SmartMATE: An Online End-To-End MT Post-Editing Framework

Sergio Penkale   **Andy Way**

Applied Language Solutions
Delph, UK
`firstname.lastname@appliedlanguage.com`

Abstract

It is a well-known fact that the amount of content which is available to be translated and localized far outnumbers the current amount of translation resources. Automation in general and Machine Translation (MT) in particular are one of the key technologies which can help improve this situation. However, a tool that integrates all of the components needed for the localization process is still missing, and MT is still out of reach for most localisation professionals. In this paper we present an online translation environment which empowers users with MT by enabling engines to be created from their data, without a need for technical knowledge or special hardware requirements and at low cost. Documents in a variety of formats can then be post-edited after being processed with their Translation Memories, MT engines and glossaries. We give an overview of the tool and present a case study of a project for a large games company, showing the applicability of our tool.

1   Introduction

The amount of content that needs to be translated and localised is increasingly growing (DePalma and Kelly, 2009). With the current focus on user-generated content and an increasing commercial interest in emerging economies, the contents which are available for translation and the amount of languages into which this content is published are set to continue increasing. However, the high costs associated with translation and localisation mean that only a fraction of this content actually ends being translated, even more so given the current global economic difficulties.

It is hardly surprising then that, as evidenced by SDL's acquisition of Language Weaver, Language Service Providers (LSPs) are turning to automation in a bid to reduce translation costs at the same time as increasing the volume of translated content. However, while large LSPs are benefiting from the increased productivity associated with state-of-the-art Statistical Machine Translation (SMT), this technology remains out of reach for smaller organizations and individual translators. In particular, a tool that integrates all of the components required in a typical translation workflow (cf. Figure 1 for a sketch, and Section 3 for details on each of the steps in this workflow), and which allows users to easily exploit MT and postedit its output is crucial to enable mass adoption of MT.

In this paper we present one such tool. SmartMATE (Way et al., 2011) is a self-serve translation platform which supports File Filtering, Machine Translation, Terminology management, and which has an integrated Editor Suite. Crucially, SmartMATE enables both individuals and companies to train an MT engine using their own data, at the press of just a few buttons. By doing so, SmartMATE effectively removes the main barriers against exploiting MT technology. Expensive hardware requirements and technical knowledge are done away with, and so is computational linguistics expertise. In addition, SmartMATE supports unique capabilities such as concurrent translation and proofreading, terminology-aware MT, and integrated QA control inside the editor. We present all of SmartMATE's
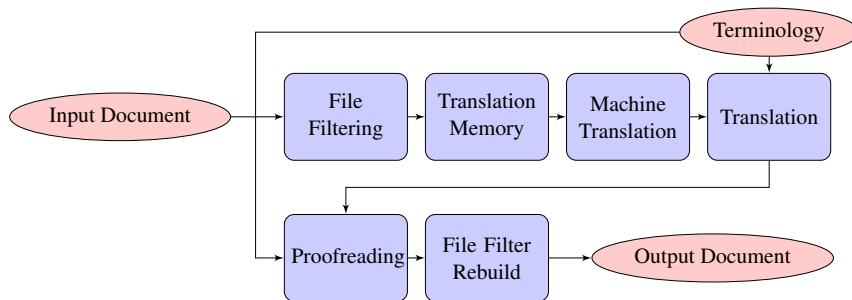
Figure 1: Typical translation workflow

capabilities, and discuss a case study of a large translation project carried out using our tool.

The remainder of this paper is organized as follows. Section 2 provides a brief review of translation platforms of a similar nature to the one presented in this paper. Section 3 presents SmartMATE and gives an overall introduction to all of its capabilities. In Section 4 we analyse a project currently being run for one of our customers using SmartMATE. We conclude and give avenues for future work in Section 5.

2 Related Tools

Although a few products which enable MT output to be postedited have been made available over the last few years, we are not aware of any tool which integrates all the capabilities offered by Smart-MATE. Google Translator Toolkit[1] allows users to upload documents and pre-translate them using Google Translate. However, unlike SmartMATE only generic MT engines are used, providing no facility for the user to train an engine adapted specifically to their data. In addition, although terminology is supported in the post-editing environment, the MT engines are not aware of glossaries, making the pre-translated content unaware of the user's terminology requirements.

Unlike Google's offer, Microsoft Translator Hub[2] does enable user-specific engines to be created. It does not, however, provide postediting facilities, making the need for an external tool a requirement in order to allow a linguist to correct the MT output.

Finally, an interesting tool which finds itself in the

opposite situation is PET (Aziz et al., 2012), which was designed specifically to post-edit the output of MT systems, and to collect various kinds of statistics from the process. However, the tool comprises only the editor part, and no actual MT services are provided.

3 SmartMATE

SmartMATE (Way et al., 2011) is an online self-serve translation platform. It is designed to be a one-stop portal where users can upload their Translation Memory (TM) files, and create user-customized MT engines trained using these TMs. It integrates all the capabilities needed in a typical translation workflow.

Figure 1 gives a sketch of a typical translation workflow in SmartMATE. Assume an input document which needs to be translated arrives. Since there is a variety of file formats in which this document can be encoded, it is first sent to File Filtering, which produces an XLIFF[3] (XML Localisation Interchange File Format) file containing only the translatable text, without additional elements such as images or page formatting information. Except for File Filtering, all of the components in SmartMATE take an XLIFF file as input and produce a modified one as output. This XLIFF can then optionally be sent through Translation Memory for leveraging of any previous translations, and through MT for segments which do not match any TM entry. At this stage, the document becomes available for editing. Smart-MATE provides an online multi-user Editor Suite. Users can utilise the editor themselves to translate the document, or they might delegate this to a third party who receives an invitation email which enables

them to work on the document using the online editor. After translation has finished, the translated XLIFF file is sent back to File Filtering to recover the original file format. The following sections provide details on each of these components.

It is important to note that SmartMATE's terms and conditions explicitly state that any data uploaded into SmartMATE will be kept confidential. TMs, input documents, glossaries and MT engines are kept in the user's password-protected area, being unreachable by other users, and ALS will not exploit any of this data for other purposes without the users's consent.

3.1 File Filtering

SmartMATE accepts a wide range of input document formats, including Microsoft Office Suite file formats (e.g. .doc, .xls, .ppt), as well as other popular formats such as .rtf, .html, .ttx and .txt.

In addition to text which needs to be translated, input documents will likely contain additional data such as formatting information, formatting tags, images, etc. The File Filtering process involves identifying the (textual) localizable content. This content is extracted and decoupled from any non-translatable content (the exception are in-line formatting tags, such as the ones used to indicate italics or boldface, which are preserved and encapsulated), resulting in a clean text version of the content which is ready to be translated, and which a linguist can edit without needing to purchase a license for the software the original document was saved in, e.g. Microsoft Office.

In addition to producing an XLIFF file, the File Filtering module also produces a skeleton of the document which contains information complementary to that in the XLIFF and which is needed to rebuild the original file format. This is used in the last stage of the workflow to produce a final document which has the same formatting as the original, but where the content has been translated.

3.2 Translation Memory

Users can upload TM files containing their previously translated data. SmartMATE is able to import TMs stored in the standard TMX[4] (Translation

Memory eXchange) format, which can be exported from any Translation Management System software.

TMs inside SmartMATE can be exploited in two different ways. Firstly, they can be used as traditional Translation Memories. When a new document is ready for translation, any segment in the document which exactly matches a TM entry will appear in the editor suite as pre-translated using the target side of this entry. In addition to exact matches, SmartMATE also leverages entries which only match above a predefined match threshold (Fuzzy Matches) (Sikes, 2007), and is able to identify In-Context Exact (ICE) matches, i.e. segments which are an exact match and which are preceded and followed by an exact match segment. After a document has been translated and signed-off by the proofreader, TMs can be automatically updated to include the newly translated content.

In addition to being used as traditional TMs, any TMX uploaded by the user can be used to train an MT engine, as explained in the following section.

3.3 Machine Translation

After TM files have been uploaded, these can be used to train MT engines. After the user has completed a simple form with the details of their requested engine, a process starts which requires no human intervention and which produces a state-of-the-art SMT engine. The process begins by extracting plain bilingual text from the TMX files, thus creating a parallel corpus. This is then subject to multiple stages of corpus cleaning which include:

- ensuring the correct character encodings are being used,

- removing any formatting tags so that they do not interfere with the training process,

- removing duplicate sentence pairs,

- removing sentence pairs which exceed certain source:target length ratio,

- replacing entities such as URLs and e-mails with placeholders to improve the generalization of the statistical models.

After the corpus has been cleaned, 1,000 randomly selected sentence pairs are kept apart for

---

[4]http://www.gala-global.org/oscarStandards/tmx/tmx14b.html

evaluation purposes, and an additional 500 sentence pairs for tuning. The remaining data is used to train SMT models using the Moses (Koehn et al., 2007) toolkit. The user is then presented with the built engine along with automatically obtained BLEU (Papineni et al., 2002) scores, which are calculated over the 1,000 randomly held-out sentence pairs and which give an indication of the level of translation quality that could be expected from this engine when used to translate documents of a nature similar to those used when training the engine.

The process of building an engine involves creating phrase-based translation models (Koehn et al., 2003) and lexicalized reordering models (Koehn et al., 2005) as well as a Language Model (LM), for which the IRSTLM toolkit (Federico and Cettolo, 2007) is used. In addition, the model weights are optimized using Minimum Error Rate Training (Och, 2003) so as to maximize the BLEU score over the 500 sentence pairs randomly held out from the original TMs for tuning. All of this complexity, as well as the significant hardware requirements needed to host the engine training, are hidden from the user.

It is worth noting that since these engines have been built using the user's own data, they are specialized engines from which a better translation quality can be expected[5] when compared to general-purpose engines such as those provided by services such as Google Translate[6] or Microsoft Bing Translator,[7] which in addition might not offer the same data privacy guarantees as SmartMATE.

3.4 Terminology

SmartMATE is able to import multilingual glossaries containing user-specific terminology. The accepted formats are CSV (Comma-Separated Values) files, which are obtainable from any spreadsheet software, or the standard TBX (TermBase eXchange) (ISO 30042, 2008).

These glossaries can be exploited in several ways. Firstly they can be used as a complement of TMX files during MT engine building. This has the effect of improving word alignment (and subsequently

phrase-alignment), as it provides reference points for the SMT alignment algorithms (Och and Ney, 2000). Secondly, they can be used for glossary-injection during MT. Once an engine has been trained, glossaries can be used while the engine is processing an input document to ensure that the MT output adheres to the terminology specified by the glossary. When using multiple glossaries which provide conflicting entries for the same source term, all of the possible target translations are provided to the engine, which uses its LM to determine which translation option provides the most fluent target sentence.

Finally, the editor suite supports the use of glossaries as well, by highlighting any source term which matches a source segment, and providing to the linguist the available target terms. The editor is also able to detect whether the target term specified in the glossary has been used in translating the segment, and to flag with a warning segments which do not conform to entries in the glossary.

3.5 Editor Suite

The editor suite integrates all of SmartMATE's capabilities, effectively providing the user with a single tool that can be used for the complete translation workflow. SmartMATE is cloud-based, as it is hosted on Amazon's cloud. This has several beneficial implications. Firstly, data is automatically saved at segment level, which means that any technical problem on the user's computer will not affect the integrity of the translated data. Secondly, the user is able to access their data from any computer which is equipped with an internet connection. Even though a collection of TMs and MT engines can easily require several Giga Bytes of disk space to be stored, the user can quickly access this data from any computer with an internet browser. Finally, its cloud-based nature means that SmartMATE is able to scale virtually arbitrarily. Regardless of the amount of users currently accessing the system or running MT engines, each user is assigned a dedicated virtual PC in the cloud so that system performance is unaffected.

The editor provides two operation modes: translation and proofreading, which we discuss in the following sections.

---

[5]This is mainly due to the ambiguity introduced by out-of-domain data (Sennrich, 2012), and is a known effect in the domain adaptation literature, e.g. (Foster et al., 2010)

[6]http://translate.google.com

[7]http://www.microsofttranslator.com

Figure 2 (Translation mode):

Translation mode...   Editing Document : company.es.xlf • Job Name : Title

Completed by me : **3.57%** Total completed : **3.57%**

Search & Replace

Pass All Segments   Next Available Segment

0002 EXACT — Founded in 2003, Applied Language Solutions (ALS) is now the world's fastest growing language services provider. | Fundada en 2003, Applied Language Solutions (ALS) es ahora el proveedor de servicios lingüísticos con mas rápido crecimiento en el mundo

0003 FUZZY — We offer fast, efficient and accurate services, including translation, interpreting, localisation and proofreading, to a portfolio of well known customers and niche operators across multiple industries - all of which is delivered by helpful and personable localisation experts. | Ofrecemos servicios rápidos, eficientes y precisos, que incluyen traducción, interpretación, localización y edición, a un porfolio de clientes bien conocidos y operadores nichos a lo largo de múltiples industrias. Todo esto es entregado por expertos en localización útiles y dispuestos a ayudar

0004 MT — ALS takes great pride in the values that continue to drive the success of the company, which are: | ALS takes great pride en el que desea continuar con los valores duro el success de la empresa, que están:

0005 MT — RESPECT, PRIDE and approaching our work with a CAN DO attitude. | respect, PRIDE y approaching nuestros work con un NO puede attitude.

0006 EXACT — The company has grown significantly since its modest beginnings, generating just over £6.2 million ($10.18 million) in global sales in 2009 and now employs more than 120 people across eleven international offices in the UK, Bulgaria, India, France, Germany, Spain, Guatemala, Australia and the US (x3). | La compañía ha crecido significativamente desde sus humildes comienzos, generando apenas sobre £6,2 millones ($10,18 millones) en ventas globales en 2009 y ahora emplea a más de 120 personas entre once oficinas internacionales en el Reino Unido, Bulgaria, India, Francia, Alemania, España,

0007 EXACT — In December 2009, ALS featured in the Sunday Times Fast Track 100 for the second year running. | En diciembre de 2009, ALS apareció en el Sunday Times Fast Track 100 por segundo año consecutivo.

Figure 2: Translation mode in the editing environment

Figure 3 (Proof Reading mode):

Proof Reading mode...   Editing Document : company.es.xlf • Job Name : Title

Completed by me : **0.00%** Total completed : **0.00%**

Search & Replace

Pass All Segments

0002 EXACT — Founded in 2003, Applied Language Solutions (ALS) is now the world's fastest growing language services provider. | Fundada en 2003, Applied Language Solutions (ALS) es ahora el proveedor de servicios lingüísticos con mas rápido crecimiento en el mundo

0003 FUZZY — We offer fast, efficient and accurate services, including translation, interpreting, localisation and proofreading, to a portfolio of well known customers and niche operators across multiple industries - all of which is delivered by helpful and personable localisation experts. | Ofrecemos servicios rápidos, eficientes y precisos, que incluyen traducción, interpretación, localización y edición, a un porfolio de clientes bien conocidos y operadores nichos a lo largo de múltiples industrias. Todo esto es entregado por expertos en localización útiles y

0004 MT — ALS takes great pride in the values that continue to drive the success of the company, which are: | ALS takes great pride en el que desea continuar con los valores duro el success de la empresa, que están:

0005 MT — RESPECT, PRIDE and approaching our work with a CAN DO attitude. | respect, PRIDE y approaching nuestros work con un NO puede attitude.

0006 EXACT — The company has grown significantly since its modest beginnings, generating just over £6.2 million ($10.18 million) in global sales in 2009 and now employs more than 120 people across eleven international offices in the UK, Bulgaria, India, France, Germany, Spain, Guatemala, | La compañía ha crecido significativamente desde sus humildes comienzos, generando apenas sobre £6,2 millones ($10,18 millones) en ventas globales en 2009 y ahora emplea a más de 120 personas entre once oficinas internacionales en el Reino Unido, Bulgaria, India,

0007 EXACT — In December 2009, ALS featured in the Sunday Times Fast Track 100 for the second year running. | En diciembre de 2009, ALS apareció en el Sunday Times Fast Track 100 por segundo año consecutivo.

Figure 3: Proofreading mode in the editing environment

3.5.1 Translation

Figure 2 shows SmartMATE's editor suite in translation mode. There are two main columns, with the left one showing the translatable source content which was extracted from the original file, and the right one the corresponding target segments. Depending on which modules were activated by the user, the initial content in the target segments will change. In this particular example, both TM and MT were activated, as can be observed from the information displayed to the left of each segment. Segments are labelled according to whether they resulted in a TM match (either exact, fuzzy or in-context exact), or whether they were sent to MT.

This figure also illustrates the use of glossaries within the editor. Segments 2 and 3 contain source terms which have been highlighted, meaning that these terms matched a glossary entry. Hovering the mouse over these terms will show the translations suggested by the glossary. In addition, when editing the target side of a segment, linguists have access to a Glossary tab from which they can easily incorporate glossary terms into the translation. The red warning sign in segment 3 illustrates how Smart-MATE indicates that a segment contains glossary matches but the target terms specified in the glossary have not been used in the translation.

Once a translator has finished editing a segment, the segment can be locked. This is automatically done by the Editor when switching to a different segment, or can be explicitly triggered by clicking on the dedicated button which separates source from target segments. In Figure 2, only segment 4 has been locked, which is indicated by a different background colour and a lock symbol. When a segment is locked, it instantly becomes available for the next stage of the workflow, e.g. proofreading. See Section 3.5.2 for the concurrency implications of being able to lock an individual segment, rather than the complete document.

Finally, segment 5 shows how in-line formatting can be protected. In the original file, the words "RE-SPECT, PRIDE" were typed in boldface. Smart-MATE's editor hides this formatting to the user, but explicitly shows that there is formatting information which should be preserved. Linguists can drag and drop these protected tags from source to target so as
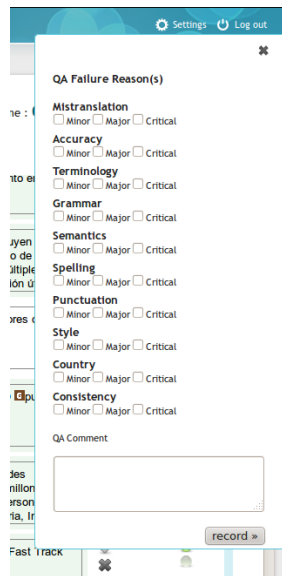

Figure 4: LISA QA-compliant feedback form

to keep the formatting. The same principle can be applied to preserve tags when translating structured documents such as HTML or XML files.

3.5.2 Proofreading

In addition to allowing the post-editing of MT output (and/or fuzzy TM matches, depending on which modules were activated for a particular job), SmartMATE also supports a proofreading stage were a different linguist can asses the work done by the translators, ensuring the coherence of the complete document, the adherence to client-specific policies and terminology, etc.

Figure 3 shows the proofreader's perspective of the document which is being translated in Figure 2. As can be seen, only segment 4 has become available for proofreading, as this is the only segment which has so far been locked by the translator.

Proofreaders are able to edit the target segments, and mark each segment as finished. If a translated segment contains severe errors, the proofreader can send the segment back to the translation phase, by clicking on the red cross next to it. When doing so, they can record detailed information about the linguist's reasons why the segment has been rejected, by using the form shown in Figure 4. This form conforms to the Localization Industry Standards Association (LISA) QA Model.
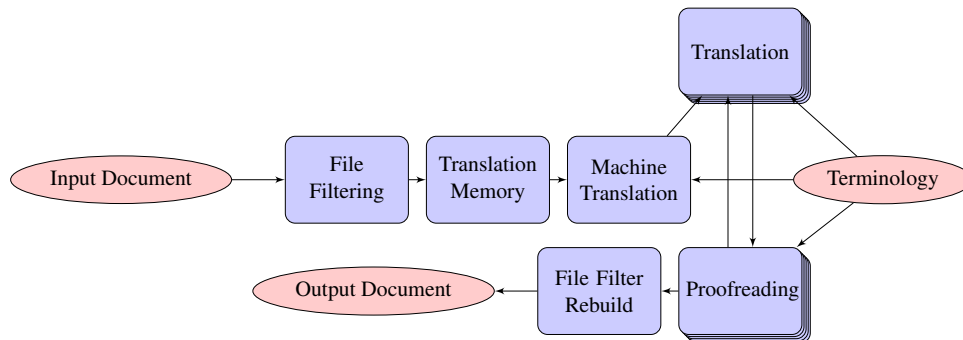
Figure 5: Possible translation workflow in SmartMATE

Figure 6: Character length limit being enforced to a segment by the Editor

Note that content becomes available for proofreading at *segment* level. That is, as soon as a translator has locked a segment, the proofreader is able to post-edit it and send it to the next stage, or send it back to translation. This means that, in addition to supporting the traditional (linear) workflow of Figure 1, the editor in SmartMATE enables proofreading to be done simultaneously to translation, effectively reducing proofreading time to zero. While some projects are best suited by the traditional linear workflow, there certainly are situations in which this concurrency model is desirable. In effect, Smart-MATE allows a workflow such as the one in Figure 5, where translation and proofreading run concurrently. Additionally, SmartMATE allows multiple users to collaborate on the same document at the same time, enabling further reductions in translation time.

4   Case Study

In order to demonstrate the robustness and usefulness of our tool, we discuss in this section a translation project which is being carried out for Spil Games,[8] a large online games developer and publisher of the type seen on social networking sites.

Games are originally written in English, and are subsequently localized into over 15 languages for a global audience of more than 180 million monthly active users.

Spil Games provides the localizable content to the author's institution (ALS), which is in charge of File Engineering, Project Management, TM/MT application and translation. Reviewing, however, is outsourced to a third party (VistaTEC).[9] The whole process is supported by and hosted in SmartMATE. ALS creates a new translation job in SmartMATE, and assigns the reviewing task to VistaTEC. Once the translation stage is complete, VistaTEC can itself delegate the reviewing to an arbitrary number of SmartMATE users from within the tool. The identity of the linguists who review the content is not revealed to ALS, thus ensuring VistaTEC's commercial confidentiality.

During the first stages of the project, only TM and Glossaries are used. However, after each new document has been translated, SmartMATE automatically updates the Translation Memories so that this newly created content can be matched against future documents. During the course of the project, as more content is translated the TM files will eventually reach a size substantial enough to allow customer-specific engines to be trained from them. We expect significant improvements in translation speed to be achieved once this happens.

The content translated for company A must satisfactorily be displayed inside the User Interface of a game, which means that some segments must conform to length restrictions. This requirement is ac-

---

[8]http://www.spilgames.com/

[9]http://www.vistatec.com/

Target Language	Segments	Source Words	Target Words	Exact	Fuzzy
Portuguese (Brazilian)	262	3,997	4,110	24%	6%
Russian	257	3,810	3,294	25%	6%
Turkish	250	3,608	3,183	24%	7%
Indonesian	256	3,787	3,327	24%	6%
Dutch	295	4,286	3,728	24%	5%
Portuguese (Portugal)	211	2,663	2,866	28%	8%
German	264	3,951	3,869	23%	6%
French	242	3,538	3,845	22%	5%
Swedish	289	4,089	3,923	21%	6%
Spanish	258	3,914	4,344	24%	6%
Italian	208	2,796	3,083	30%	6%
Polish	238	3,059	2,944	26%	7%
Arabic (Modern Standard)	111	2,353	1,851	0%	0%

Table 1: Statistics for each language pair in the project

commodated in SmartMATE by allowing a character limit to be specified in an XLIFF element at segment level, using the `maxwidth` property. Spil Games can then specify the desired limit, and this is enforced by the editor, as illustrated in Figure 6.

We give in Table 1 statistics gathered during one of the first weeks in the project. During this week, an average of 241 segments were translated from English into 13 language pairs, which amount to 45,851 source words among all language pairs. Although the average sentence length among all of the English segments is 14.6 words, there is a large variance. Most of the content to be translated consists of titles and descriptions. Titles tend to be quite short, while descriptions are longer. We see that for most language pairs, an exact match rate of between 20% and 30% is achieved. Although this means that a significant amount of translation work is reduced due to SmartMATE exploiting our customer's TMs, we noticed that most of the matching segments were titles rather than descriptions. We expect, however, that as TMs grow in size, a larger number of long segments will be able to be matched, and that the incorporation of post-edited MT into the project will significantly reduce turn-around times.

5 Conclusions and Future Work

In this paper we have presented SmartMATE, an online self-serve MT translation platform, which integrates TM, MT and Terminology into a powerful editing environment. We have shown not only how the complete localisation workflow can be accommodated using this single tool, but also how the concurrency capabilities of the editor enable additional workflows to be considered. In addition we have studied the first stages of a particular project from a large client which is currently being run using SmartMATE, showing that our product is robust enough to be used in large-scale production environments. We believe that SmartMATE has the capability of empowering non-technical users with MT technology, and of advancing the standards in the localisation industry.

There are many areas in which we can continue to improve SmartMATE. In the short term, we will focus on extending the number of file formats supported by our file filtering module (e.g. pdf), and on enabling advanced modules when training MT engines, such as named entity recognizers, segmenters, tokenizers and compound splitters.

Acknowledgments

References

Wilker Aziz, Sheila Castilho Monteiro de Sousa, and Lucia Specia. 2012. PET: a tool for post-editing and assessing machine translation. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3982–3987, Istanbul, Turkey.

Donald A. DePalma and Nataly Kelly. 2009. The business case for machine translation. Common Sense Advisory. http://www.commonsenseadvisory.com/AbstractView.aspx?ArticleID=859.

Marcello Federico and Mauro Cettolo. 2007. Efficient Handling of N-gram Language Models for Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 88–95, Prague, Czech Republic.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 451–459, Cambridge, MA.

ISO 30042. 2008. *Systems to manage terminology, knowledge and content – TermBase eXchange (TBX)*. ISO, Geneva, Switzerland.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 48–52, Edmonton, Canada.

Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proceedings of the International Workshop on Spoken Language Translation*, Pittsburgh, PA.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the ACL, demonstation session*, pages 177–180, Prague, Czech Republic.

Franz Josef Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, pages 1086–1090, Saarbrücken, Germany.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA.

Rico Sennrich. 2012. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 539–549, Avignon, France.

Richard Sikes. 2007. Fuzzy matching in theory and practice. *Multilingual*, 18(6):39–43.

Andy Way, Kenny Holden, Lee Ball, and Gavin Wheeldon. 2011. SmartMATE: Online self-serve access to state-of-the-art SMT. In *Proceedings of the Third Joint EM+/CNGL Workshop "Bringing MT to the User: Research Meets Translators", JEC 2011*, pages 43–52, Luxembourg.