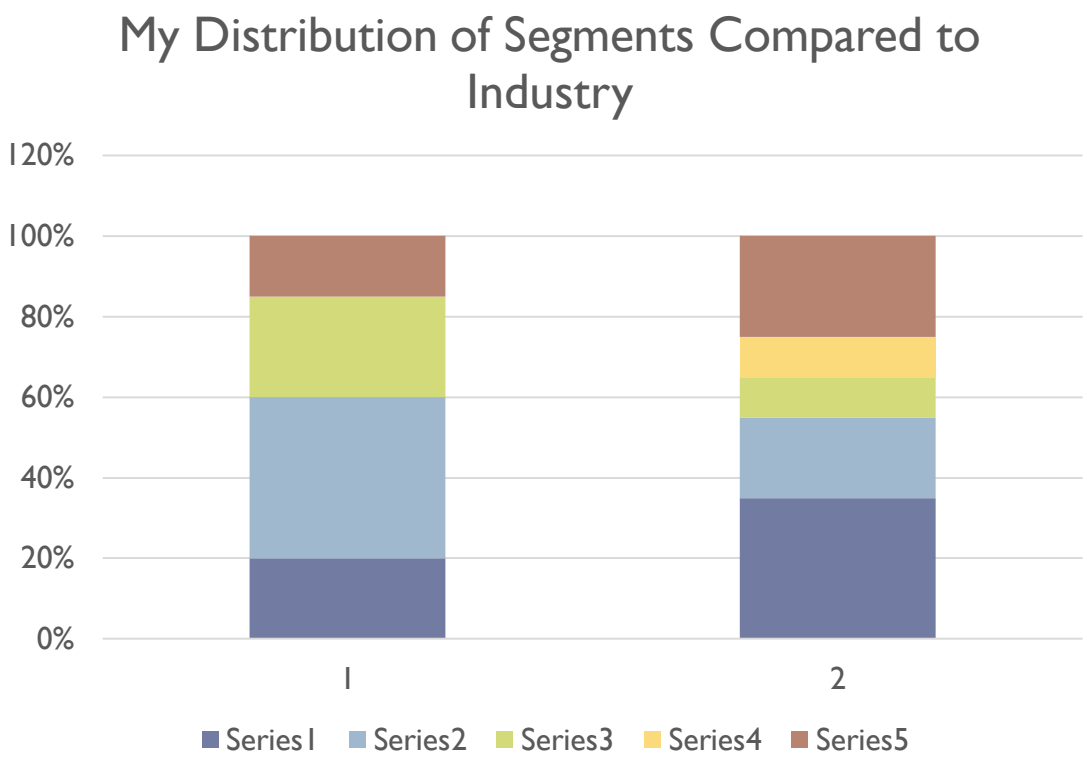
Where do my translations come from vs. industry?

TAUS
Enabling Better Translation

Home Services Academy Events Blog Review Membership My TAUS

Quality Dashboard

- **Productivity**
- Efficiency
- Adequacy
- Fluency
  - Language
  - Time
  - Technology
  - Process
  - Content
  - Industry
  - Project
  - Translator/vendor
  - Customer
- Statistics
  - Language
  - Time
  - Technology
  - Process
  - Content
  - Industry
  - Translator/vendor
  - Customer
- Distribution of segments
  - Language
  - Time
  - Content
  - Industry
  - Project



August 31, 2015: 566,987,756 words have been measured

More information

Benchmark

Quality Dashboard
Content profiling
Quality evaluation
Adequacy/Fluency
Error review
MT Ranking
Productivity measurement
DQF API

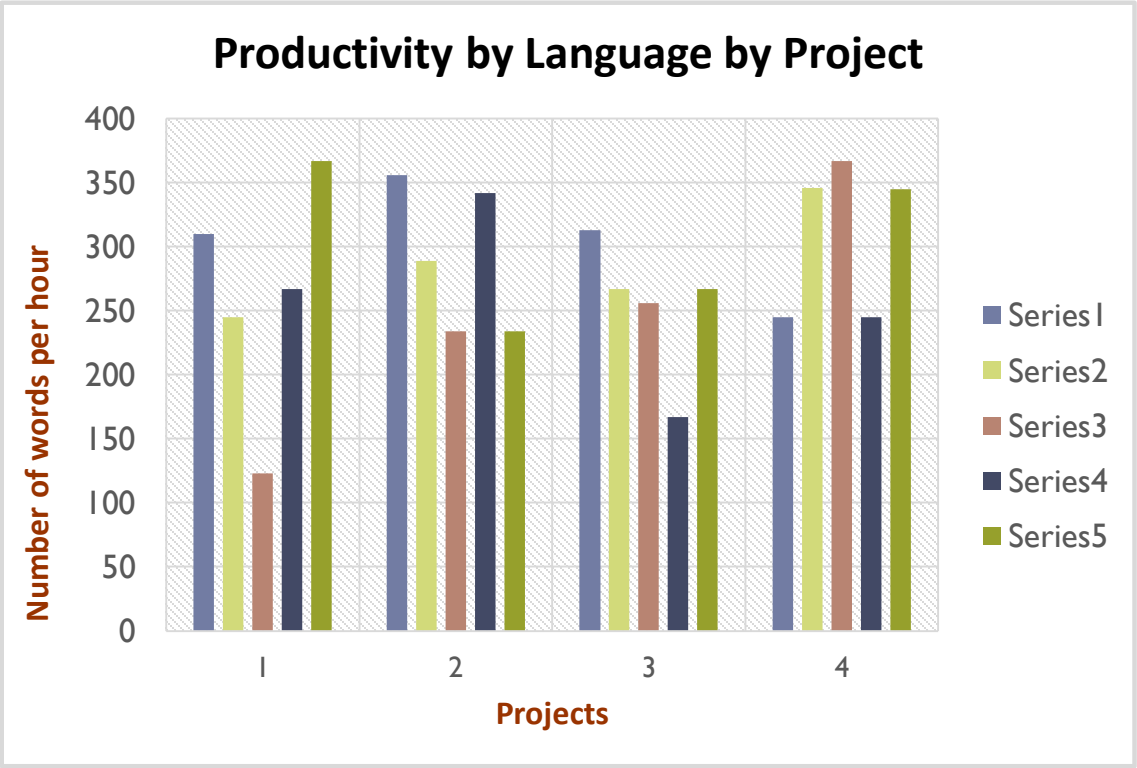
What is my productivity by language and by project?

TAUS Enabling Better Translation

Home Services Academy Events Blog Review Membership My TAUS

Quality Dashboard

- **Productivity**
- Efficiency
- Adequacy
- Fluency
  - **Language**
  - Time
  - Technology
  - Process
  - Content
  - Industry
  - **Project**
  - Translator/vendor
  - Customer
- Statistics
  - Language
  - Time
  - Technology
  - Process
  - Content
  - Industry
  - Translator/vendor
  - Customer
- Distribution of segments
  - Language
  - Time
  - Content
  - Industry
  - Project



August 31, 2015: 566,987,756 words have been measured

More information

Benchmark

Quality Dashboard

Content profiling

Quality evaluation

Adequacy/Fluency

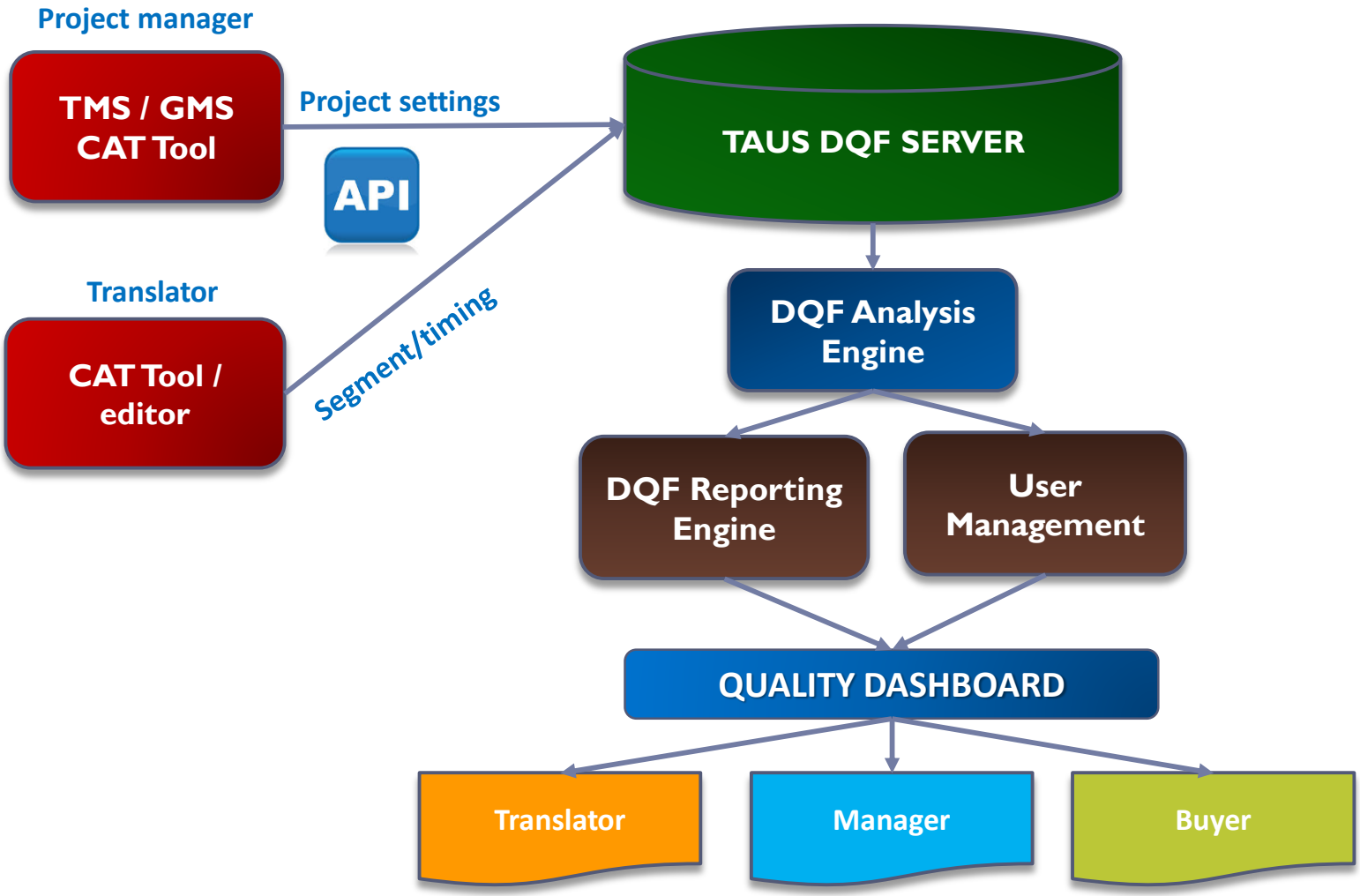
Error review

MT Ranking

Productivity measurement

DQF API

TAUS DQF Infrastructure



DQF Data Instrumentation

- Milliseconds per segment
- Source segment
- Target segment
- Edited target segments
- Time
- Language pair
- Project key
- Translator key

Open API

- Test Environment
  - https://dqf.taus.net/assets/api/v1/index.html
- Open API on GitHub
  - http://github.com/TAUSBV/dqf-api
  - Specification
  - Test Code
  - Documentation
  - Issue Tracker
- Available under the MIT Open Source License

Quality Dashboard Integrators

SDL WorldServer

matecat™

XTM International

SDL Trados Studio

Lingotek
THE TRANSLATION NETWORK

SDL Translation Management System

MEMSOURCE

Unbabel

"Microsoft Office International team is committed to the DQF model and approach and are actively partnering with TAUS to investigate how best to integrate TAUS Quality Dashboard API into our translation tool set."

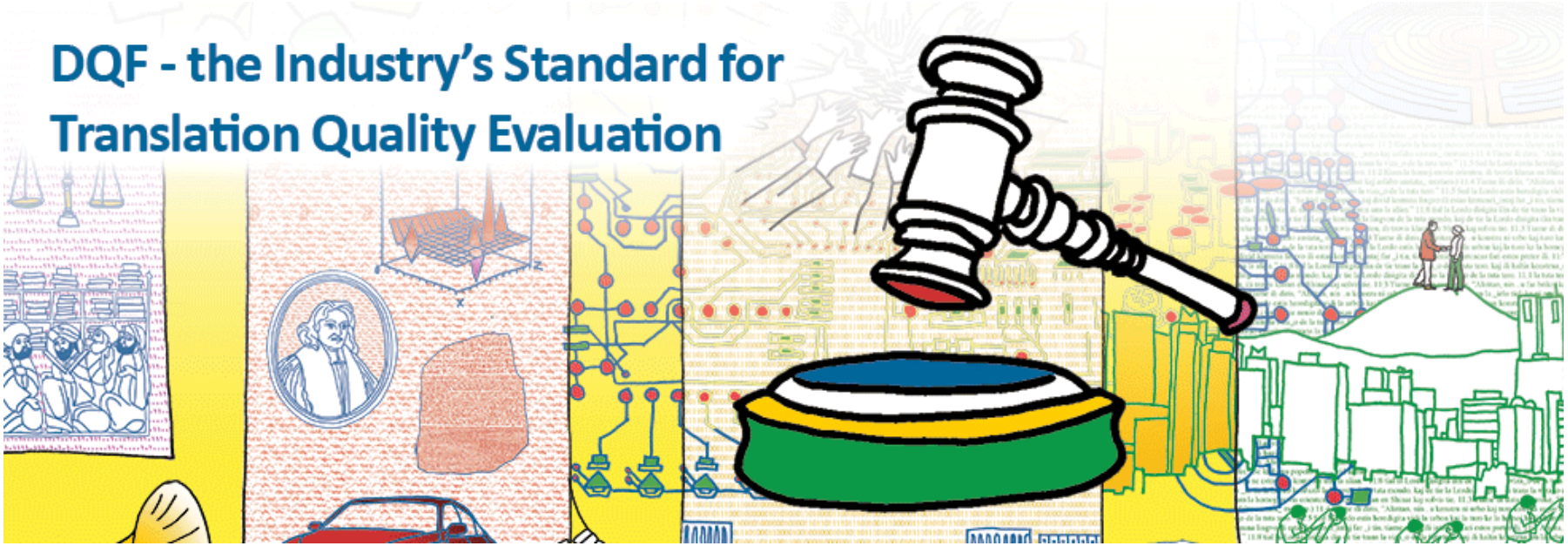
TAUS

The TAUS Efficiency Score

Introducing a new score for measuring productivity

- **2 Core variables:**
  - Words per Hour - WPH
  - Edits per Hour - EPH
- **Efficiency =** WPH + EPH
  - **Normalized** using Min-Max
- **Credit:** Nikos Argyropoulos

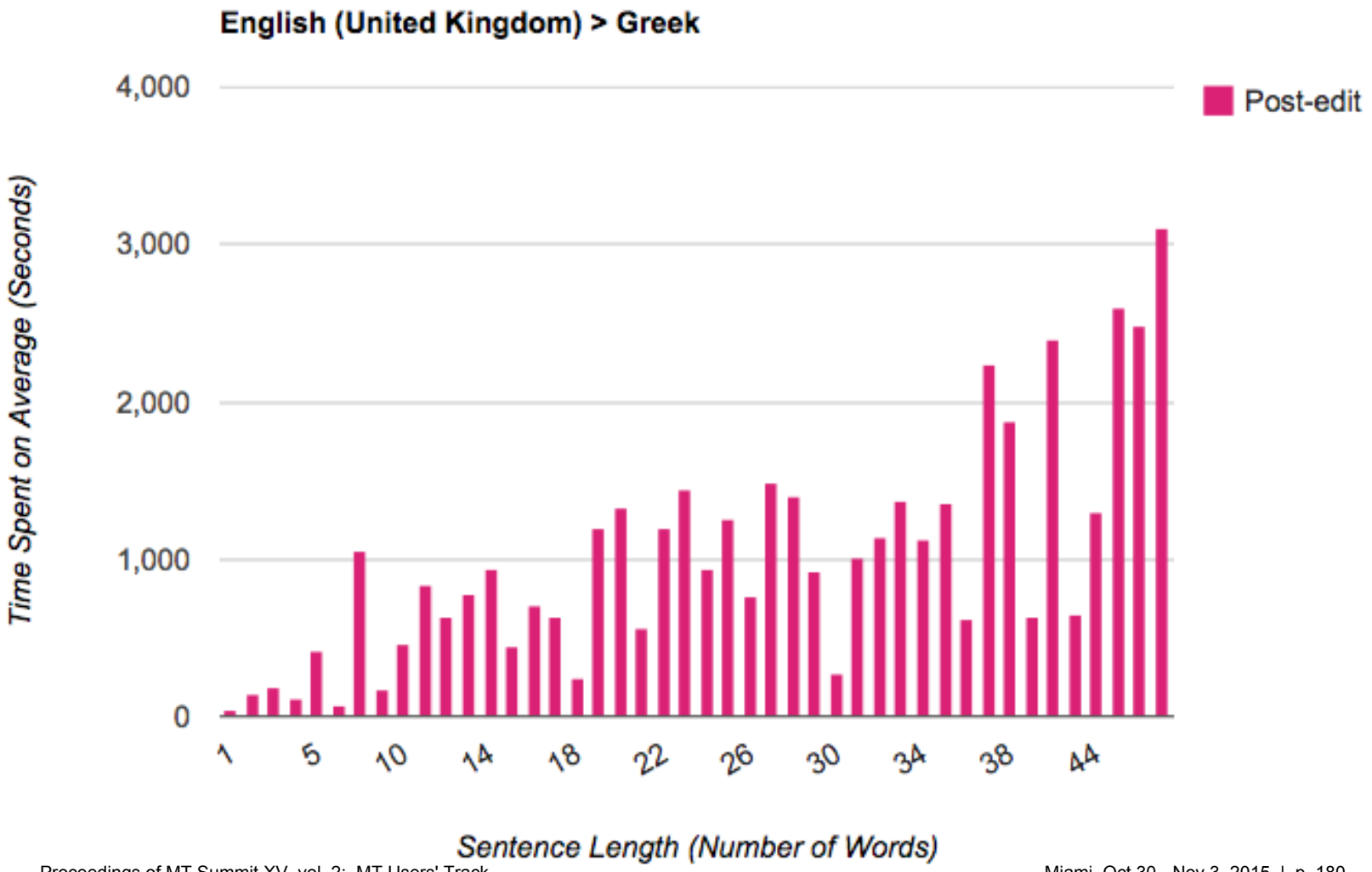
Productivity



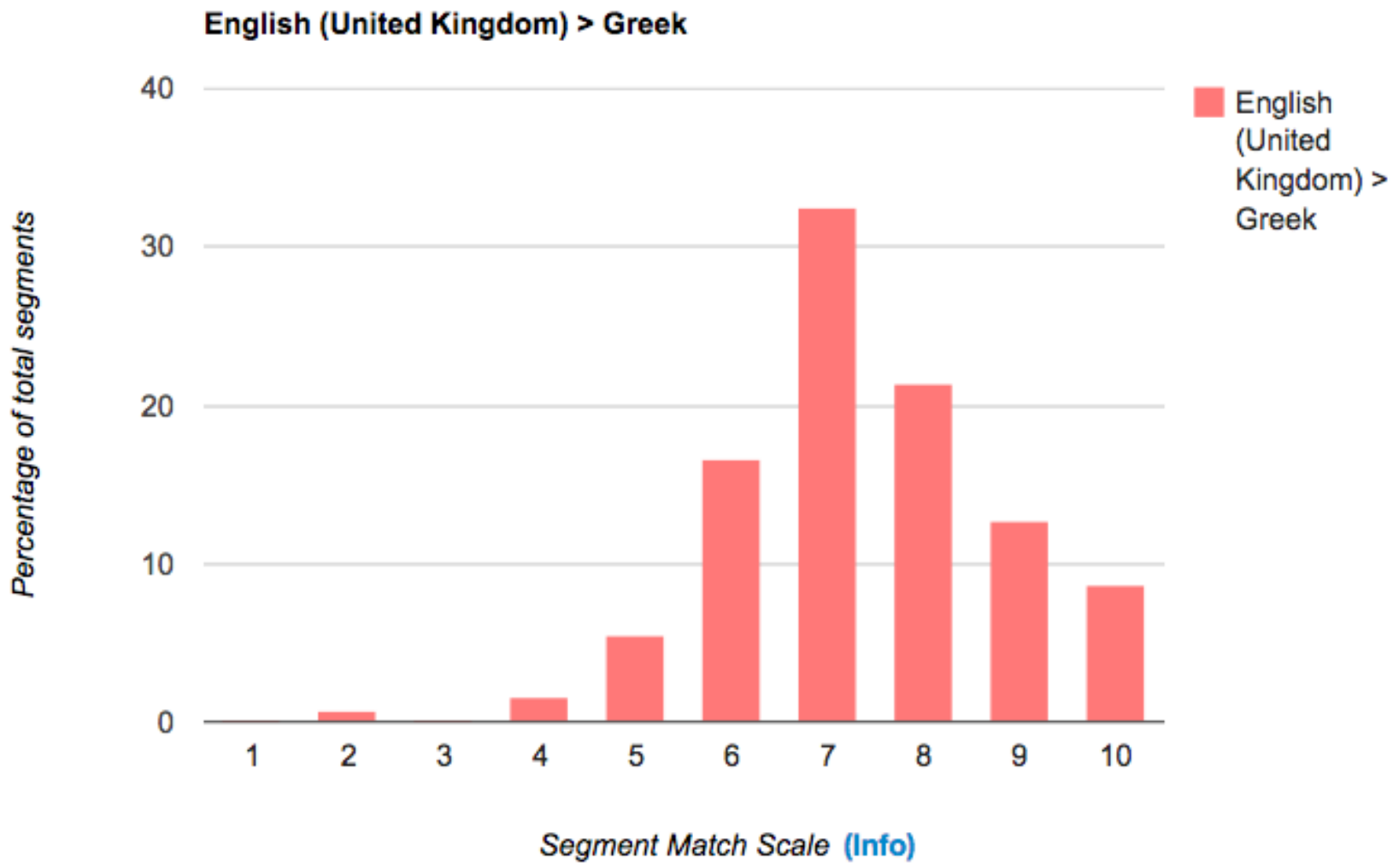
Average Productivity (Info)

Language Pair	Number of Segments	Number of Words	Post-edit (WPH)	Time Spent for Post-edit (seconds)	MT Engine
English (United Kingdom) > Greek	126	2,941	838	12,621	Not Specified

Average Time Spent by Sentence Length (Info)



Edit Distance Graph (Info)



Edit Distance

Levenshtein distance

The Levenshtein distance calculates how many operations are necessary to modify one sentence into another one. The number of single ***character edits*** (insertion, deletion, replacement) needed, is called the Levenshtein distance.

Efficiency = WPH + EPH

Name	Number of Words	Time (seconds)	Words per hour (WPH)	Edit distance	Edits per hour (EPH)
Translator 1	100	120	3000	50	1500
Translator 2	150	140	3857	80	2057
Translator 3	80	120	2400	70	2100
Translator 4	120	130	3323	30	831

Min-Max Normalization

$$X' = \frac{X - min_{WPH}}{max_{WPH} - min_{WPH}}(new_{max} - new_{\min}) + new_{min}$$

To normalize the value 3000 to a new range [0.0, 1.0] the following should be calculated,

$$X' = \frac{3000 - 2400}{3857 - 2400}(1.0 - 0) + 0 = \frac{600}{1457}1.0 = 0.411$$

So, by min-max normalization, the value 3000 in the WHP metric will be transformed to 0.411.

Normalized scores & Efficiency Score

Name	WHP normal	EPH normal.	Sum	Efficiency Score
Translator 1	**0.411**	**0.527**	0.938	**0.469**
Translator 2	**1.0**	**0.966**	1.966	**0.983**
Translator 3	**0.0**	**1.0**	1.0	**0.5**
Translator 4	**0.633**	**0.0**	0.633	**0.317**

Post-editor profiles

Evaluate Post-editors

The evaluation of the translators is based on the following selection:

Content Type: Website Content

Industry: Computer Software

Source Language: English (United States)

Target Language: Dutch (Netherlands)

Evaluator Name	Email	Number of Segments	Number of words	Time Spent for Post-edit (seconds)	Words per Hour	Total Edit distance	Score
User 1	user1@email.com	4	11	59.01	671.07	20	**1**
User 2	user2@email.com	4	11	61.48	644.11	8	**0.33**
User 3	user3@email.com	4	11	66.96	591.4	15	**0.29**

Post-editor profiles

	WHP	Post-editing	PE quality	Post-editing type	Result
1	High	High	Full	Fast & Aggressive	Good
2	High	Low	Light	Fast & Passive	Good
3	High	High	Light	Fast & Aggressive	Average
4	High	Low	Full	Fast & Passive	Average
5	Low	High	Full	Slow & Aggressive	Average
6	Low	Low	Light	Slow & Passive	Average
7	Low	High	Light	Slow & Aggressive	Not suitable
8	Low	Low	Full	Slow & Passive	Not suitable

Limitations and further work

- More data for benchmarking
- From relative to absolute scores
- 0 score theoretically possible = discouraging
- Eliminating outliers
- **Additional variables to include**

Additional variables to include

- **Keystrokes** – number of keystrokes
- **Mouse clicks** – number of clicks
- **TM fuzzy** – 0-100%
- **MT confidence** – 0-100%
- **Quality** – Review, automatic QA or manual QE
- **Difficulty of Source**
- **Experience** – number of words produced

Harmonized error-typology

DQF & MQM Harmonization

Cooperation with **DFKI** to harmonize **DQF** with **MQM** and standardize **Error categories and metrics**. A deliverable in the **EU project Q21**.

QT21
QUALITY TRANSLATION

TAUS

Accurately Predicting Post-editing Time & Labor for Cost-Management

Carla Schelfhout cschelfhout@sdl.com
SDL International, Maidenhead, UK

Abstract

This paper will describe a way of assessing the post-editing effort for a specific project, language and engine combination. This serves as a tool for LSPs to estimate the necessary effort on the project and quote accordingly.

.

1. Introduction

Over the last decade, there has been an upsurge in the use of machine translation, both for the purpose of gisting and for the purpose of post-editing. There are various definitions of post-editing: the "term used for the correction of machine translation output by human linguists/editors" (Veale and Way 1997), "…checking, proof-reading and revising translations carried out by any kind of translating automaton" (Gouadec 2007) and "In basic terms, the task of the post-editor is to edit, modify and/or correct pre-translated text that has been processed by a machine translation system from a source language into (a) target language(s)." (Allen 2003) are among them. All definitions center around the notion that a human applies changes to machine translation output in order to create a final translation that reaches a previously agreed quality level. We will refer to this process as PEMT (post-editing machine translation).

More and more clients ask their Language Service Providers (LSP) to apply PEMT. Their underlying assumption is that PEMT is faster than conventional translation, as part of the final translation is already there. The demand to use PEMT is mostly combined with a request to reduce the financial rates paid for translation. LSPs have a vested interest in determining if they can afford to comply with such requests. This paper will outline one way of validation.

2. Productivity of PEMT

Various studies have shown (Laubli et al 2013, Plitt & Masselo 2010) that PEMT as a process can provide productivity benefits over conventional translation, at least for the conditions examined in those papers. However, the productivity in a specific case depends on a large number of factors. To mention just a few:

- The engine quality as such. This depends on the technology that was used, but also on the complexity of the source-target language combination (some combinations are harder than others).
- The applicability of this specific engine to this specific project. Was the engine geared towards this particular content in any way, or is it a generic engine? Is the project in question terminology-rich and specific, or very generic?
- How much experience with PEMT do the selected vendors have?

- How has the technical preparation of the files been handled? Any wrong segmentation will likely have a detrimental effect on the machine translation quality.

As the LSP generally needs to provide a quote to the client before starting the job, and needs to negotiate rates with its own vendors as well, it is vital to know in an early stage whether the productivity increase from using PEMT is sufficient to allow a rate reduction to both the client and the vendors. Doing this fully automated would be ideal, but technology is not quite there yet. A human factor in productivity testing is still needed, but the testing needs to be both cost-effective (the cost of testing should not negate the gain of the project) and time-efficient (the LSP needs to have the results in time to quote to the client within his set time limits). The next sections will discuss SDL's approach to this dilemma.

3. Approach

The recommended process has three stages: assessing the project, creating the test set and running the test.

3.1. Assessing the project

The first step, which is still entirely human, consists of an analysis of the project. Aside from the normal steps used for conventional translation, the assessment for PEMT adds a few more questions. The main ones are:

- Does the expected gain of this project justify the cost of PEMT testing? Only if so, the next questions come up.
- What languages are in the project? Do we need/want to test them individually, or would it be possible to group them – for example, assume that the performance of English>French is a good indicator for English>Italian?
- Is the project homogeneous, or does it have flows? For example: a large car manufacturer could offer the user guides for buyers, the marketing brochures for prospective buyers, the repair manuals for the mechanics in the garage and the assembly instructions for the workers in the factory. It seems likely that the linguistic characteristics of these documents will differ, and so will the MTPE productivity. Does the client offer the flows split, or all together? Does it make sense to test them separately?

3.2. Creating the test set

Once a decision has been made, a test needs to be set up for the intended language and content type. This test set needs to be representative and varied.

The representativity of a test starts from the project sample delivered by the client. If not done before, now it needs to be confirmed with the client that their sample is in no way exceptional. The sample has to mimic the total project in (stylistic and terminological) complexity, content and technical characteristics (markup and segmentation). Ideally it will also be large and consist of several outtakes of the total project. This gives a larger variety in topics and the related terminology. Any non-representative or invalid content needs to be removed from the client sample before taking the next step.

In order to select the most representative test set, it is recommended to randomly select segments that have the average segment length of the sample, give or take one or two words. This can be automated, which saves time, and it increases the chance that the linguistic complexity of the test set will mimic the complexity of the sample. A couple of longer and shorter segments can be added to test on the less frequent segments and to add variety to the test. In

order to test the engine´s coverage of client-specific terminology, it is recommended to select a number of segments from various parts of the project rather than use running text, which will mostly cover only one or two topics and its related terminology.

For financial reasons, the smaller the test set can be while still giving meaningful results, the better. The smallest possible number of segments depends on the tool used for the testing and the margin of error this tool gives. While we recommend involving a statistician to assess the minimum for use with a specific tool, around 100 segments seems a good rule of thumb.

3.3. Running the test

The next step is to split the test set and have the two parts processed. One part will be done as conventional translation, the other part as PEMT. Both parts have to have the same average segment length to keep the times spent on them comparable. Both parts need to be done by the same resource(s), to ascertain the impact of the PEMT for this resource.

In order to increase the predictive value of the test for the project, it is advisable to use (some of) the resources who are likely to be used on the live project in the test. This will also help when it comes to negotiating rates for this particular project – having performed the test, they will have a better idea of its validity and of the MT value.

The key factor is the registration of times and actions to obtain meaningful results. Ideally, the interpretation of the test will consist of automated indicators to such a degree, that the test can be validated without having to read source or target language.

4. Tools

The more steps in the process can be automated, the cheaper each test becomes. Two steps are candidates for automation: the test bed selection and the hosting and analysis of the test. For the test selection, SDL has created a proprietary tool. It makes an automated selection out of one or more sample file(s), based on characteristics like segment length and linguistic characteristics like question/confirmation etc. Using this tool makes the test bed creation much faster, but as there is a development cost, it is only recommended if enough tests are needed in an LSP to recoup this cost.

The second step, running and analyzing the test, has been automated to a large degree in SDL. The ME tool is an online, proprietary tool, which has been used and further customized for a couple of years. In the meantime, similar functionality has been embedded in freeware like the qualitivity Studio plugin and the TAUS DQF framework. For the purpose of this paper, we will focus on the ME tool.

4.1. Characteristics of the ME tool

The tool is online, which means that resources can be onboarded by simply registering. This saves the overhead of resources downloading a tool, and prevents most of the complications of local PC setups causing incompatibilities. A test is uploaded as a tmx file. For the part of the test that is conventional translation, the source is copied into the target. For the PEMT part of the test, the MT is copied into the target field.

The tool displays the segments to the tester one at a time, only offering the next segment if the user clicks « Done ». It allows users to interrupt the test by clicking a button

« Continue later ». They can pick up the test at any later time. This keeps the time registration free of disruptions like telephone calls which would otherwise create noise in the results. Besides these buttons, the conventional part of the test only contains source and target fields, where the target field is editable. The PEMT part of the test has an extra field for the MT output, which remains in view for reference. The target field starts containing the MT output and can be edited. The PEMT UI also contains a button « Use MT » to indicate that the MT is correct as is and needs no changes. The button « Done » is only enabled after either selecting this checkbox or making edits, so as to prevent accidentally skipping segments. A screenshot is shown in Figure 1 :

Figure 1. Screenshot of PEMT part in ME tool

The number of segments needed in a test depends on the exact tool and setup used. In consultation with a statistician, it was decided that for the ME tool 80 segments is the minimum. The tool allows for larger tests, and enables the comparison of up to 5 different engines using up to 5 resources per test.

4.2. Analysis of the ME results

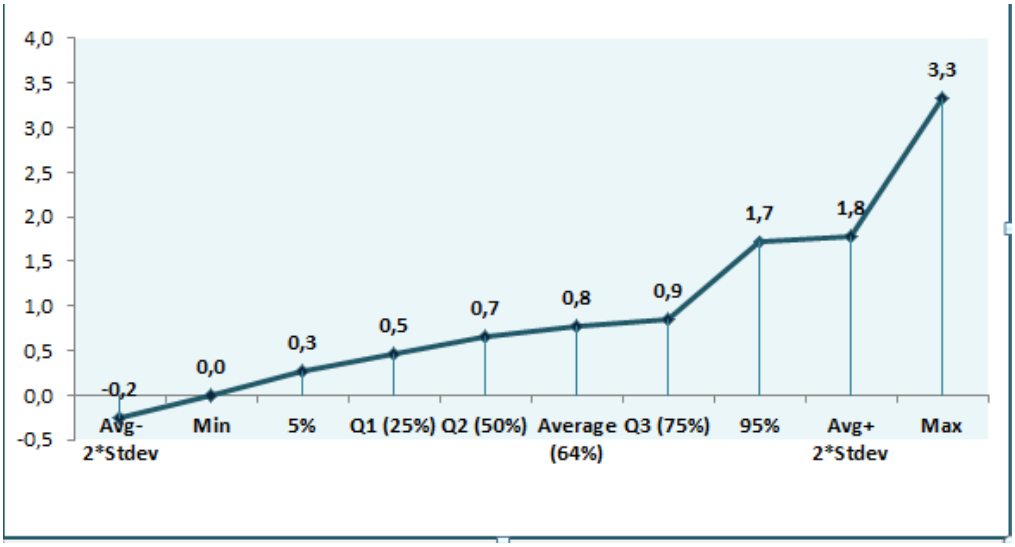
The tool delivers an automated analysis with a number of indicators. These serve to ascertain the validity of the test as well as the actual productivity increase. Among the indicators are:

- The translation speed for each resource in both conditions (conventional human translation, henceforth HT, and PEMT). This speed is not relevant for production as the tool is not mimicking the production environment, but extremely high or low numbers can point to a problem in the test.
- The actual increase in productivity as a percentage of the original speed for each resource. So if for example the HT speed is 800 words per hour and the PE speed is 880 words per hour, the productivity increase is 10%.
- The Levenshtein distance[1] per segment and on average for both conditions (human translation and PEMT).

[1] https://en.wikipedia.org/wiki/Levenshtein_distance

- Aside from the total time per segment, the tool also delivers the typing time per segment. The difference between the two is the time needed for reading source and target and thinking about the changes to make to the MT output.
- The number of times the resource interrupted the test. Depending on the size of the actual test, a low number could indicate that interruptions were accepted while the test was running, which would point to less reliable figures.
- The actions taken by the resources. Especially pasting actions without a copy-action are relevant, as they could point to copying from an outside source, like a Translation Memory. If so, the purpose of the test would be defeated.
- An overview of the time spent for the segments. An example is given in Figure 2.

Figure 2. Times spent on all segments



The Q-values indicate what percentage of the segments was done in less time than this value. So in Figure 2, the first quartile of the segments was done in less than 0.5 minutes. Half of the segments was done in 0.7 minutes. The relevant bit is the far right of the figure. If there is a sharp angle upwards, it indicates that one or very few segments took far more time than the others. This may indicate that the resource was distracted, or that the segment in question was very difficult. Such segments require further attention and if the segment is deemed unreliable, it needs removing from the test. The tool will automatically recalculate all values.

4.3. How to use the ME results

The approach discussed above was designed to give a prediction of the productivity of PEMT for a project. The careful selection of the segments is meant to enhance representativity, while the ME tool will give precise numbers. However, the selection of segments remains just a spotcheck of the total number of segments in the project. The productivity change coming out of the test will not be replicated exactly on each and every job in the project, even if the overall productivity is likely to be similar. For this reason, it is recommended to interpret the productivity figures in bands. For example: a gain of 20%-40% indicates a decent productivi-

ty gain for the project and could therefore give an LSP grounds to reduce the rate to the vendor by a commensurate amount.

When vendors have been introduced to the testing process, and have taken part in some tests, they will be better able to interpret test results and any associated rate discounts. Depending on the local vendor market, this can save quite some overhead on discussions about how valuable MT will be in this case and what rates vendors are willing to accept.

Please note that the PEMT process is only part of the overall translation delivery process. While PEMT may increase the productivity of this one step, compared to human translation, it will not have any beneficial impact on overhead like the downloading of files or engineering and desktop publishing effort. Also the step of reviewing translations is not sped up as such. The LSP will need to assess for every individual project what impact the productivity increase in PEMT will have on the total project before deciding on any rate reduction.

5. Impact of the tool

SDL have found that using the two tools described above (select data tool and ME tool), the time spent on creating and analyzing tests was reduced to one-fifth of what it was when humans built the tests, while following the same selection guidelines. Of course, developing both tools came at a cost as well. For the amount of tests SDL processes, this cost was recouped in a year.

The main advantage is that tests can be built, sent out and analyzed within one working day. Of course, in real life there may be delays for practical reasons – without sufficient heads-up, resources may not be available on demand, and depending on the languages, some resources may live in different time zones. But the amount of work to be done within the LSP is restricted, and that allows a fast turnaround time for any tenders or sales opportunities requiring a fast response. In a highly competitive market, this offers a huge benefit, as the following example will illustrate.

5.1. Example of ME usage

In order to illustrate the point : SDL was asked to quote on a PEMT project. The deadline was a week. ME testing indicated that the existing off-the-shelf engine would not offer sufficient productivity gain to bid with reduced rates. SDL requested the client TMs and built a new engine, which was tested on ME as well. This engine offered sufficient productivity gain. SDL was able to offer reduced rates and return the tender within the given timeframe and won the bid. Thanks to the tool, we could run two tests as well as build a new engine within the week.

6. Conclusion

In a market where PEMT is increasingly becoming a standard part of translation workflows, LSPs need to be more and more aware of what they can quote for a particular PEMT project. This paper has described how the ME tool, and the processes around it, can help determining the future gain for a particular project in an acceptable timeframe, thus giving LSPs safer ground to quote to their clients and a firmer stand with their resources.

References

Allen, Jeffrey (2003). Post-editing. In *Computers and translation: a translator's guide*, edited by Harold Somers. Pages 297-317.

Gouadec, Daniel (2007). *Translation as a profession.* John Benjamins Publishing.

Läubli, Samuel, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk (2013): Assessing post-editing efficiency in a realistic translation environment. In *Proceedings of the 2nd Workshop on Post-editing Technology and Practice (WPTP),* pages 83–91.

Plitt, Mirko and Francois Masselot (2010): A Productivity Test of Statistical Machine Translation Post-Editing in A Typical Localisation Context. In *Prague Bulletin of Mathematical Linguistics, 9.* Pages :7–16.

Veale, Tony and Way, Andy (1997). Gaijin: A Bootstrapping, Template-driven Approach to Example-based MT. In *Proceedings of the Recent Advances in Natural Language Processing.*

Adjusting Interaction Levels in a Speech Translation System for Healthcare

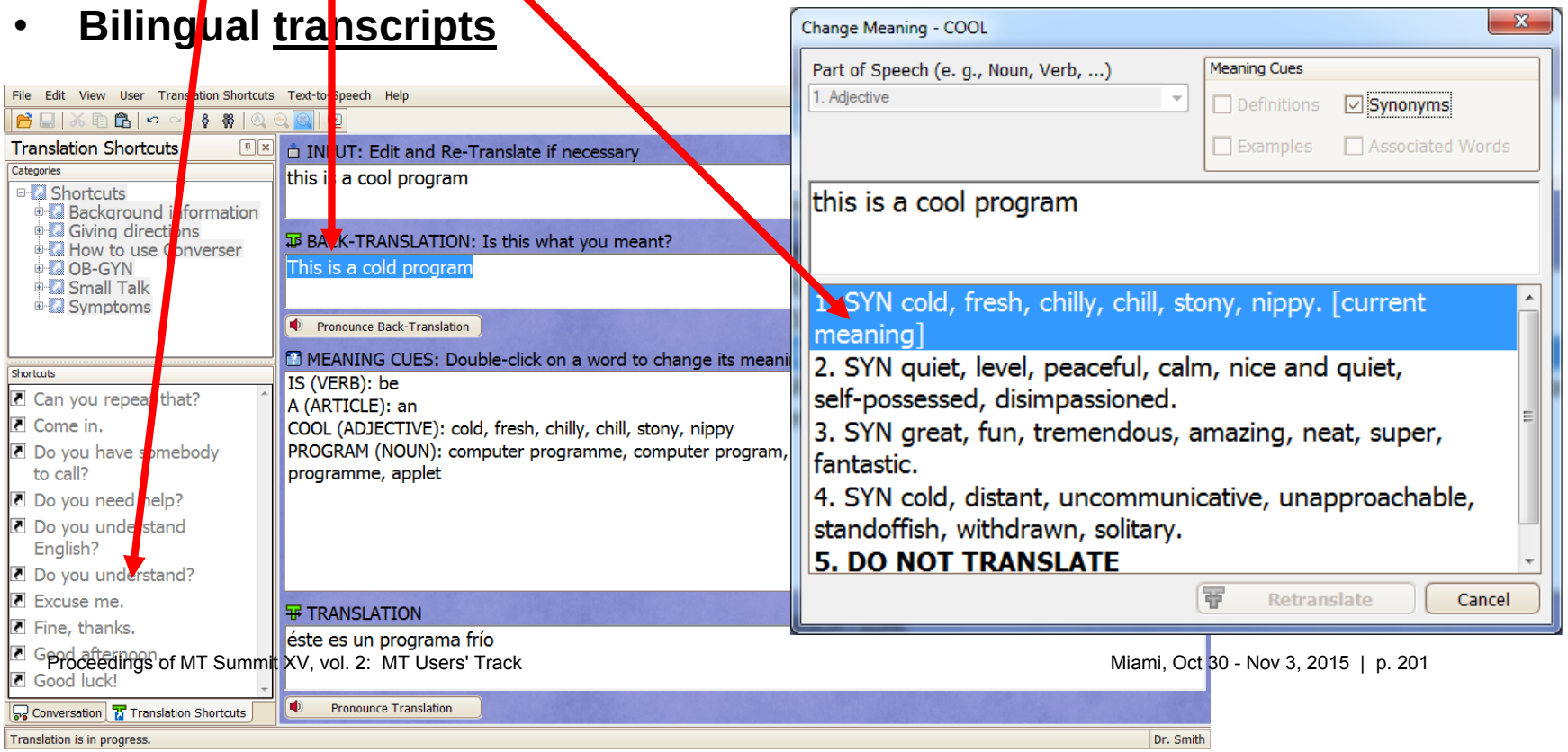
Mark Seligman
Mike Dillinger
Spoken Translation, Inc.
mark.seligman@spokentranslation.com

Introduction

- **Converser for Healthcare**
  - **Intro**
  - **Demo**
- **Kaiser Permanente pilot project**
  - **Needs and setup**
  - **Kaiser's evaluation (with numbers!)**
- **System revision**
  - **Especially … adjustment of interaction levels**
- **Future need for adjustment of interaction**
  - **Telepresence**
  - **Emergency response**
  - **Law enforcement**
  - **Military**

Converser for Healthcare: Intro

- **Patented verification and correction of translation**
  - Reliable Retranslation™
  - Meaning Cues ™
- **Customizable Translation Shortcuts™**
- **Bilingual transcripts**



Kaiser Permanente Pilot

- **Three departments at San Francisco Medical Center**
  - **Pharmacy:**
    - **Consulting** or **Drop-off** use case
      - *Shortcuts:* Consultation: Typhoid Vaccine
    - **Pickup** use case
    - **Greeter** use case
  - **Inpatient Nursing**
    - *Shortcuts:* IV, External Catheter, Pain Assessment
  - **Eye Care**
    - *Shortcuts:* Informed Consent for Cataract Surgery

Kaiser's Goals

Problem Project is Solving:

- Members' language needs remain unmet in many situations throughout the KP organization.

- Since the needs vary from situation to situation, no single solution can be expected. Different interpretative solutions need to be tested and analyzed to determine their best fit on multiple variables such as setting, situation, type of patient, etc.

- Accuracy of translation and member/patient acceptance of technology-based interpretive services vs. in-person interpretation need to be assessed.

Equipment (1): EliteBook Setup

- ***Good points:***
  - **EliteBook:** Fast; has touchscreen; runs standard image; foldable for portable use; has own keyboard
  - **Wacom Pen Display**: no handing computer back and forth
  - **TableMike:** excellent noise cancellation; hands-free operation; on-signal; easy to switch between staff and patient
- ***Drawbacks:***
  - Too much equipment for crowded areas
- ***Conclusions:***
  - Best for roomy over-the-counter situations with infrequent movement of equipment

Equipment (2): Motion Computing F5v Setup

- ***Good points:***
  - All functionality contained for one-handed portability
  - Liquid-tight for leak-proof sanitation
- ***Drawbacks:***
  - Sound volume too low for noisy settings
    - Aux speakers are unwelcome extra items
  - Docking station heavy, so stationary
  - Peripherals (keyboard, etc.) connect thru clip-on dock
  - Standard image not yet available
- ***Conclusions:***
  - Upgrade to MC J3500
    - Twin speakers for added volume
    - Portable clip-on keyboard: no need for dock
    - Touchscreen: minimize stylus use

Member/Patient Evaluation Comments

+	△
The system was described as: • "cool" • Useful – **5 mentions** • "looks good" "well done" • Would help • Good tool – **2-3 mentions** • I would recommend it • Even if translation was not 100%, it was always understood • "Perfect and clear" – **2 mentions** • Saving time – don't have to wait for an interpreter • "I like it" • "I like the idea of it" • Good for emergencies – **2 mentions**	• GUI too complicated (need larger buttons, crowded screen, …) – **6 mentions** • Literacy issues: some immigrants can't read or write – **6 mentions** • Font size too small - **3 mentions** • "Too technical for me" "I don't like computers": family say elderly can't use – **8 mentions** • Quality of Sound/Volume issues – **6 mentions** • Handwriting didn't work – **6 mentions** (Note: usage limited) • Worries about quality of translation – **2 mentions** • Keyboard issues (hard to use, pen is faster …) – **5 mentions** • Problems with English voice – **2 mentions** • System slow or froze – **6 mentions** • Hard to use tablet in hospital – **1-2 mentions**

General Member Comments

- Training (for users) would be needed – **4 mentions**
- Product would be “ideal” with voice recognition – **4 mentions**
- A lot of mixed comments – they like the system but worry others (elderly, less literate) will struggle with it (these comments came largely from partial or full English speaking members).
- Would rather have an in person interpreter – **4-5 mentions**

Staff Evaluation: 10 staff provided feedback

+	△
The system was described as: • Good for short interactions • Writing is easier than talking • Typing was easier than talking • You can verify translations better vs. Language Line – **2-3 mentions** • I would use it if no other options • Portability is good	• Occasionally missed a sentence • Computer literacy of members is a real issue – **3 mentions** (also elderly can't double click fast enough) • User Interface – buttons crowded • Translations were a bit odd • Slow • Hard for patients to write on the tablet, in bed – **2 mentions** • Takes (valuable) time for the system to process

- Training of patient's voice for DragonNaturallySpeaking would be needed.
  - But time is limited already (i.e. no time in visit to train patients) – **4 mentions**
- Training for staff and providers needed – **3 mentions**
- This product is really (more) needed for Cantonese/Mandarin here in San Francisco.
- The system needs a formal introduction (so system can describe itself, for English provider to use it with Spanish member)

Summary of Member/Patient & Staff Evaluations

- High praise for the “idea.” Higher than the actual experience of it.
- Translation quality definitely “good enough” as rated by Members/Patients.
- Limited English speakers (who can get along) would still use to verify the conversation and ensure completeness
- Issues of literacy and computer literacy impact applicability.
- Even though the system had issues (low to fair GUI, slow processing, lack of recognition of voice etc.), members partial or full English speakers thought it was “cool.”
- Most people, and especially those who lacked English skills, preferred an in-person interpreter. Although one person noted it saves time waiting for an interpreter, and a provider commented it saved the wait for Language Line.
- Good for emergencies
- Hard for members to use tablet in the hospital
- A number of patient declined to use in hospital but lacking data as to why.

Member/Patient Evaluation Summary

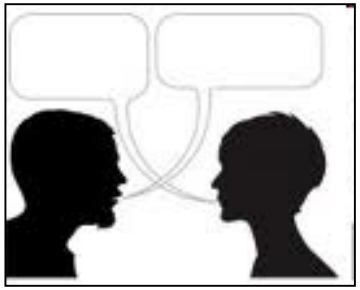
Member/Patient Evaluation	% answered question*	Rated (5) Completely and (4) Most
Did this meet your needs?	79%	94%
Was it accurate?	79%	90%
Was it easy to use?	72%	57%
Prefer handwriting question	67%	68%
Prefer using keyboard	67%	17%
Prefer to use handwriting and keyboard	67%	12%

Includes input from all settings: Outpatient Pharmacy, Hospitalized Patients, Outpatient Optometry.

Total of 61 interactions observed. Some patients declined to answer the question or were not asked the question.

Converser 4.0 Features (1)

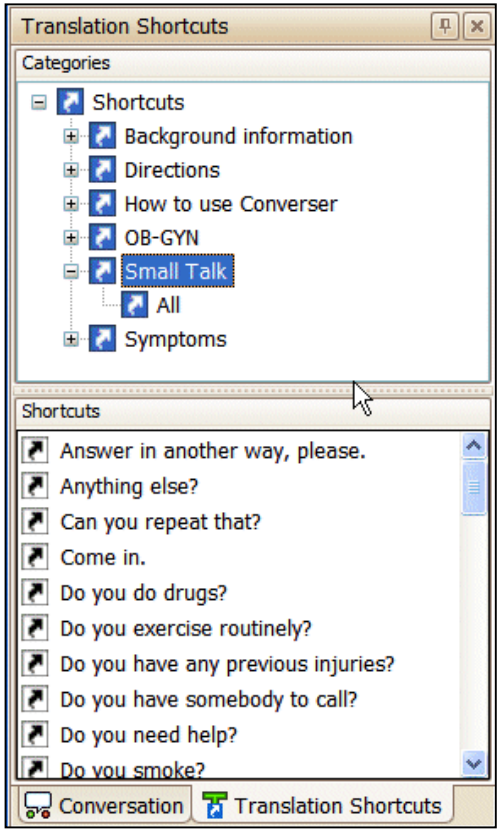
- **Speech recognition:**
  - ***Training-free* speech for both sides!**
    - **Spanish speech input enabled!**
  - On-screen push-to-talk button



- **Interface, training:**
  - Improved English<>Spanish switching
  - Large fonts for all windows
  - Eliminate in-person training
  - No-check Mode: can bypass MT verification

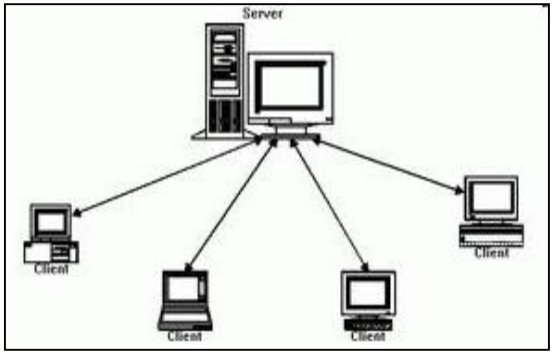
Converser 4.0 Features (2)

- **Translation Shortcuts:**
  - Many new categories
    - Emergency Room
    - Nutrition
    - …
  - New *Introducing Converser* Shortcuts
- **Text-to-speech:**
  - Speed controls for TTS

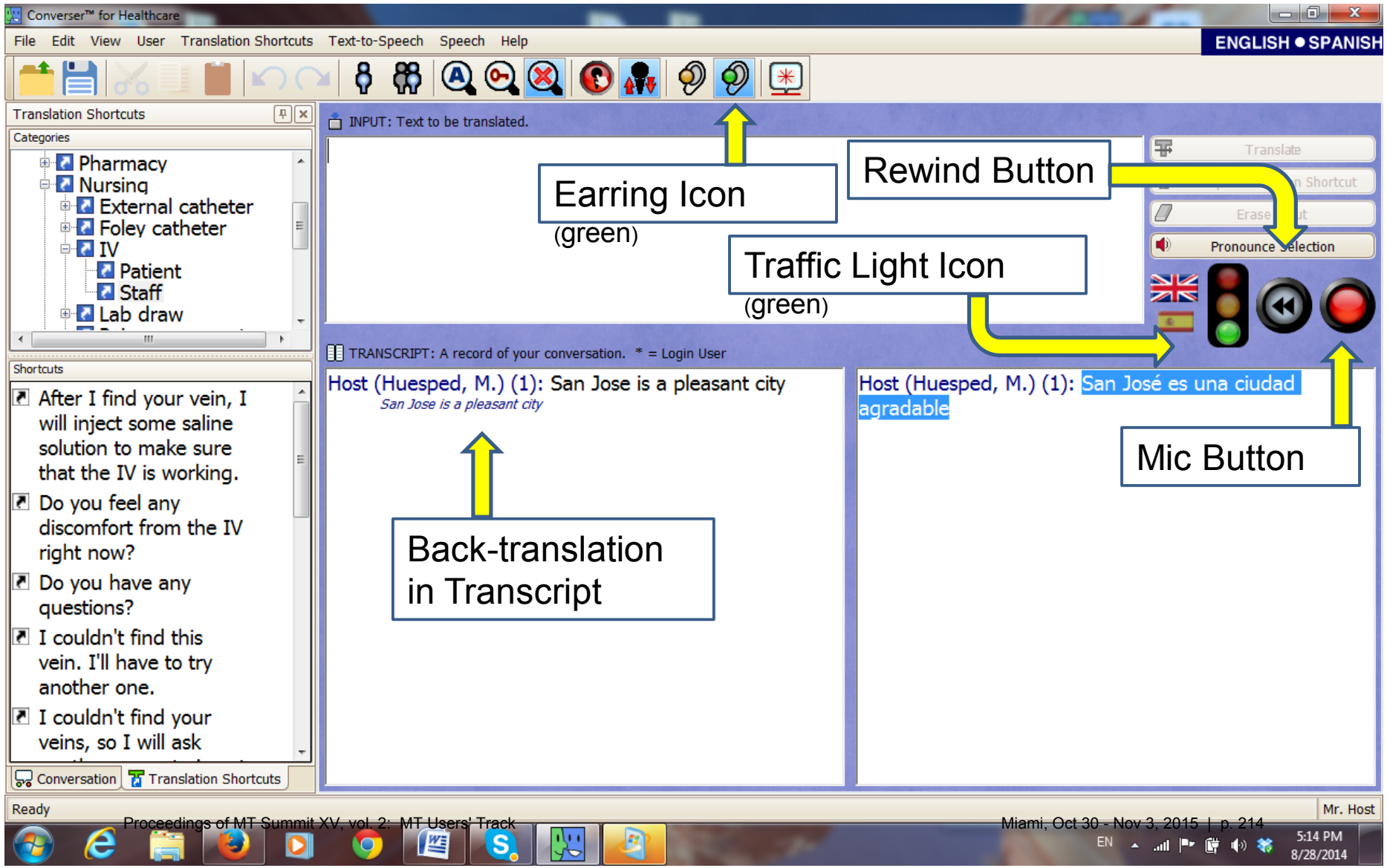


Converser 4.0 Features (3)

- **Handwriting:**
  - ➢ Improved correction interface
- **Typing:**
  - ➢ Onscreen keyboard with larger keys
  - ➢ Text entry by finger
- **Centralized installation, maintenance:**
  - ➢ Web-based delivery
  - ➢ Eliminate in-person maintenance



New Interaction Tools



Verification Controls

Yellow: Proceed with caution!
(Do pre-check.)

Red: Stop!
(Lock to prevent accidental use.)

Green: Full speed ahead!
(Don't pre-check … but transcript shows back-translation!)

Translation

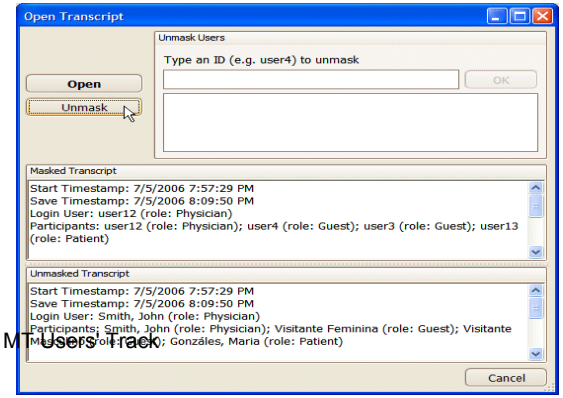
Speech Recognition

Future Need for Adjustment Tools

- **Beyond healthcare …**
- Telepresence
  - e.g. for business
- Emergency response
- Law enforcement
- Military

Future Features

- **Converser 5.0:**
  - Mobile delivery: e.g. iPhone, iPad
  - Other languages
    - E.g. English<>Chinese (Mandarin, Cantonese)
  - Transcripts:
    - Direct download to EMR
  - Personal, shared Shortcuts



Sendoff

Mark Seligman

- mark.seligman@spokentranslation.com
- spokentranslation.com

Beyond Text, Machine Translation and NLP for e-discovery

Jean Sennelart jean.senellart@systrangroup.com
SYSTRAN Global CTO/SYSTRAN SA CEO, Paris, 75002, France
Denis Gachot denis.gachot@systrangroup.com
SYSTRAN Software Inc President, San Diego, 92121, United States
Joshua Johanson joshua.johanson@systrangroup.com
SYSTRAN Software Inc, San Diego, 92121, United States

Abstract

As the amount and variety of digital data in different languages has increased, e-discovery processes need to evolve in order to streamline the processing of data and display crucial in-formation into an intuitive interface in a way that is scalable to any size user. It needs to be able to adapt to the user's needs and deal with any mix of media, such as image, voice, video, emails, blogs, social media posts and documents from any language in a manner that is con-sistent and intuitive. It needs to synthesize all of this information in a way that improves translation and exhibits the facts and evidences that the users need in the language of a digi-tal forensic examiner.

We will specifically describe how the tools implemented by SYSTRAN can be used to accomplish these more sophisticated tasks through several use cases. We will demonstrate how these tools can deal with multiple types of data, extract and normalize text, analyze lan-guage, create terminology lists, and customize the translation to better suit a variety of do-mains. We will illustrate how the tools accomplish this by self-tuning using unstructured and noisy corpora from the individual user and user-generated content written in approximate and sometime coded languages. This can be done across several languages, several types of data and multimodal documents, and can be scaled to suit the user's needs. We will show how these techniques can be used in combination to improve the overall translation quality and user experience.

Integration of SYSTRAN language libraries within an existing e-discovery platform will be presented to illustrate the presentation.

We will conclude by showing how these approaches can be generalized for big data analysis introducing challenges in real-time large scale data processing, but also processing of multi-topic and volatile information threads.

The full presentation can be found at http://static.systran.net/internaldocs/mtsummit-xv-systran.pdf

Designing User Experience for Machine Translated Conversations

Tanvi Surti tsurti@microsoft.com

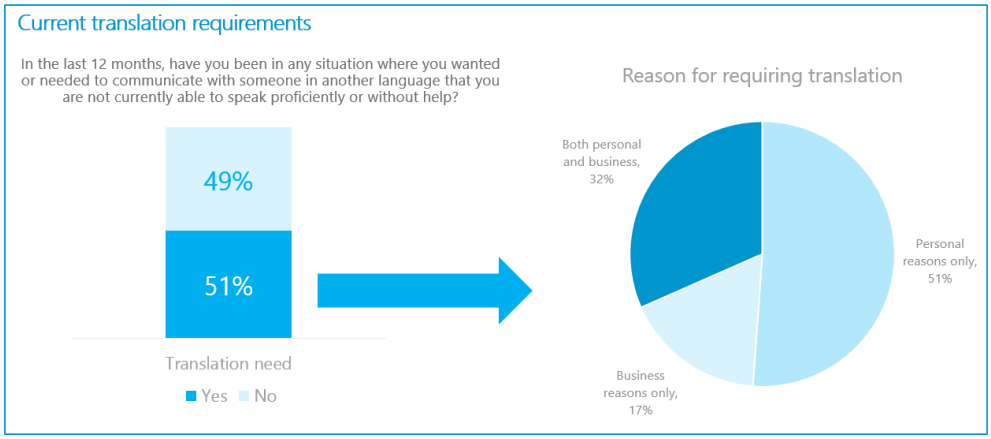
Abstract

Speech Translation technology in Skype enables users to have a live translated conversations across language barriers. From data collected from usability studies and thousands of Skype users, we've uncovered unique user experience challenges of a translated call that dissuade users from having conversations. This talk summarizes these findings and details how we iterate our designs to maintain a semblance of normalcy in translated conversations.

1. Introduction

The advent of deep neural networks in Automatic Speech Recognition (ASR) enabled researchers to reduce the word error rate in recognized speech by a third, and made it feasible to use ASR beyond the limited scope of SMS dictation, personal assistants, voice navigation, and made it applicable to the wider domain of every-day conversational speech. And by chaining together the ASR models with existing text translation, it now became possible to build automatic speech translation software for human conversations with previously unachievable accuracy.

This breakthrough in ASR resulted in the Skype Translator project, released in December 2014, enabling users to have automatic translated conversations over Skype. Skype Translator logged over 700 thousand app downloads over 9 months and clocked hundreds of hours in call time. There was clearly a need and interest in automatically-translated speech conversations, both in the personal and business sphere.



Yet, when users and reviewers tried Skype Translator for the first time, feedback for improvement was surprisingly equally concentrated around the experience of using Skype Translator as it was on the quality of translation on Skype Translator. In fact, users were more willing to forgive translation mistakes, acknowledging that speech translation was nascent technology; and were less patient with user experience issues as evidenced by the following excerpts taken from a usability study in February 2015 from first-time Skype Translator users –

"It (the call) was very chaotic."

"I zoned out waiting for the translations."

"I tried listening to the voice in the beginning, and when it wasn't working, I turned to the text."

"A bad translation is a conversation killer"

"I know that this is a monumental task and will revolutionize technology… but there isn't a flow in communication …"

"I felt like four people were speaking - two in English and two in Spanish"

Skype Translator Usability Lab, Mountain View – February 3rd 2015

2. User Experience areas of focus

Based on our usability studies and data from real-world users using Skype Translator, we identified that user experience of a translated call was a top pain-point for Skype Translator users, and over several design iterations, here are the top aspects of translated call experience we've addressed with some success -

2.1. First-Run and Learning Curve

Early usage data for the Skype Translator showed that ~40% of users had not made more than two calls on Skype Translator and that most calls on Skype Translator were under several minutes. This can be attributed to several issues such as poor translation quality and connectivity problems; but one underlying issue that emerged was that users didn't

know how to conduct a translated call. Having a translated Skype call was dissimilar to a normal Skype call, because included learnt behaviours such as waiting for the translation audio to play, remembering to pause between sentences and avoiding interruptions.

This first-run issue was addressed with two UX solutions.

Solutions

- User Education Video – all first-run users were taken through a two-minute explanation video to walk them through how to conduct a translated call on the first use of Skype Translator
- Tooltips – first-run users were given useful tips during their first translated call which provided context for how to have a successful translated call such as reminders to wear a headset.

2.2. Sensory overload

After his first call on Skype Translator, one male user study participant sat back and proclaimed - "You have to be a woman to be able to multitask in this thing…"

The sentiment he expressed referred to the multitude text and audio output the user receives during a translated call. First, there are four voices in the call – the caller's, the caller's translated voice, the callee's and the callee's translated voice. This gets cacophonous, especially when sequences of utterances are said in quick succession. Secondly, along with the audio, the user is also reading along to the translated transcript for her utterance and her partner's utterance in both languages. Many users complained that this was a lot of feedback to follow at once while trying to conduct a normal conversation.

Solution

- Audio Ducking – A technique used on radio, where if two audio clips are played at the same time, the volume is lowered on the less relevant once. Similarly, for Skype Translator, if translated audio and original audio is played at the same time, audio ducking is used to reduce the volume on the audio in the foreign language.

2.3. Perceived Translation Speed

Another frequently heard area of feedback from users was around the slowness of translation. Users felt they had to wait a long time to hear and read their partner's translated utterance and therefore made the conversation seem stretched out and awkward.

To a large extent, this delay is a *perceived* speed issue, on account of the fact that a user's translated audio could not be played until the user had completed their utterance, so as to not interrupt the user in the middle of their speech.

Solutions

- Partial recognition – Partial recognition enabled Skype to return partially understood utterances before the user had finished speaking. These "partials" are displayed in the transcript pane so that the user could follow along minimally to what their partner is saying.

- Silence interval – This advanced-user setting let users changed the value of the amount of time Skype would wait before translating their utterance. This allowed for users with a faster cadence of speech to set a low silence interval value that allowed their speech to be translated quicker.

2.4. Misrecognitions and Mistranslations

The most frequent problem users encounter during a translated conversation is misrecognitions and mistranslations. Some users see misrecognitions more than others, usually users with regional accents or children because of the lack of training data for these types of speech.

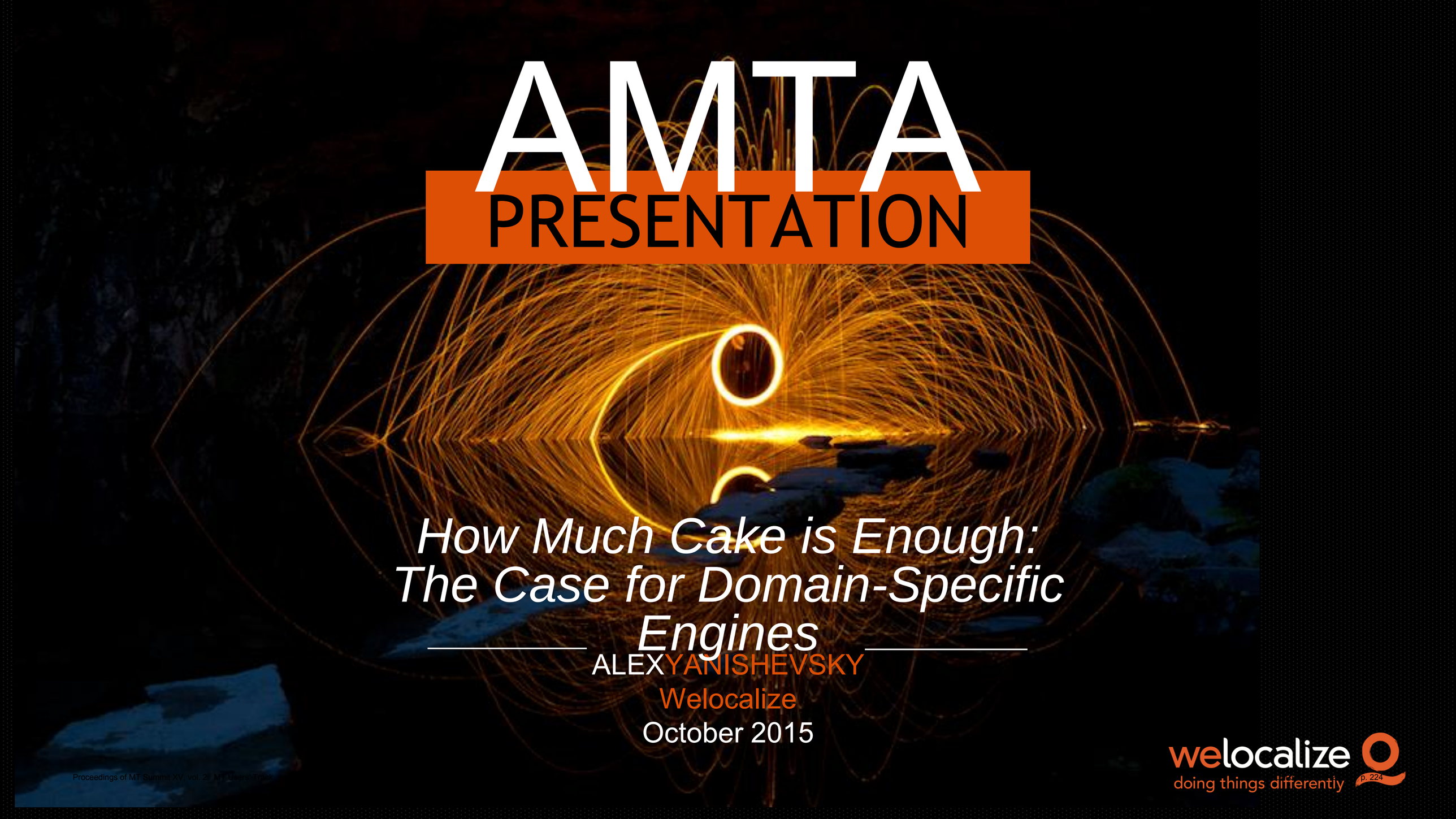
We reviewed several unsuccessful approaches to equip users to address misrecognitions and mistranslations. In the first iteration of the design, we tried to get users to cancel out wrong recognitions by clicking on a cancel button which their partner would also be able to see. In another iteration, we attempted to get users to correct the mistranslation by typing in the correct recognition instead by clicking on an Edit button. However, subsequent user studies demonstrated that users were generally unwilling to switch modalities from speaking to typing and clicking.

Solutions

- Basic user education – During the first-run setup video, users were told to repeat themselves when they were misrecognized or to rephrase their statement.
- IM prompt – Skype Translator tracked the confidence scores in the last five user utterances. If Skype Translator saw repeated low-confidence recognitions from users, the user was told that they should use the chat window to type to communicate instead. Therefore users with consistently bad recognitions were prompted towards a workaround.

3. Conclusion

Our research around Skype Translator revealed the importance of good user experience and design during a translated speech conversation. Users can be taught, over time, how best to leverage translation capabilities without expecting perfection, if the translation software sets the right context for them. Over time, users can learn to use speech translation tools in day-to-day communication along with a healthy caution to not expect perfection.



AMTA PRESENTATION

How Much Cake is Enough: The Case for Domain-Specific Engines

ALEXYANISHEVSKY
Welocalize

October 2015

welocalize
doing things differently

How Much Cake Is Too Much Cake?

What is the Tipping Point?

AGENDA

- How Many Engines
- How to Split Domains
- How to Measure Success
- How to Improve

welocalize
doing things differently

HOW MANY ENGINES: CRITERIA

- Environment: Elegant Deployment?
- Cost
- How Different are They From Each Other?
- Maintenance: Engineering + Linguistic Feedback Implementation

HOW TO SPLIT DOMAINS: CRITERIA

- Content Owner Feedback
- Historical Experience Based On Business Unit or Portfolio
- Naming Convention
- Style Analysis: Difference in Characteristics Based on Lexical Diversity, Sentence Length + Syntactic Complexity

HOW TO SPLIT DOMAINS: TOOLS

HOLISTIC APPROACH BASED ON SEVERAL TOOLS:

- Build Domain-Specific Language Models + Select TUs for Domain by PPL
- Source Content Profiler – Helps Identify Domain Based on Language Models, as well as Other Stylistic Characteristics
- Style Scorer – Higher Score Indicates Better Match to Style Established by Client's Documents

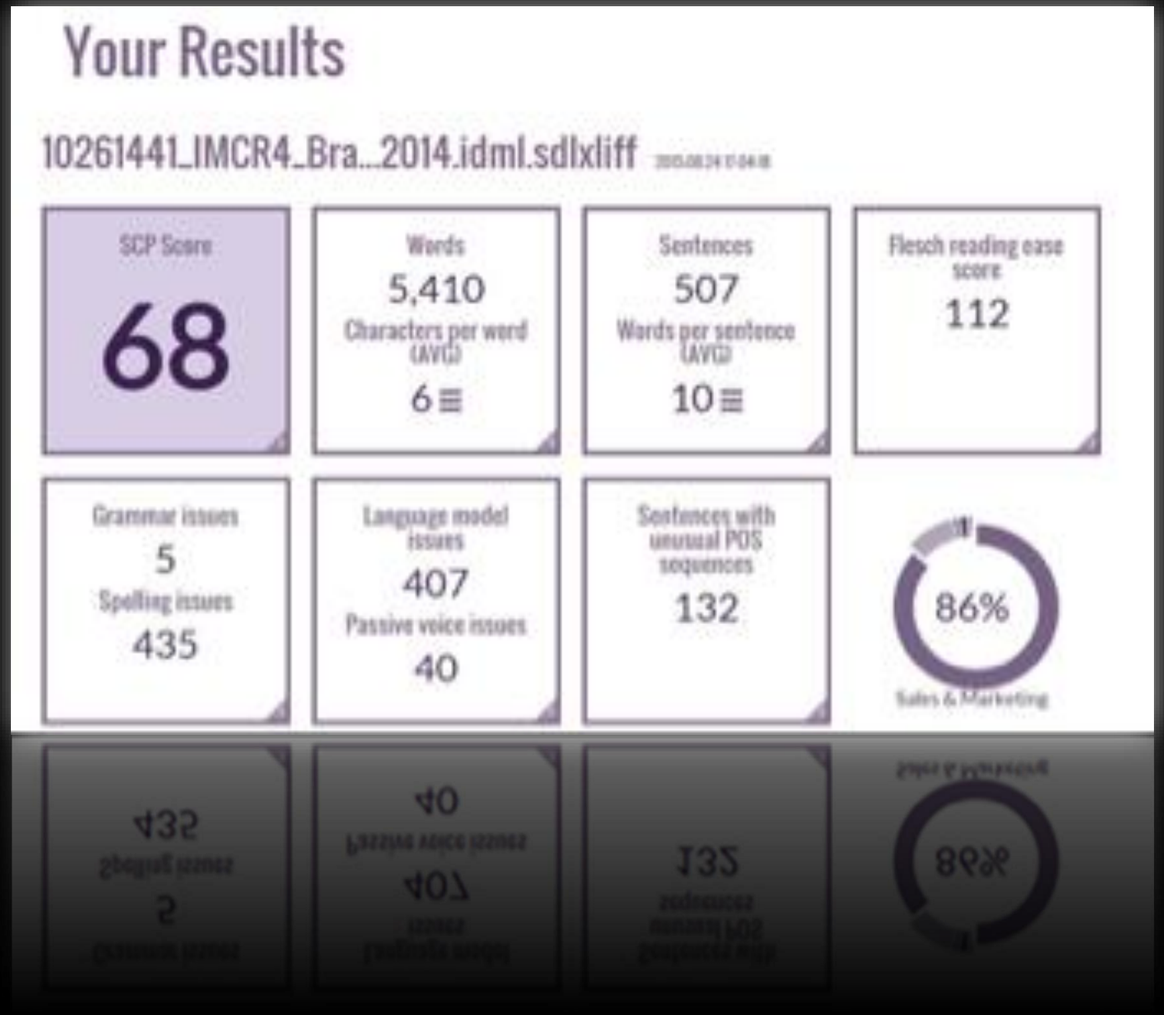
TOOLS: PERPLEXITY EVALUATOR

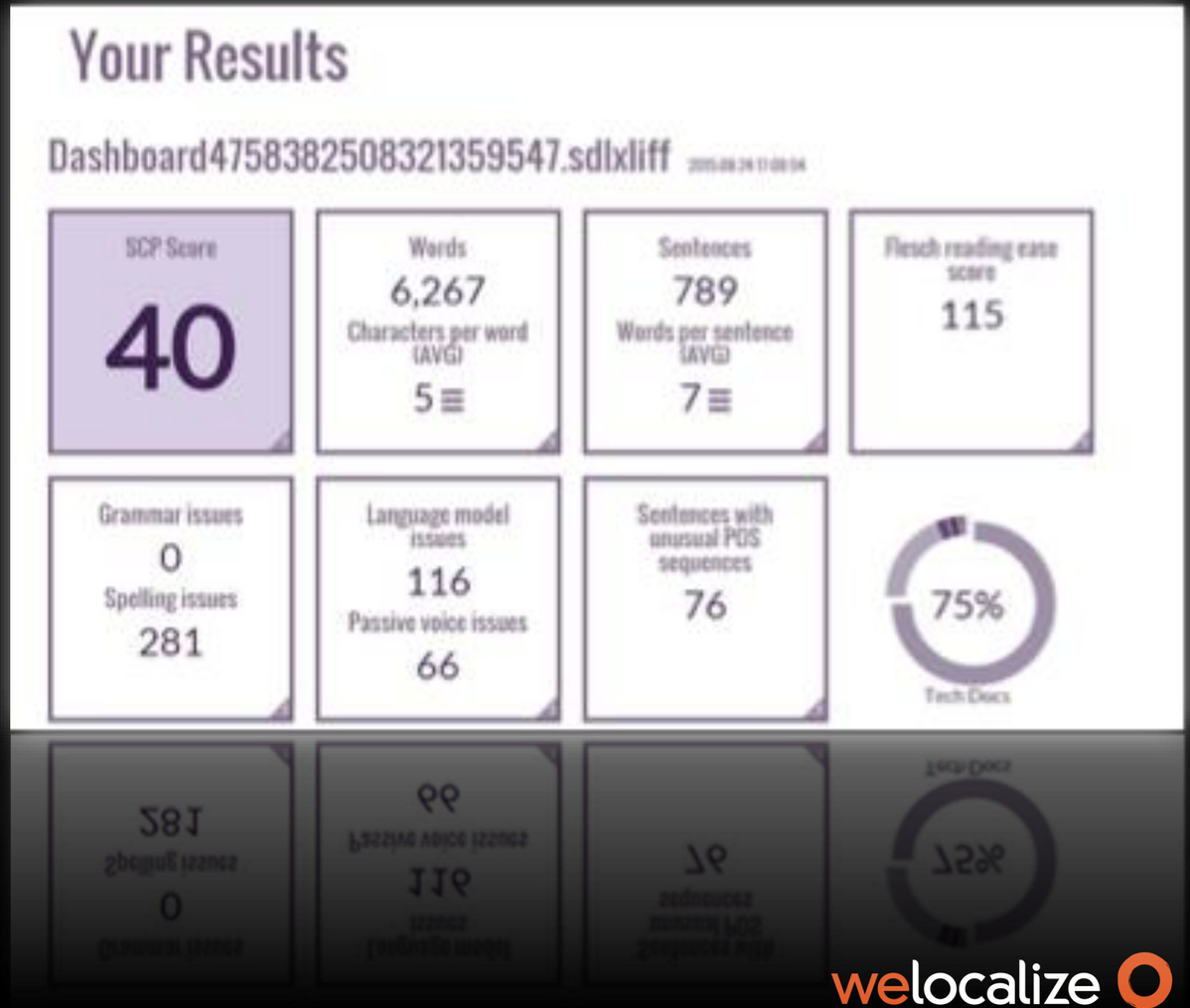
TU LEVEL

<tu srclang="EN-US" tuid="75438"> <prop type="x-ppl:train2">208</prop><prop type="x-ppl:techdoc6">191.025</prop><prop type="x-ppl:support2">325.983</prop><prop type="x-ppl:sales1">97.0736</prop><prop type="x-ppl:productLoc1">396.398</prop><prop type="x-ppl:legal1">617.876</prop><tuv xml:lang="EN-US"> <seg>Consistent feature set across multiple platforms (Windows, Mac, iOS, Android).</seg> </tuv> <tuv changedate="20140325T122530Z" changeid="serviceaaa" creationdate="20140325T122530Z" creationid="serviceaaa" lastusagedate="20140325T122530Z" usagecount="0" xml:lang="ES-XL"> <prop type="x-ALS:Context">TEXT</prop> <prop type="x-ALS:Source File">\\DATA\TC\39720\SRC\EN-US\co-02__battle-card_en\co-02__battle-card_en.inx</prop> <seg>Conjunto de características coherente en varias plataformas (Windows, Mac, iOS, Android)</seg> </tuv> </tu>

welocalize
doing things differently

TOOLS: SOURCE CONTENT PROFILER

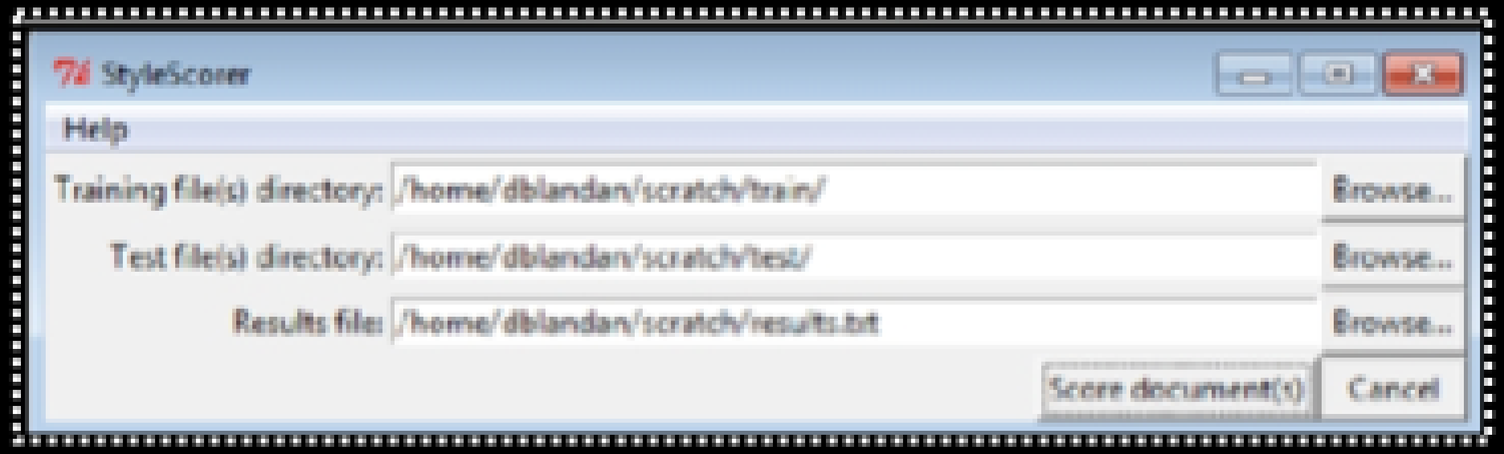




welocalize
doing things differently

TOOLS: STYLE SCORER

COMBINES PPL RATIOS, DISSIMILARITY SCORE + CLASSIFICATION SCORE



TEST CATEGORY	TRAINING CATEGORY	SCORE
SUPPORT	TECH DOC	3.16
TECH DOC	TECH DOC	2.94
TECH DOC	LEGAL	,02

welocalize
doing things differently

WHY USE STYLE SCORER?

- Identify similarity of source document to “gold standard” documents from that domain and other domains
- Identify similarity of target document to “gold standard” documents from that domain and other domains
- Example: Is this really a support document? To what degree is it similar to other support documents, tech doc documents, etc.?
- Dissimilarity can point to worse quality for raw MT and/or reduced post-editing productivity

STYLE SCORER + SCP

- SCP Helps Classify a Document
- Style Scorer Tells You How Good a Match a Document is to a Profile
- SCP Only Works on English Source
- Style Scorer Works on English Source + Non-English Target

CASE STUDY
ONE DOMAIN?

One ring
to rule them all

Three Rings for the Elven-kings under the sky,
Seven for the Dwarf-lords in their halls of stone,
Nine for Mortal Men doomed to die,
One for the Dark Lord on his dark throne
In the Land of Mordor where the Shadows lie.
One Ring to rule them all, One Ring to find them,
One Ring to bring them all and in the darkness bind them
In the Land of Mordor where the Shadows lie.

welocalize
doing things differently

CASE STUDY: HOW MANY DOMAINS?

- Started With 6 Domains: Technical Documentation, Legal, Support, Training, Product UI, Sales/Marketing
- Found that Technical Documentation, Support + Training Were Very Similar Based on LMs Scores Against Each Other, Length of Sentences, Similar Grammatical Structures
- Found that Product UI was Close Enough to Above 3 That Making a Separate Engine was Not Warranted
- Found that Legal + Sales/Marketing Were Different Enough from Above Domains and From Each Other Based on LMs Scores Against Each Other + Length of Sentences

CASE STUDY: GATHERING ASSETS

TMs

- Old
- Somewhat Recent
- Current

- Termbases in MultiTerm
- Existing User Dictionaries + Normalization Dictionaries
- New User Dictionaries Based on Term Extractions + Auto-Import for Some Languages

CASE STUDY: CURATING ASSETS

- Cleaned TMs
- Based on LM Perplexity
- Kept the UDs + Normalization Dictionaries As Is
- Additional term extraction for weak languages or languages with insufficient assets

CASE STUDY: ENGINE ITERATIONS

Based on options in Systran:

- RBMT only
- Hybrid with Stemming, LM Order, Distortion, etc.
- SMT only

HOW TO MEASURE SUCCESS

- Automatic scores
- Human evaluations
- Decrease in PE distance
- Decrease in linguistic issues reported

CASE STUDY: AUTOMATIC SCORES SALES/MARKETING1

	ar-SA		de-DE		es-ES		fr-CA		fr-FR		it-IT	
	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc
BLEU:	29.41	26.82	50.17	41.01	65.06	60.57	49.36	47.10	53.40	54.47	56.57	56.55
NIST:	6.89	6.53	9.23	7.87	11.10	10.64	9.37	9.11	9.76	9.83	10.17	10.10
METEOR:	14.69	13.98	60.07	53.43	79.66	76.44	64.80	62.98	67.41	68.11	70.06	70.04
GTM:	57.37	54.67	72.06	64.52	83.75	81.27	73.22	71.50	75.14	75.50	78.14	77.63
Avg. PE Dist.	29.27%	32.32%	29.07%	38.89%	22.56%	25.68%	27.51%	29.37%	28.87%	28.82%	24.85%	24.91%
TER:	54.09	57.68	38.61	49.53	24.08	27.26	36.53	38.76	34.51	34.17	30.34	30.99
Precision:	0.59	0.56	0.76	0.65	0.86	0.84	0.78	0.76	0.78	0.78	0.81	0.80
Recall:	0.56	0.53	0.69	0.64	0.81	0.79	0.69	0.68	0.72	0.73	0.75	0.75
Length (Mean Ref./Cand. Len.)	0.95	0.95	0.91	0.98	0.94	0.94	0.89	0.90	0.92	0.93	0.92	0.95
Sample size (Segments):	999	999	999	999	999	999	999	999	999	999	999	999
(Target Words):	9808	9808	12670	12670	12063	12063	11836	11836	11956	11956	11927	11927

CASE STUDY: AUTOMATIC SCORES SALES/MARKETING2

	ja-JP		ko-KR		pl-PL		pt-BR		ru-RU		zh-CN	
	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc	SalesMktg	Techdoc
BLEU:	62.04	45.96	59.05	44.23	52.38	30.98	64.38	53.56	53.42	43.86	55.01	49.39
NIST:	10.10	7.92	9.84	7.95	9.37	6.68	10.97	9.81	9.52	8.30	10.01	9.33
METEOR:	71.79	58.09	70.63	58.28	65.22	44.20	76.55	68.62	66.69	58.16	69.23	64.54
GTM:	78.99	68.00	77.60	67.42	74.07	56.02	83.30	77.21	74.04	66.07	77.99	74.23
Avg. PE Dist.	40.48%	56.25%	39.48%	53.61%	26.97%	45.44%	18.20%	24.82%	27.62%	35.72%	36.09%	42.32%
TER:	33.64	48.17	34.79	48.76	35.33	54.26	23.31	31.40	35.13	43.67	33.85	38.73
Precision:	0.84	0.73	0.81	0.69	0.79	0.59	0.86	0.79	0.78	0.69	0.83	0.79
Recall:	0.75	0.64	0.74	0.66	0.70	0.53	0.81	0.76	0.71	0.63	0.74	0.70
Length (Mean Ref./Cand. Len.)	0.89	0.88	0.91	0.94	0.89	0.91	0.94	0.96	0.91	0.92	0.89	0.90
Sample size (Segments):	991	991	931	931	999	999	999	999	999	999	999	999
(Target Words):	11951	11951	10834	10834	11153	11153	11776	11776	11770	11770	11542	11542

CASE STUDY: AUTOMATIC SCORES LEGAL1

	ar-SA		de-DE		es-ES		fr-CA		fr-FR		it-IT	
	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc
BLEU:	46.90	30.73	48.57	32.59	62.58	46.10	56.79	38.45	61.24	38.04	56.76	48.12
NIST:	8.94	7.03	9.04	6.90	10.77	9.03	10.15	8.07	10.49	7.99	10.09	9.10
METEOR:	60.90	45.67	59.12	46.80	77.41	65.67	70.19	56.27	72.95	55.52	69.88	62.75
GTM:	70.81	59.47	71.04	58.53	81.97	72.50	77.28	65.59	79.07	65.14	77.53	71.65
Avg. PE Dist.	28.00%	42.63%	30.90%	46.05%	20.36%	30.69%	25.42%	37.86%	27.87%	42.46%	24.64%	31.90%
TER:	40.69	54.53	41.48	57.82	26.04	37.57	31.60	45.96	29.39	46.69	30.68	38.11
Precision:	0.74	0.63	0.74	0.59	0.86	0.75	0.81	0.69	0.82	0.67	0.80	0.73
Recall:	0.68	0.57	0.68	0.58	0.79	0.70	0.74	0.63	0.77	0.63	0.75	0.70
Length (Mean Ref./Cand. Len.)	0.92	0.91	0.93	0.99	0.92	0.94	0.91	0.91	0.94	0.93	0.93	0.95
Sample size (Segments):	1000	1000	999	999	999	999	999	999	999	999	999	999
(Target Words):	9412	9412	10648	10648	10845	10845	11893	11893	10835	10835	10665	10665

CASE STUDY: AUTOMATIC SCORES LEGAL2

	ja-JP		ko-KR		pl-PL		pt-BR		ru-RU		zh-CN	
	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc	Legal	Techdoc
BLEU:	55.99	41.51	57.62	37.40	50.78	25.04	59.06	45.89	46.77	29.94	65.13	43.77
NIST:	9.26	7.38	9.60	7.03	9.21	5.93	10.35	8.99	8.66	6.55	10.92	8.64
METEOR:	67.02	54.00	69.28	52.16	64.30	38.05	72.31	62.13	60.51	45.18	76.15	59.33
GTM:	75.52	64.79	76.39	62.03	73.61	51.58	79.94	72.31	68.78	55.18	82.09	69.72
Avg. PE Dist.	43.46%	61.59%	38.14%	60.56%	26.83%	50.97%	22.10%	30.00%	35.02%	49.37%	28.73%	44.97%
TER:	38.14	52.43	36.09	55.43	36.15	59.89	27.62	37.03	41.44	56.23	28.22	44.44
Precision:	0.82	0.69	0.81	0.63	0.78	0.54	0.82	0.74	0.73	0.58	0.84	0.73
Recall:	0.70	0.61	0.73	0.61	0.70	0.49	0.78	0.71	0.65	0.53	0.80	0.67
Length (Mean Ref./Cand. Len.)	0.85	0.88	0.90	0.96	0.89	0.90	0.95	0.96	0.89	0.90	0.95	0.91
Sample size (Segments):	966	966	999	999	999	999	999	999	999	999	998	998
(Target Words):	10469	10469	10119	10119	9485	9485	10601	10601	10805	10805	9073	9073

HOW TO IMPROVE

OPPORTUNITIES FOR RESEARCH

- Eradicate High-Frequency Inconsistencies Between TMs, Termbases + User Dictionaries (UDs)
- Create Domain-Specific UDs
- Pre-MT Source Check: Was This Content Properly Categorized?
- Send Best Reply: TMT Prime, Send Best Translation Irrespective of Domain

SUMMARY

- Domain-specific Engines Yield Better Results as Evidenced by Auto Scores, Human Evaluations and Reduced PE Distance
- Group Closely-related Content into One Domain
- Determine How Many Engines Your Infrastructure Can Support

ALEXYANISHEVSKY
Welocalize
October 2015

Productivity Promotion Strategies for Collaborative Translation on Huge-Volume Technical Documents

Guiping Zhang zgp@ge-soft.com
Na Ye yena_1@126.com
Human Computer Intelligence Research Center, Shenyang Aerospace University, Shenyang, 110136, China
Fang Cai fangcai@berkeley.edu
Department of Statistics, UC Berkeley, Berkeley, 94704, U.S.
Chuang Wu wuchuang@ge-soft.com
Human Computer Intelligence Research Center, Shenyang Aerospace University, Shenyang, 110136, China
Xiangkui Sun sunxk@ge-soft.com
Jinfu Yuan yuanjf@ge-soft.com
Shenyang Global Envoy Software Co., Ltd., Shenyang, 110136, China
Dongfeng Cai caidf@vip.163.com
Human Computer Intelligence Research Center, Shenyang Aerospace University, Shenyang, 110136, China

Abstract

Automatic machine translation systems are seen unable to produce publishable quality translation, so various computer-assisted translation systems that emphasize human-machine cooperation have been proposed. However, translator collaboration technologies are underdeveloped, an area of great importance for large volume translation tasks. Ideally, all human translation knowledge is shared among translators in order to maximize productivity. In a knowledge engineering manner, our collaborative translation platform collects translation knowledge and actively pushes in real time. The mutual learning between translators and machine simultaneously builds the knowledge base and improves translators' proficiency. This paper introduces the collaboration strategies used in our platform that not only promote productivity but also ensure the translation quality. Comparative experiments by 36 professional translators prove the effectiveness of our collaboration strategies. A sounding result is that 22 professional translators completed a 97,000 page Chinese-English technical manual translation task within 42 months.

1. Introduction

With the advent of big-data era, the amount of technical documents (patents, standards, specifications, manuals) that need translation to different languages explosively increases. Huge-volume technical document translation suddenly became a bottleneck for the globalization of technology. Human translation is inefficient, whereas machine translation (MT) outputs are far from being satisfactory. Recently, the computer-assisted translation (CAT) technology aiming at improving the human translation productivity achieved great progress. The most

popular two CAT modes are post-editing (PE) and interactive machine translation (IMT). For huge-volume technical document translation, however, the core issues are still unresolved.

Besides all the problems in traditional translation tasks, there are three additional challenges particular for huge-volume technical document translation tasks. First, high-volume means that the task requires many professional translators collaborating, so progress management and knowledge sharing technologies play essential roles and can fundamentally affect the overall speed. Second, when there is more than one translator, it is hard to enforce consistent word choices and consistent sentence structure within or across documents. Technical manual normally requires translation of at least publication level, where details like consistent word choice and sentence structure are required. Finally, technical documents require highly specialized knowledge during translation, like technical term knowledge and relevant technical reference knowledge. Without special design, the cost on terminology looking-up by itself will fail our task.

Attempting to deal with all these challenges, our collaborative machine translation platform/pipeline incorporates a new thought of the integration of knowledge management and machine translation, which centralizes on a user model. Section 3 and 4 describe the thinking, design and realization of the platform.

In the rest of this paper, we select several strategies adapted in our platform tackling two issues: speed and quality. Before starting translation, the high frequency terms and sentences are pre-translated to ensure the accuracy and consistency of important technical concept translations and reduce translation difficulty. During the process of collaborative translation, the reliability of each fragment in the reference translation is color-encoded according to the source of its reference material, so as to help the translators make decisions rapidly. The translations by other translators on the same or similar sentences are pushed in real time, enabling the whole team to share the results. The translators' progress ranking is displayed, informing them of the team progress knowledge and encouraging them to speed up. Automatic proofreading tool is provided to help translators quickly verify their translation. Synchronous quality checking is adopted to control the translation quality in time.

Comparative experiments in section 5 show that these strategies can effectively improve the translation productivity while maintaining high quality. With these strategies, 22 professional translators accomplished a 97,000 page technical manual translation task within 42 months (each translator worked for 19.4 months on average). The quality requirement is higher than publishable level.

2. Related Work

More than thirty years ago, Kay (1980) proposed the idea of integrating machine translation and other assistant tools into human translation work (finally published in 1997). And it is predicted that the enhancement of such a system will finally lead people to achieve the goal of machine translation. With the continuous progress of the technologies such as machine translation, information retrieval and knowledge management, human-machine synergetic translation has replaced the traditional human translation mode and evolved several new modes.

Translation memory (TM) is the language processing technology which is earliest adopted in the translation process. Up till now, many professional translators still work by retrieving translations of similar fragments in the TM base. With the rapid development of statistical machine translation (SMT) technology, performing post-editing on SMT output becomes a new translation mode. It has been proved that both TM and PE can improve the translation productivity and quality (Mandreoli et al., 2006; Garcia, 2011; Arenas, 2014). Another pilot translation mode is the interactive-predictive machine trans-lation (Barrachina et al., 2009;

Sanchis-Trilles et al., 2014), in which the human gives the longest correct prefix of the translation and the system accordingly performs new decoding. The above research mainly focuses on how to improve the translation performance of an individual translator.

In recent years, how to achieve highly efficient and high quality collaborative translation among multiple translators became a new interest. Some researchers studied the methods of having Internet users to perform crowdsourced translation (Zaidan and Callison-Burch, 2011; Yan et al., 2014), having community members to perform community post-editing on the user generated content (Mitchell et al., 2014), or having monolingual users cooperate to translate (Hu et al., 2010). These studies focus on non-professional translators or even non-bilingual users, and aim at making the quality of translation achieve comprehensible level or specialized level. But publishable-level translation task is still difficult to accomplish.

In terms of large-scale collaborative translation among professional translators, the most relevant work is that of Karamanis (2011). The localization practice in two Language Service Providers is thoroughly investigated. The translator team's activities of manually establishing terminology glossary (Esselink, 2003; Wittner and Goldschmidt, 2007), searching the TM, sending emails and constant messages, and talking with other team members to communicate and share translation results are introduced. In this paper, we further developed these spontaneous and naïve collaboration activities. Automatic analysis tools are used to fully mine the important terms and fragments in the whole translation task, allowing the platform to actively share the translation results and team progress in real time. Besides, the translation quality is controlled more timely through automatic proofreading and synchronous quality checking. These strategies help the translators to better understand the translation task, the team decisions and progresses, so that they can accomplish precise and consistent translation more rapidly.

3. Collaborative Translation Practice

3.1. Project Background

In 2010, we started a 97,000 page publication-level Chinese-English technical manual translation project. A project team consisting of a translating group, a quality checking group, a R&D group, and a technology storming group was formed.

The members of the translating team are all full-time professional translators. The members of the quality checking team are all full-time professional and experienced translators. They are paid by the amount of translations that meet the quality requirement. At the beginning, the R&D team mines the requirement and configures a series of systems and tools that support the translation. Then they continuously receive feedback from translators during the collaborative translation process, rapidly develop new functions and perform small scale trials. If a new function is satisfactory, then it will be applied in the platform.

We made system developers and translators sit next to each other, so that translators can keep communicating with the technicians and the technicians can watch the real translating scenario to improve the platform in time.

3.2. The Collaborative Translation Process

High-volume technical document translation is a well-known difficult task. Our approach is to break down large pieces of work into smaller, simplified and more manageable parts. On the basis of the collaborative translation platform, we built a translation pipeline consisting of 3 main stages: pre-translation analysis stage, translation stage and post-translation management stage. Before translation, deep and fragmented analysis is performed. During translation, mul-

ti-dimension knowledge view, multi-aspect translation collaboration, multi-channel knowledge pushing and multi-layer quality controlling are provided. After translation, fine-grained management is performed. In this way, the pipeline decomposes the difficulties in the source texts and refines the translation step by step, thus achieving the effect of mutual knowledge increment between human and machine. The overall collaborative translation process is illustrated in Figure 1.

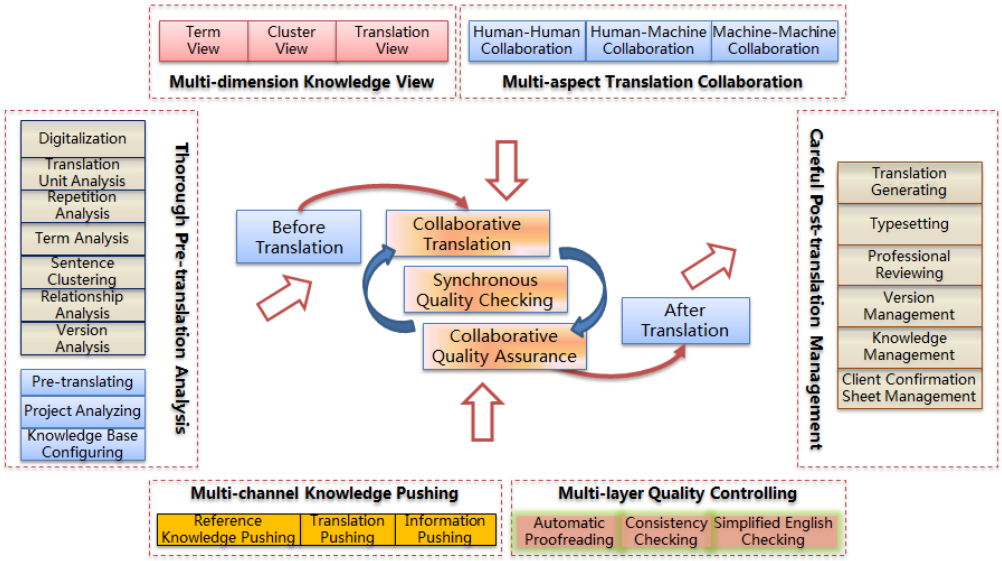


Figure 1. Overview of the collaborative translation process.

In the above figure, during the pre-translation stage, translation unit analysis is to split the source text in the manuals into basic translation units such as paragraphs and sentences. In this project, we take sentences as the basic units. Sentence clustering is to cluster sentences with similar contents. The clustering results are used for extracting translation templates and checking sentence-level consistency. In this project, sentences are clustered with a complete-linkage hierarchical clustering algorithm. Cosine distance is used to measure word-level similarity. Version analysis is designed to deal with the frequent changes in the document contents caused by the progress of technologies and the update of products. The differences among different versions are identified to avoid unnecessary repetitive work. Project analysis involves personnel recommendation, cost estimation and progress estimation.

During the translation stage, information pushing involves displaying the current translator's speed, his/her progress on the current document and all translators' progress ranking. Term view is for listing all the translation units that contain a certain term and their translations. It is designed for integrative viewing of the term translations. Clustering view is for listing all the similar translation units and their translations. It is designed for integrative viewing of the translations of similar units.

4. Platform Architecture

The work of this paper is based on a large collaborative translation platform. The platform includes six layers, namely knowledge layer, basic tool layer, interface layer, system layer, application layer and cloud service layer.

(1) The knowledge layer stores and manages the linguistic knowledge for translation such as terminology, bilingual sentences, rules and templates, process knowledge (e.g. translation history, quality checking errors and experience exchanges of translators) and domain knowledge (e.g. relevant technical references and term definitions).

(2) The basic tool layer provides the basic component set, including functional tools (such as data storage, network communication and data encryption), language analysis tools (such as lexical analysis, chunk analysis, parsing, text similarity computation and clustering), collaborative translation tools (such as machine translation, translation memory and translator activity recording) and knowledge management tools (such as knowledge collecting, accumulating, main-taining and sharing).

(3) The interface layer uniformly packages the tools of the previous layer. Popular network communication interfaces are provided and popular protocols such as HTTP, RESTful, SOAP and CMIS are supported to enable distributed management and concurrent access to the basic tools of the platform.

(4) The system layer provides all kinds of assistant systems for translators, including task management system, collaborative translation system, collaborative quality checking system, TM retrieval system, term management system and resource management system.

(5) The application layer configures the sys-tems according to the task requirement, and also realizes other applications such as translation data mining and pushing, enterprise-customized translation project management, translation skill teaching and crowdsourced translation.

(6) The cloud service layer makes use of the cloud computing and cloud security technologies to provide cloud-based translation service, online trading service and translator training service, finally achieving the goal of multiple translator collaboration under the cloud environment.

It is hard to describe every single technology used in our collaborative translation pipeline in one paper. In the next section, we will introduce several novel strategies for increasing translation productivity in the high-volume technical document translation context. As far as we believe, these strategies can be used in general large-scale translation situations. Of course, these strategies are far from being comprehensive. All the proposed strategies are implemented under the condition of ensuring quality. That is to say, if the translation cannot meet the quality requirement, then it will be returned to the translator for revision before it can be included in the productivity calculation.

5. Productivity Promotion Strategies

5.1. Pre-translating

Before starting translation, the technical terms in all the input documents are identified automatically. Since our practice is on a Chinese-English translation task, we trained a Conditional Random Fields (CRFs) model using 2000 manually labeled sentences for each domain to extract Chinese terms. The features are the context (word and part-of-speech) within a 3-word sized window. Experimental results on 568 documents show that the precision of Chinese term recognition is 75.06% and the recall is 79.30%. Then the frequencies of the terms in the whole translation task are counted and the terms are ranked according to the frequency. Table 1 gives some examples.

Term	Frequency
连接件(connector)	1559
制冷组件(cooling component)	1519
混合装置(mixing equipment)	1330
高压分离器(high pressure separator)	1220
…	…

Table 1: Examples of term analysis result.

The frequencies of the sentences in the whole task are also counted. The high frequency terms and sentences are considered to be important technical concepts and fragments. They are given to human experts to translate. And the corresponding fragments in the source texts are replaced with the decided translations. During the process of collaborative translation, any revision on these translations is prohibited.

To verify the influence of pre-translating on productivity, we divided 30 translators into two groups[1]. Each group has 3 teams, and each team has 5 members. A document of 10,000 characters is offered for translation. The teams in group A evenly split the document into 5 pieces and each member translates 2,000 characters. The high frequency terms/sentences are translated individually and review together after translation. The teams in group B perform pre-translating at first, and then evenly split the document for individual translation and review after translation. The average translation time and reviewing time are compared[2]. Table 2 shows the results (in minutes).

	Group A	Group B
Translation Time	241	282
Reviewing Time	182	70
Overall	423	352

Table 2: Comparative result of the pre-translating strategy.

It can be seen that the pre-translating of high frequency terms/sentences increased the translation time of group B, but greatly reduced the reviewing time. Therefore the overall time is less. For large scale translation tasks, pre-translating needs to be done only one time before starting translation, and will consequently save much more time. In terms of quality, pre-translating ensures that the translations of important concepts and fragments are highly consistent.

5.2. Translation Reliability Marking

In our human-machine interactive translation interface, a reference translation is provided for translators. Generally, a phrase translation model and a reordering model are both adopted in the phrase-based SMT systems. This brings about a mixture of phrase translation errors and reordering errors in the SMT output as illustrated in Figure 2.

[1] While dividing translators, we considered their translation capabilities and tried our best to divide evenly. Section 5.2-5.6 has the same consideration.

[2] When the translator needs to stop temporarily, he/she can click to stop the timing and click to continue when he/she starts again. Section 5.2-5.6 has the same setting.

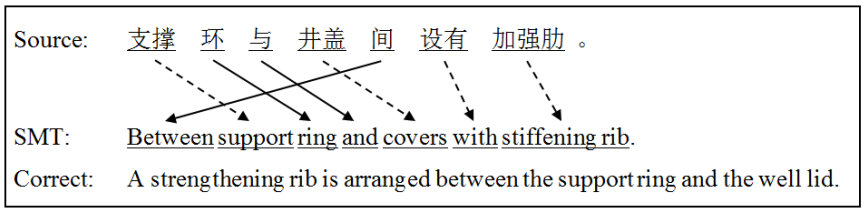


Figure 2. Mixed types of errors in the SMT output.

In the above example, it is relatively easy for the translators to judge and correct the phrase translation errors denoted with dashed lines. But the confusing reordering results of SMT may disturb the translators' train of thought. Discussions with 20 translators show that they need deeper analysis to identify reordering errors. And if the phrase alignment is labeled by arrows as in Figure 2, the reference translations will become too chaotic, especially for long sentences. Due to the above reasons, the reference translations are given in the monotone format as shown in Figure 3. The phrase translations are output in their original order as in the source sentence.

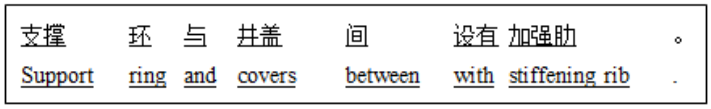


Figure 3. Example of monotone reference translation.

The phrase translations are marked with different colors to indicate their reliabilities. Figure 4 gives an example.

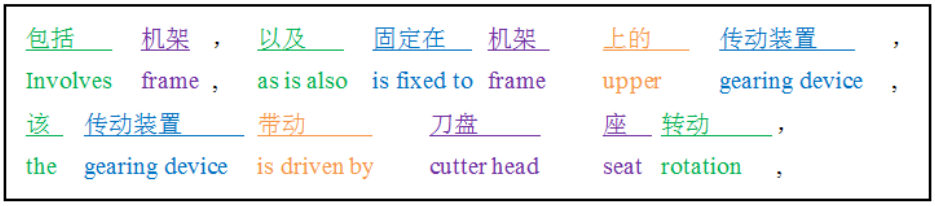


Figure 4. Example of translation reliability marking.

Purple font indicates that the translation comes from relevant reference X. Blue font indicates that the translation comes from relevant reference Y. Green font indicates that this is a translation used by other translators[3]. Orange font indicates this is a machine translation result.

Marking translation reliability can influence the translation productivity. We prepared a document of 1,000 characters and divided the translators into two groups, 18 in each group. Every translator is asked to translate the whole document. The reference translations in the interaction windows of group B are marked with reliability color.

Experimental results show that the average translation time of group A is 125 minutes, and that of group B is 111 minutes. The reliability color helps the translators to know the

[3] During post-editing, when a translator needs to revise the current translation of a phrase, he/she can right-click it, then a menu containing other options will pop up and he/she can left-click the correct one to accept it. These activities are recorded by the platform. And the option with the highest frequency of being left-clicked is displayed in the next time.

sources or reasons for the reference translations and make decisions easier, thereby increased the productivity.

5.3. Translation Pushing

After a translator completes a sentence, his/her translation is pushed to the same sentences waiting for translation in the whole task (directly replaced in the translation task window and labeled with its original translator) in real time to avoid translating the same content. Translation pushing includes two types. One is complete matched pushing (for exactly the same sentences), the other is fuzzy matched pushing (for the sentences with minor differences such as letters and digitals). In the latter type, the different parts are automatically revised. For example, when a translator completes the sentence "工作状态" (Operating Condition), all the sentences "CPU工作状态" in the remaining tasks will be automatically replaced with "CPU Operating Condition". After that, when the other translators face the pushed translation, he/she can accept it or revise it according to the context. If a translator finds that the pushed translation is wrong or problematic, then he/she can also tell its original translator or discuss with him/her to find out the best decision.

We prepared a document of 1,000 characters containing repetitive and similar sentences and divided the translators into two groups (18 in each). Every translator is asked to translate the whole document. Group B is provided with the translation pushing function.

Experimental results show that the average translation time of group A is 141 minutes, while that of group B is 111 minutes. The speed of group B is 1.27 times as fast as group A. For the technical documents with strongly related content, translation pushing can solve the translation of many sentences, improve the consistency and help the translators to make decisions. Meanwhile, because the results from other translators can be seen, this strategy also partly realizes collaborative quality checking among translators.

5.4. Progress Ranking

In this strategy, the real-time translation progress ranking of translators are displayed above the interaction window, including the translators who translate the most and the second most in the current month and the translators who translate the most and the second most in the current week.

We give the same set of documents to two translator groups. Each group has 18 translators. The members of group B can see the ranking in real time. The translators' performance in a week (5 workdays) is observed. Table 3 gives the average speed (character per day) of each day.

	Group A	Group B	Improvement
Day 1	560.2	560.7	0.09%
Day 2	550.8	555.9	0.93%
Day 3	576.6	586.4	1.70%
Day 4	556.9	571.1	2.55%
Day 5	547.8	555.4	1.39%
Average	558.5	565.9	1.32%

Table 3: Comparative result of the progress ranking strategy.

Experimental results show that the average speed of group B is 1.32% higher than that of group A. Displaying the fastest translators can inform the translators of the team progress and every one tends to try his/her best to catch up with the others' progress.

5.5. Synchronous Quality Checking

In the traditional translation process, the quality checkers usually start working until the translators finishes their tasks. In contrast, the process of synchronous quality checking is having the quality checkers and translators work at the same time. When the translators start working, the quality checkers can immediately see the results and perform checking.

We prepared a document of 1,000 characters containing repetitive and similar sentences and divided the translators into two groups. Each group has 9 teams, and each team has 2 members (one translator and one quality checker). Group A follows the traditional process of checking after translation, and group B adopts synchronous quality checking.

Experimental results show that the average translation time of group A is 175 minutes, and that of group B is 119 minutes. The speed of group B is 1.47 times as fast as group A. The reason for the obvious improvement is due to two aspects. First, in this way the traditional sequential working process is transformed into parallel working. Second, the translators can get to know their mistakes as early as possible and solve them, therefore the translation quality and speed afterwards are ensured.

5.6. Automatic Proofreading

When a translator completes the current translation unit, an automatic proofreading tool works to check the frequently appeared grammatical mistakes in the translation, including capitalizations, articles, punctuations, missed translation, spelling errors and some usages prohibited by the specification of the task. Whenever a mistake is detected, the tool labels the corresponding part to alert the translator. Mistake detection is implemented by a rule-matching strategy, in which rules are written manually in the form of regular expressions.

We give the same set of documents to two groups of A and B. Each group has 18 translators. The members of group B are provided with the proofreading tool. The translators' performance in a week (5 workdays) is observed. Table 4 gives the average speed (character per day) of each day.

	Group A	Group B	Improvement
Day 1	971.4	985.2	1.42%
Day 2	950.1	990.1	4.21%
Day 3	975.2	1032.4	5.87%
Day 4	956.3	1025.3	7.22%
Day 5	950.7	1022.7	7.57%
Average	960.7	1011.1	5.25%

Table 4: Comparative result of the automatic proofreading strategy.

Experimental results show that the average speed of group B is 5.25% higher than that of group A. With the increase of time, the difference in speed increases continuously. Through talking with the translators, we find that after adding the automatic proofreading function, once the system doesn't find mistakes, the translator will submit his/her translation confidently, thus improving the productivity.

5.7. Results and Analysis

Through 42 months collaborative work among 22 translators, this huge-volume technical document translation task was accomplished. Each translator worked for 19.4 months on average. During this time, more than 20 translation specifications are established, covering all

aspects of the project from pre-translation analysis to post-translation management, from collaborative translation to collaborative quality assurance. Several million knowledge entries including bilingual sentence pairs, translation process logs, technical terms, proofreading knowledge, reference knowledge are accumulated.

The six strategies described in this paper played an important role in the accomplishment of the project. They realized the dynamic accumulation, real-time transformation and simultaneous increment of knowledge during the process of collaborative translation. Using this collaborative translation platform, the overall translation productivity increased by more than one time on this project.

In the stage of pre-analysis, the translation task is deeply understood as a whole. In the process of collaborative translation, the platform continuously accumulates the translation results of the whole team, and provides the translators with the newest translation knowledge, translation decisions, and the team progress knowledge in different ways with a fine-grained manner in real time, so that the translators can rapidly make the decisions. At the same time, the inner knowledge structure of the platform is also continuously optimized. The human and the machine make the most of their advantages and learn from each other to make common progress. With the increase of the collaborative translation time, the knowledge scale and the translation ability of both the translators and the platform are improved constantly.

6. Conclusion and Future Work

The translation of huge-volume technical documents is a task that requires multiple professional translators to collaborate. How to fully increase the productivity while maintaining high quality is a crucial problem. This paper proposed several strategies used in our translation pipeline to promote productivity, helping 22 professional translators to accomplish a 97,000 page Chinese-English publishable technical manual translation within 42 months. This paper also gave some clues to the translators' psychological activities and processes during collaborative translation, which help people to deepen the understanding of cognitive translation activity and psychology.

In the future, we will make use of the large amount of translation knowledge and quality checking knowledge to conduct collaborative translation on the same type of manuals. We will also study the assistant compilation technology of the same type of manuals, and the interactive interface customization technology for translators with different levels and different characteristics.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61402299).

References

Arenas, A G. (2014). Correlations between Productivity and Quality when Post-editing in a Professional Context. *Machine Translation*, 28(3-4): 165-186.

Barrachina S, Bender O, Casacuberta F, et al. (2009). Statistical Approaches to Computer-assisted Translation. *Computational Linguistics*, 35(1): 3-28.

Esselink B. (2003). Localisation and Translation. *Computers and Tanslation: a Tanslator's Guide*. John Benjamins, Amsterdam, pages 67-86.

Garcia I. (2011). Translating by Post-editing: is it the Way forward? *Machine Translation*, 25(3): 217-237.

Hu C, Bederson B B, Resnik P. (2010). Translation by Iterative Collaboration between Monolingual Users. In *Proceedings of Graphics Interface 2010*. Canadian Information Processing Society, pages 39-46.

Karamanis N, Luz S, Doherty G. (2011). Translation Practice in the Workplace: Contextual Analysis and Implications for Machine Translation. *Machine Translation*, 25(1): 35-52.

Kay M. (1997). The Proper Place of Men and Machines in Language Translation. *Machine Translation*, 12(1-2): 3-23.

Mandreoli F, Martoglia R, Tiberio P. (2006). EXTRA: a System for Example-based Translation Assistance. *Machine Translation*, 20(3): 167-197.

Mitchell L, O'Brien S, Roturier J. (2014). Quality Evaluation in Community Post-editing. *Machine Trans-lation*, 28(3-4): 237-262.

Sanchis-Trilles G, Alabau V, Buck C, et al. (2014). Interactive Translation Prediction versus Conventional Post-editing in Practice: a Study with the CasMaCat Workbench. *Machine Translation*, 28(3-4): 217-235.

Wittner J, Goldschmidt D. (2007). Technical Challenges and Localisation Tools. *Localisation Guide 2007*. Multilingual Computing Inc, Sandpoint, pages 10-14.

Yan R, Gao M, Pavlick E, et al. (2014). Are Two Heads Better than One? Crowdsourced Translation via a Two-Step Collaboration of Non-Professional Translators and Editors. In *Proceedings of the 52nd Annual Meeting of the Association of Computational Linguistics*, pages 1134-1144.

Zaidan O F, Callison-Burch C. (2011). Crowdsourcing Translation: Professional Quality from Non-Professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Lin-guistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 1220-1229.

A Machine Assisted Human Translation System for Technical Documents

Vishwajeet Kumar vishwajeetkumar86@gmail.com
Ashish Kulkarni kulashish@gmail.com
Pankaj Singh pr.pankajsingh@gmail.com
Ganesh Ramakrishnan ganesh@cse.iitb.ac.in
Department of Computer Science and Engineering, IIT Bombay, Mumbai, India

Ganesh Arnaal ganesh.arnaal@smcapitaladvisors.in
SM Capital Advisors,Mumbai,India

Abstract

Translation systems are known to benefit from the availability of a bilingual lexicon for a domain of interest. A system, aiming to build such a lexicon from source language corpus, often requires human assistance and is confronted by conflicting requirements of minimizing human translation effort while improving the translation quality. We present an approach that exploits redundancy in the source corpus and extracts recurring *patterns* which are : *frequent, syntactically well-formed, and provide maximum corpus coverage*. The patterns generalize over phrases and word types and our approach finds a succinct set of good patterns with high coverage. Our interactive system leverages these patterns in multiple iterations of translation and post-editing, thereby progressively generating a high quality bilingual lexicon.

1 Introduction

The problem of language translation has been in focus for many decades and has seen contributions from both linguistic and computer science communities. Linguistic contribution (Streiter (1996)) has come in the form of several language resources comprising of dictionaries, grammar and studies on units of translation. Computer science community has contributed in coming up with formal machine translation (MT) models (Vogel et al. (2003)) that leverage corpus statistics along with linguistic features and resources. There is a body of work (Federico et al. (2014); Alabau et al. (2014)) that studies the complementary contributions of humans and MT models and present "machine-centric" translation systems that leverage human input. These systems, referred to as computer aided translation (CAT) systems, typically employ a statistical MT model to translate text and provide a post-editing tooling to enable humans to correct the resulting translations. Human feedback and corrections are used to adapt and retrain the translation model. What constitutes the right unit of translation and how can the human feedback be incorporated in the underlying translation model, pose interesting research challenges.

A domain corpus is often replete with redundancy arising due to the choice of vocabulary and syntax. Translation memory-based systems (Sato and Nagao (1990)) exploit this redundancy and store recurring phrases and their translations. We are fur-

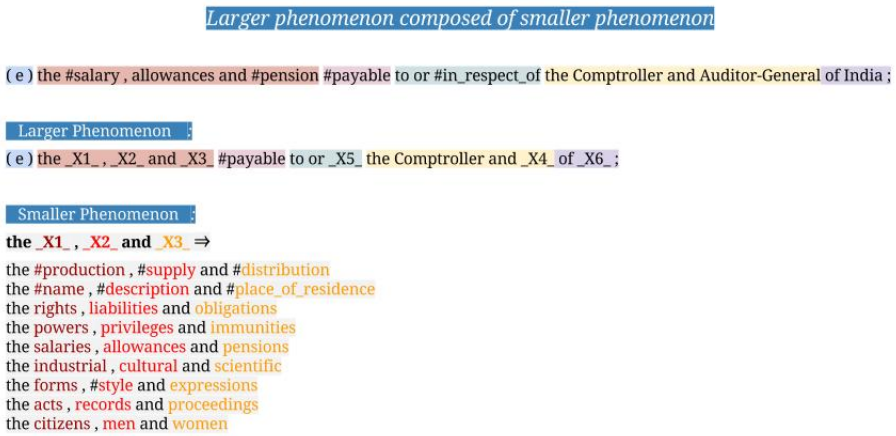


Figure 1: An example illustrating the principle of compositionality and higher order patterns in a domain corpus

ther motivated by Frege's principle of compositionality (Pelletier (1994)), which states that the meaning of a compound expression is a function of the meaning of its parts and of the syntactic rule by which they are combined. Figure 1 shows an example, taken from legal domain, of a compound expression and its constituent expressions. Some of these expressions comprise of categories that generalize over several tokens, thus, forming higher order recurring patterns in the corpus. Extraction of these patterns and using them as the unit of translation might enable us to better capture the structure and semantics of the domain.

An in-domain (especially technical, legal) corpus often adheres to a certain lexical and syntactic structure and is often less amenable to creative or "free" translation. These domains, therefore, might be good candidates for translation using rule-based systems Terumasa (2007), comprising of source and target language dictionaries, grammars and translation rules. Grammatical Framework (GF) (Ranta (2004)) provides the necessary formalism to theorize rule-based translations and also provides a system to author abstract and concrete language syntax.

We present an approach and a system that builds on these ideas to extract meaningful patterns from a domain corpus, gather human feedback on their translation and learn a rule-based translation system using the GF formalism. The system is "human-centric", in that, it heavily relies on manually curated linguistic resources, while the machine continuously prompts the human on *what* to translate. This interactive human-machine dialog produces a translation system that aims to achieve high precision in-domain translations and might find application in several technical domains including medical, education, legal etc.The system is available for demo at http://mtdemo.hostzi.com.

2 Related Work

There has been a lot of research on automated statistical machine translation (SMT) and several systems (Wang and Waibel (1998); Vogel et al. (2003); Och and Ney (2000);

Koehn et al. (2007)) have been proposed. While they are all typically based on a combination of a translation model and the target language model, the difference lies in their units of translation (word-based, phrase-based *etc.*) and translation decoding. The statistical approach to MT itself falls under the general category of example-based MT (EBMT) (Somers (1999)) or memory-based MT (Sato and Nagao (1990)). These approaches rely on the availability of a corpus or a database of already translated examples, and involve a process of matching a new input against this database to extract suitable examples and then determine the correct translation. These corpus-based approaches suffer from two major drawbacks - (1) parallel corpus is often expensive to generate and is often scarce or unavailable for certain language pairs or domains; (2) their quality of translation is not as good as that of human translation and therefore not suitable for certain applications like those involving translation of government documents or academic books.

Rule-based machine translation systems (RBMT) like Apertium (Forcada et al. (2011)) alleviate the need for a sentence aligned parallel corpus but require explicit linguistic data in the form of morphological and bilingual dictionaries, grammars and structural transfer rules. Apertium is a free and open-source machine translation platform with liguistic data for a growing number of language pairs along with the necessary tools and a translation engine. However, these systems typically involve a complex pipeline and statistical tools, making it difficult to track and correct errors.

Many researchers in the past have claimed and suggested that we cannot remove humans completely from the translation pipeline (Kay (1980)). In order to cater to applications requiring a high-quality translation, the output of MT systems is often revised by a *post-editing* phase. Several computer-aided translation (CAT) tools exist that are either desktop-based (Carl (2012); Aziz et al. (2012)), iOmegaT[1] or web-based (Federico et al. (2014); Denkowski and Lavie (2012); Roturier et al. (2013)). As an alternative to pure post-editing systems, interactive machine translation (IMT) (Toselli et al. (2011)) combines a MT engine with human, in an interactive setup, where, the MT engine continuously exploits human feedback and attempts to improve future translations. Daniel Ortiz-Martınez (2011); Ortiz-Martínez et al. (2010), for instance, talk about online learning in the machine translation pipeline, where, human feedback on translations is used to re-estimate the parameters of a statistical machine translation model. Bertoldi et al. (2013) address the problem of dynamically adapting a phrase-based SMT model from user post-editing by means of a caching mechanism. Their cache-based model combines a large global static model with a small local and dynamic model estimated from recent items in the input stream. Lavie (2014) incorporate human feedback and propose three online methods for improving an underlying MT engine based on translation grammar, Bayesian language model and parameter optimization. Anusaarka (Bharati et al. (2003)), a hybrid machine translation system for English to Hindi, also involves interaction but is restricted to authoring rules for word sense disambiguation. Ranta had proposed Grammatical Framework (GF) (Ranta (2004)) which is a grammar formalism and a programming language for multilingual grammar applications. One good example of applications using GF[2] is Molto (Cristina Espa˜na-Bonet (2011)), a machine translation system for patent translation.

While our approach builds on existing work, our primary contribution is a framework and a system for high quality domain corpus translation. Our system gathers manual translation of redundant patterns in an interactive setting and uses these to

[1] http://www.omegat.org/

[2] http://www.grammaticalframework.org/

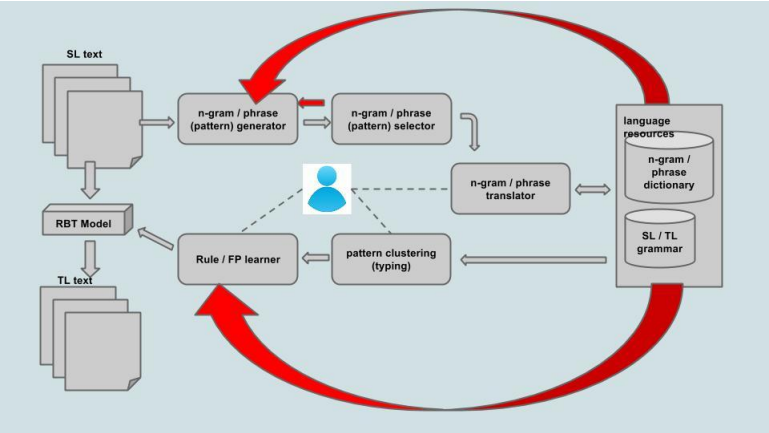


Figure 2: System Architecture

build language resources like grammars and bilingual lexicons. These are realized using the GF formalism and the translation system continues to benefit from more human feedback.

3 System Architecture

Our system follows an iterative pipeline architecture where every component is modular. The system is interactive and takes human feedback on translations. The feedback is used to build linguistic resources and is incorporated into the underlying translation model. The translation model itself is expressed using the grammatical framework formalism, which is based on functional programming and type theory. This expressivity and abstraction makes the model easily programmable by humans.

3.1 Pattern Extraction

This module captures redundant translation units present in the corpus. It takes as input a domain specific corpus and monolingual typed dictionaries and produces frequently occurring translation units as output. It uses frequent pattern mining technique to capture exhaustive set of frequent translation units. In order to extract more general translation units, we extract patterns with gaps where a gap might be of varying length. A gap could be considered as a generalized form of an entity and is represented as "X" or "_X_". The length of the gap controls the generalization. As output, the module produces a directed acyclic graph of frequent translation units in the corpus. Algorithm 1 contains details of our frequent pattern mining algorithm. The module also supports filtering of invalid translation units. An invalid translation unit is the one that does not honor pattern compositionality.

3.2 Pattern Selection

Pattern Extraction (Section 3.1) mines a large number of redundant patterns as potential translation units. Since getting manual translations for these candidate translation units is a costly operation, we identify a subset of patterns that are both diverse and maximally cover the in-domain source language corpus. The pattern selection algorithm (Refer Algorithm 2) provides details on this selection of a subset of "good" patterns, where, goodness of the subset is measured in terms of corpus

Algorithm 1: Algorithm: FPM algorithm

Data: Corpus C, Pattern length L, Frequency threshold T, Maximum consecutive gaps of tokens G
Result: Set F of frequent patterns
Maintain a dictionary structure globalPatternList where key is pattern and value is list of span
for *each sentence s in C* **do**
  maintain an array of list, slist, of size $|s|$, such that, $slist[i]$ stores all one length pattern along with its span in the sentence which starts from s_i
  using slist, create a 2D array of list, smatrix, of size $|s|xL$ such that, $smatrix[i][j]$ stores all patterns, along with its span in s, which starts from s_i and of pattern length j
  Filter pattern from smatrix whose span is syntactically incomplete
  Add these patterns to globalPatternList
end
Initialize patternWithGap dictionary
for *i in* $1 \cdots L$ **do**
  for *valid mask v of length L* **do**
    for *pattern p of length i in globalPatternList* **do**
      apply v on pattern p and create a new pattern $\hat{p}$
      if $\hat{p}$ *is present in patternWithGap* **then**
        update its spanlist by doing union with span list of p
      else
        add $\hat{p}$ in patternWithGap with its spanlist as spanlist of p
      end
    end
    remove patterns of length i and with gap position according to mask and having spans count less than T
  end
end
remove patterns from globalPatternList whose number of spans is below **T**
output $patternWithGap \cup globalPatternList$

coverage. Figure 3 provides an example, where, the first column contains sample text from a corpus and the other columns show the extracted patterns and the patterns (in bold) after the selection step.

3.3 Pattern Translator

Translator module involves users to provide translations of translation units. In this module five system generated translations are displayed to translator out of which he can select best translation for a particular translation unit or he can even write a new translation.

3.4 Generalization of Translation Units

At each iteration we identify important non-terminals present at that level and use this information while generating translation units at the next level. This module helps in generalizing translation units by clustering them together. This in turn helps in reducing

Sample input text	Sample patterns subset		
	on the expiration of the X period	the fixed period	one year
on the expiration of every second year in accordance	on the expiration of the said period	***the X period(5)***	**his term of office(4)**
on the expiration of every second year in accordance	on the expiration of the fixed period	a period of NP	his term
on the expiration of the said period	on the expiration of a period of ten years	a period of ten years	
on the expiration of the fixed period	on the expiration of every second year in accordance	a period	
on the expiration of his term of office	on the expiration of a period of six months	ten years	
on the expiration of his term of office	on the expiration of a period of one year	every second year	
on the expiration of a period of ten years	on the expiration of his term of office	***every second year in accordance(3)***	
on the expiration of a period of six months	***on the expiration of a period of NP (1)***	a period of six months	
on the expiration of a period of six months	***on the expiration(2)***	six months	
on the expiration of a period of six months	the said period	a period of one year	
on the expiration of a period of one year			

Figure 3: Example

the number of rules required to express compositionality. In terms of grammar, we can think of it as identifying LHS of productions. Arguably, this module must also serve the purpose of organizing non-terminals such that it is useful for translation task. Since we are identifying domain specific concepts (non-terminals) which can be translated, it must also keep the target language in mind.
We have observed in various sentences that if internal reordering[3] of phrases in sentences having same cannonical structure is same then their external reordering[4] also remains same. So we tried to cluster phrases having same internal reordering into one cluster. It is very clear from the objective of this module that clustering of translation units should be based on some translation in-variance phenomenon. Since the group represent all the translation units present in that group, it should also represent their translation behavior. Same external reordering help a category to generalize these translation units for higher level translation unit generation while same internal re-ordering will help in writing single translation rule for all the member translation units. We used reordering distortion score between translations of two translation units as a measure to cluster translation units.

3.5 Rule/FP Learner

Once patterns are extracted, selected, translated and stored in database, we annotate sentences with pattern name or in other words represent sentences in the form of sequence of translation units. If a sentence is completely covered by the set of patterns, it can be represented in terms of patterns. Once a sentence is represented in such a canonical form, we parse and linearize it using grammatical framework rules.
Idea of using functional programming and type theory in machine translation came from logical framework and ALF[5]. The logical framework ALF was based on the constructive type theory of Martin-Löf (Martin-Löf 1984, Nordström & al. 1990). Constructive type theory has also proven usable for meaning representation in natural languages (Ranta 1994). Logical frameworks were used to define logic in other perspectives but logic in machine translation means grammar. The type checking and proof search machinery provided by a logical framework like ALF gives tools for the kind of semantic analysis needed in machine translation. And here the missing component was parsing and linearization which was provided by Grammatical framework developed by Arne Ranta.

Grammatical framework is nothing but an extension of logical framework with a component called concrete syntax. Reordering rules and rules for handling gender, number and person information while doing look up is written using grammatical framework. The main purpose behind using grammatical framework is its functional nature. Gram-

[3]Reordering of tokens within a pattern during its translation from source to target language
[4]Reordering of a pattern within a sentence during the translation of that sentence
[5]ALF (Another Logical Framework) is a logical framework based on Martin-Lof type theory

Algorithm 2: Pattern Selection

Data: Dictionary of patterns P with its spanslist, Number of words in corpus N, Max size of selected set k
Result: Set F of diverse and high coverage (in terms of words) patterns

$F = \emptyset$
$bitCorpus \leftarrow \emptyset$
for $i \leftarrow 1$ ***to*** N **do**
  $bitCorpus[i] \leftarrow false$
end
for $i \leftarrow 1$ ***to*** k **do**
  $currentBest \leftarrow NULL$
  $currentBestCoverage \leftarrow 0$
  for *each pattern* p *in* $P \setminus F$ **do**
    $coverage_p \leftarrow 0$
    $coverage_p \leftarrow$ count of false bits in bitCorpus which is in spanlist of p
    if $coverage_p > currentBestCoverage$ **then**
      $currentBest = p$
      $currentBestCoverage = coverage_p$
    end
  end
  if $currentBest$ **then**
    $F \leftarrow F \cup$currentBest)
    set $BitCorpus[i]$ = true if i is in the spanlist of currentBest
  else
    break
  end
end
output F

matical framework also has a concept called abstract syntax which provides interlingua representation. Interlingua representation helps in linearizing in different languages very easily just by writing concrete grammar for that language.

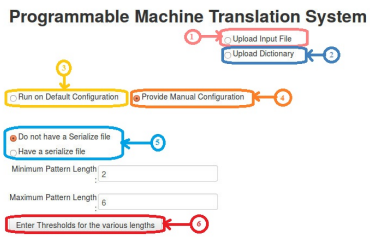


Figure 4: Interactive user interface for providing parameters to Frequent Pattern Miner

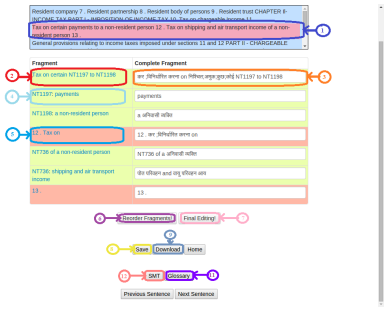


Figure 5: Interactive user interface for humans to translate patterns and n-grams.

3.6 System User Interface

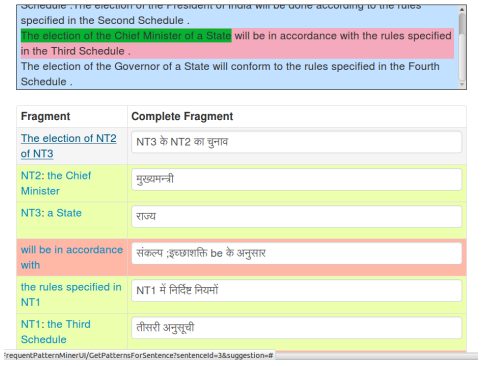
Our system has a highly interactive user interface for humans to translate patterns and n-grams. It also has provision for expert users to configure pattern length and frequency threshold for pattern extraction. Figure 4 depicts the features provided to expert users. Users can upload a new corpus using the *Upload Input File* option marked with label 1 in the figure. The *Upload Dictionary* option (labeled 2) enables users to upload bilingual dictionaries for the system to perform lookups and provide translation suggestions. Users can either choose to run the system and extract patterns on the optimized default configuration (labeled 3) or they can manually configure the pattern length and frequency (labeled 4).

Once patterns are extracted, filtered and validated by the system, users use the web-based system shown in the Figure 5 for providing translation feedback. Human translators are shown the current sentence (labeled 1) along with the previous and next sentences as context information. Patterns are displayed below column labeled *fragment* (label 2). On hover over patterns or untranslated n-grams, the span covering that pattern or n-gram in the sentence gets highlighted (refer to figure 6a). For patterns containing generalized non terminals (labeled 2), translators can view all the instances of non terminals by hovering over the NTs. Instance of a non terminal is represented by label 4 in Figure 5. Initially a translation of patterns and untranslated n-grams (labeled 5) is suggested by the system using translated patterns database, glossary look-up and SMT. Translators can even configure the source for getting the suggestion (a) they can choose to get translation suggestion from SMT system by clicking on SMT button (labeled 12) or (b) they can choose to get translation suggestions from database by clicking on glossary button (labeled 11). Translators can edit the translation suggestion (labeled 3) given by the system and correct them. They can also reorder the composed translation of sentence by clicking on reorder button (labeled 6), which presents a simple drag and drop interface to the user (refer to Figure 6b). Finally, if user wish to edit the composed translation they can do that by clicking on final editing button depicted by label 7 in Figure 5. After final editing, users can save the translation by clicking on save button (labeled 8). Users can also download the translations by clicking on the *download* button (labeled 9). In order to get translation suggestions for a particular word or phrase, users can enter the text in *suggestions* panel on the right and get multiple translation suggestions for the particular word or phrase.
Important Features of the system:-

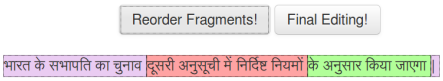
- Once a translator translates a pattern, a pattern instance or an n-gram, the system auto-translates it if next time it appears in a sentence.
- If a pattern, pattern instance or n-gram is translated differently in different sentences, the system lists all of them as choices for the user to choose from or enter a new translation.
- The system also has an integrated suggestion component that fetches translation suggestions from various sources. Users can use this to get translation suggestions for words or phrases and choose the best translation from the choices.

4 Evaluation

We evaluate the system in terms of the quality of extracted patterns, GF grammar and system efficiency. Evaluation was done on five public datasets *viz.* the Constitution of

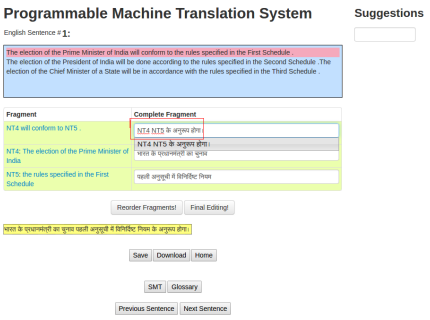


(a) On hover over patterns the part of sentence covering the pattern gets highlighted

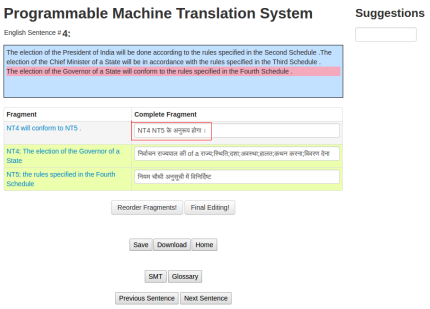


(b) Reordering composed translation of sentence

(c) Final editing



(d) Pattern translated by human highlighted in red rectangle



(e) Same pattern appearing again in another sentence

Figure 6: Illustration of various features of the system user interface

India[6], Spoken Tutorial[7], NCERT Biology[8], Income-tax Act[9], and NCERT Physics[10]. These datasets belong to the domains of government documents, technical tutorials and academic books, where, high quality translations are an imperative. Table 1 shows the corpus statistics in terms of number of sentences for each of the datasets.

4.1 Number of Frequent Patterns and Corpus Coverage

Number of Frequent patterns increase as the size of corpus increases. Corpus coverage depends on the number of syntactically well formed patterns extracted from the corpus which adhere to specified pattern length and frequency. Table 10 depicts information about number of filtered patterns extracted and coverage on five different corpus.

[6] http://indiacode.nic.in/coiweb/welcome.html
[7] http://spoken-tutorial.org/
[8] http://www.ncert.nic.in/NCERTS/textbook/textbook.htm?kebo1=0-22
[9] http://www.incometaxindia.gov.in/pages/acts/income-tax-act.aspx
[10] http://www.ncert.nic.in/NCERTS/textbook/textbook.htm?leph1=0-8

Table 1: Datasets and corpus coverage by patterns

Domain	#Sentences	#Frequent Patterns	#Frequent Instances	#Coverage %
Constitution of India	1582	12946	154218	86.62
Spoken Tutorial	16233	32974	10846	78.32
NCERT Biology	1144	615	12407	60.82
Income-Tax Act	1758	8391	104998	89.34
NCERT Physics	8013	15070	244034	79.94

4.2 Effect of Varying Pattern Length and Frequency Threshold for Pattern Extraction

One of the criterion to assess the quality of an individual extracted pattern is whether or not it appears in unseen data, thereby covering sentences in that data. A set of such patterns is then considered to be "good" if it collectively offers a high coverage on an unseen data. We split the datasets into MINE and TEST, where, the MINE split was used for extracting patterns and their coverage (in terms of number of words covered) was evaluated on the TEST split. We perform three-fold cross validation, varying both pattern length and frequency threshold from 2 to 6 and report coverage on MINE and TEST sets. Figure 7 captures the trade-off between pattern length, frequency threshold and coverage. For a fixed threshold, as the pattern length increases, the coverage on both MINE and TEST sets progressively decreases. Same observation applies when we fix the pattern length and increase the frequency threshold. We also observe that the gap in coverage is much smaller for varying frequency thresholds at smaller lengths and this gap progressively widens as the pattern length increases.

4.3 Effect of Varying Dictionary Size on Corpus Coverage

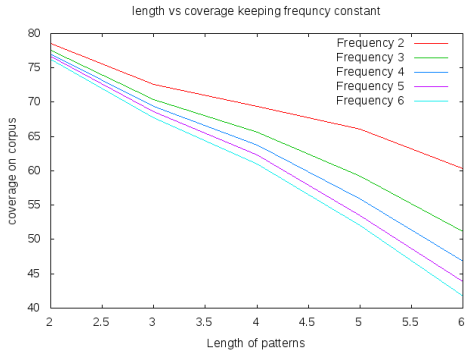
Our pattern selection algorithm constrains the cardinality of the set while maximizing a quality criteria like corpus coverage. Constraining the cardinality of the final set corresponds to limiting the size of the bilingual dictionary and this is desirable as the size of the bilingual dictionary is proportional to the human effort for translation. The corpus coverage increases with increasing size of the dictionary, however this increase is not linear but rather diminishes with increasing size of the dictionary. Figure 7d captures this relationship between coverage and fraction of patterns selected after sub-setting for different datasets.

4.4 Induced GF grammars

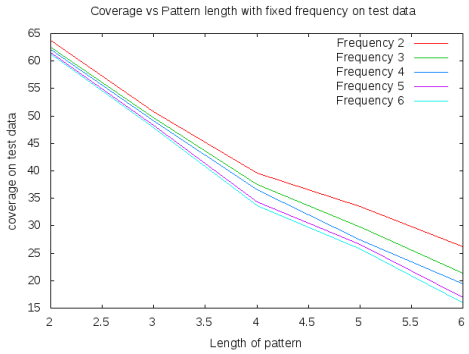
Once users provide translations of patterns, their instances, uncovered n-grams in sentences and reorders different chunks, grammatical framework rules are induced. Firstly, abstract syntax is induced which defines what meanings can be expressed in the grammar and then concrete English and concrete Hindi syntax is induced which provides mapping from meanings to strings in English and Hindi languages. Figure 8 illustrates a sample induced GF grammar. For a new sentence, extracted and translated patterns are given as input to GF grammars and if a match is found, then the sentence is reordered using the mapping from the concrete syntax. A more detailed example is available at `http://www.cse.iitb.ac.in/~vishwajeet/gf_rules.html`.

4.5 Conclusion

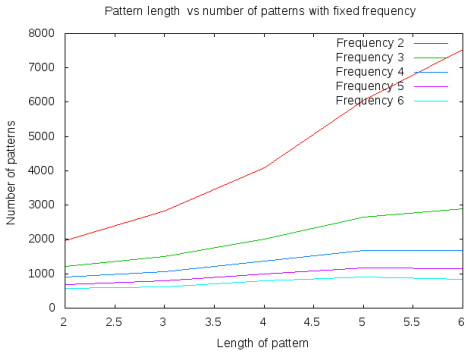
We presented an interactive machine translation approach for high quality translation of technical domain corpora. Given an in-domain source corpus, our system mines minimal



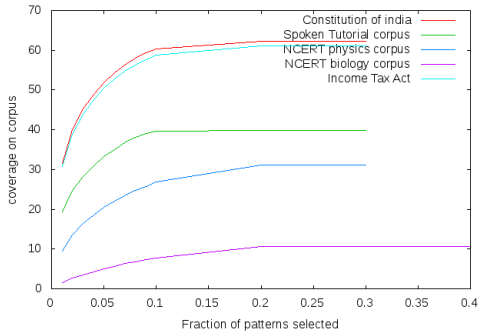
(a) Coverage vs. pattern length on the mining data



(b) Coverage vs. pattern length on the test data



(c) Pattern length vs number of patterns for a fixed threshold



(d) Coverage vs number of pattern selected after pattern selection

Figure 7: Corpus coverage for varying pattern lengths and frequency and coverage vs number of patterns selected after pattern selection

number of frequent patterns that maximally cover the corpus. Leveraging humans for their high quality translations, we continuously rebuild a rule-based translation engine that is realized using GF formalism.

```
abstract ST = {

flags coding = utf8 ;

flags startcat = Sent;

cat
    Sent; Sent1; NT2233; NT2234; P2; Ch1; Sent2; Ch2; Sent3; Sent4; Ch3; NT3402; Ch4; Ch5; P3; Sent5; P1;

fun
    NT2233_Fun : NT2233;
    NT2234_Fun : NT2234;
    Sent1_Fun : P1 -> Sent1;
    Sentence1_Fun : Sent1 -> Sent;
    Sent2_Fun : P2 -> Ch1 -> Sent2;
    Sentence2_Fun : Sent2 -> Sent;
    P2_Fun : P2;
    Ch1_Fun : Ch1;
    Sentence3_Fun : Sent3 -> Sent;
    Sent3_Fun : Ch2 -> Sent3;
    Ch2_Fun : Ch2;
    Sentence4_Fun : Sent4 -> Sent;
    Sent4_Fun : Ch3 -> Sent4;
    Ch3_Fun : Ch3;
    NT3402_Fun : NT3402;
    Sentence5_Fun : Sent5 -> Sent;
    Sent5_Fun : Ch4 -> P3 -> Ch5 -> Sent5;
    Ch4_Fun : Ch4;
    P3_Fun : NT3402 -> P3;
    Ch5_Fun : Ch5;
    P1_Fun : NT2233 -> NT2234 -> P1;

}
```

(a) Abstract Grammar

```
concrete STEng of ST = {

flags coding = utf8 ;

lincat
    NT2233 = Str;
    NT2234 = Str;
    Sent1 = Str;
    Sent = Str;
    P2 = Str;
    Ch1 = Str;
    Sent2 = Str;
    Ch2 = Str;
    Sent3 = Str;
    Sent4 = Str;
    Ch3 = Str;
    NT3402 = Str;
    Ch4 = Str;
    Ch5 = Str;
    P3 = Str;
    Sent5 = Str;
    P1 = Str;

lin
    NT2233_Fun  = "the spoken tutorial on Logical operators";
    NT2234_Fun  = "C and C++";
    Sent1_Fun p1 = p1;
    Sentence1_Fun sent1 = sent1;
    P2_Fun  = "In this tutorial";
    Ch1_Fun  = "we will learn about Logical operators like : Logical AND e.g. expression1 && expression2";
    Sent2_Fun p2 ch1 = p2 ++ ch1;
    Sentence2_Fun sent2 = sent2;
    Sentence3_Fun sent3 = sent3;
    Sent3_Fun ch2 = ch2;
    Ch2_Fun  = "Logical OR eg . expression1 OR expression2";
    Sentence4_Fun sent4 = sent4;
    Sent4_Fun ch3 = ch3;
    Ch3_Fun  = "Logical NOT eg . not ( Expression1 )";
    NT3402_Fun  = "the help of examples";
    Sentence5_Fun sent5 = sent5;
    Sent5_Fun ch4 p3 ch5 = ch4 ++ p3 ++ ch5;
    Ch4_Fun  = "We";
    P3_Fun nt3402 = "will do this with" ++ nt3402;
```

(b) Concrete English Grammar

```
concrete STHin of ST = {

flags coding = utf8 ;

lincat
    NT2233 = Str;
    NT2234 = Str;
    Sent = Str;
    Sent1 = Str;
    P2 = Str;
    Ch1 = Str;
    Sent2 = Str;
    Ch2 = Str;
    Sent3 = Str;
    Sent4 = Str;
    Ch3 = Str;
    NT3402 = Str;
    Ch4 = Str;
    Ch5 = Str;
    P3 = Str;
    Sent5 = Str;
    P1 = Str;

lin
    NT2233_Fun  = "तार्किक ऑपरेटर्स पर स्पोकन ट्यूटोरियल";
    NT2234_Fun  = "सी और सी++";
    Sent1_Fun p1 = p1;
    Sentence1_Fun sent1 = sent1;
    P2_Fun  = "इस ट्यूटोरियल में";
    Ch1_Fun  = "हम तार्किक ऑपरेटर्स के बारे में सीखना होगा जैसे : तार्किक AND उदाहरण: अभिव्यक्ति1 && अभिव्यक्ति2 ";
    Sent2_Fun p2 ch1 = p2 ++ ch1;
    Sentence2_Fun sent2 = sent2;
    Sentence3_Fun sent3 = sent3;
    Sent3_Fun ch2 = ch2;
    Ch2_Fun  = "तार्किक OR  जैसे: अभिव्यक्ति1 || अभिव्यक्ति2 ";
    Sentence4_Fun sent4 = sent4;
    Sent4_Fun ch3 = ch3;
    Ch3_Fun  = "तार्किक NOT उदाहरण: ! ( अभिव्यक्ति1 ) ";
    NT3402_Fun  = "उदाहरणो की सहायता";
    Sentence5_Fun sent5 = sent5;
    Sent5_Fun ch4 p3 ch5 = ch4 ++ p3 ++ ch5;
    Ch4_Fun  = "हम";
```

(c) Concrete Hindi Grammar

Figure 8: Induced abstract grammar, concrete english grammar and concrete hindi grammar

References

Alabau, V., Buck, C., Carl, M., Casacuberta, F., Garcıa-Martınez, M., Germann, U., González-Rubio, J., Hill, R., Koehn, P., Leiva, L., et al. (2014). Casmacat: A computer-assisted translation workbench. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 25–28.

Aziz, W., Castilho, S., and Specia, L. (2012). Pet: a tool for post-editing and assessing machine translation. In *LREC*, pages 3982–3987.

Bertoldi, N., Cettolo, M., and Federico, M. (2013). Cache-based online adaptation for machine translation enhanced computer assisted translation. *Proceedings of the XIV Machine Translation Summit*, pages 35–42.

Bharati, A., Chaitanya, V., Kulkarni, A. P., Sangal, R., and Rao, G. U. (2003). Anusaaraka: overcoming the language barrier in india. *arXiv preprint cs/0308018*.

Carl, M. (2012). Translog-ii: a program for recording user activity data for empirical reading and writing research. In *LREC*, pages 4108–4112.

Cristina Espa˜na-Bonet, Ramona Enache, A. S. A. R. L. M. M. G. (2011). Patent translation within the molto project. *MT Summit*.

Daniel Ortiz-Marfınez, Luis A. Leiva, V. A. s. G.-V. F. C. (2011). An interactive machine translation system with online learning. *ACL-HLT*, pages 68–73.

Denkowski, M. and Lavie, A. (2012). Transcenter: Web-based translation research suite. In *AMTA 2012 Workshop on Post-Editing Technology and Practice Demo Session*, page 2012.

Federico, M., Bertoldi, N., Cettolo, M., Negri, M., Turchi, M., Trombetti, M., Cattelan, A., Farina, A., Lupinetti, D., Martines, A., et al. (2014). The matecat tool. In *Proceedings of COLING*, pages 129–132.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Kay, M. (1980). The proper place of men and machines in language translation. *CSL*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lavie, M. D. C. D. A. (2014). Learning from post-editing: Online model adaptation for statistical machine translation. *EACL 2014*, page 395.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 440–447, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ortiz-Martínez, D., García-Varea, I., and Casacuberta, F. (2010). Online learning for interactive statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 546–554, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pelletier, F. J. (1994). The principle of semantic compositionality. *Topoi*, 13(1):11–24.

Ranta, A. (2004). Grammatical framework. *Journal of Functional Programming*, 14(02):145–189.

Roturier, J., Mitchell, L., and Silva, D. (2013). The accept post-editing environment: A flexible and customisable online tool to perform and analyse machine translation post-editing. In *Proceedings of MT Summit XIV Workshop on Post-editing Technology and Practice*, pages 119–128.

Sato, S. and Nagao, M. (1990). Toward memory-based translation. In *Proceedings of the 13th conference on Computational linguistics-Volume 3*, pages 247–252. Association for Computational Linguistics.

Somers, H. (1999). Review article: Example-based machine translation. *Machine Translation*, 14(2):113–157.

Streiter, O. (1996). *Linguistic modeling for multilingual machine translation.*

Terumasa, E. (2007). Rule based machine translation combined with statistical post editor for japanese to english patent translation. In *Proceedings of the MT Summit XI Workshop on Patent Translation*, volume 11, pages 13–18.

Toselli, A. H., Vidal, E., and Casacuberta, F. (2011). Interactive machine translation. In *Multimodal Interactive Pattern Recognition and Applications*, pages 135–152. Springer.

Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B., and Waibel, A. (2003). The cmu statistical machine translation system. In *IN PROCEEDINGS OF MT SUMMIT IX*, pages 110–117.

Wang, Y.-Y. and Waibel, A. (1998). Fast decoding for statistical machine translation. In *ICSLP*.