SORT: An Interactive Source-Rewriting Tool for Improved Translation

Shachar Mirkin, Sriram Venkatapathy, Marc Dymetman, Ioan Calapodescu
Xerox Research Centre Europe
6 Chemin de Maupertuis
38240 Meylan, France
firstname.lastname@xrce.xerox.com

Abstract

The quality of automatic translation is affected by many factors. One is the divergence between the specific source and target languages. Another lies in the source text itself, as some texts are more complex than others. One way to handle such texts is to modify them prior to translation. Yet, an important factor that is often overlooked is the source *translatability* with respect to the specific translation system and the specific model that are being used. In this paper we present an interactive system where source modifications are induced by confidence estimates that are derived from the translation model in use. Modifications are automatically generated and proposed for the user's approval. Such a system can reduce post-editing effort, replacing it by cost-effective pre-editing that can be done by monolinguals.

1 Introduction

While Machine Translation (MT) systems are constantly improving, they are still facing many difficulties, such as out-of-vocabulary words (i.e. words unseen at training time), lack of sufficient in-domain data, ambiguities that the MT model cannot resolve, and the like. An important source of problems lies in the source text itself – some texts are more complex to translate than others.

Consider the following English-to-French translation by a popular service, BING TRANSLATOR:[1] *Head of Mali defense seeks more arms* → *Défense de la tête du Mali cherche bras plus.* There, apart from syntactic problems, both *head* and *arms* have been translated as if they were body parts (*tête* and *bras*). However, suppose that we express the same English meaning in the following way: *Chief of Mali defense wants more weapons.* Then BING produces a much better translation: *Chef d'état-major de la défense du Mali veut plus d'armes.*

The fact that the formulation of the source can strongly influence the quality of the translation has long been known, and there have been studies indicating that adherence to so-called "Controlled Language" guidelines, such as *Simplified Technical English*[2] can reduce the MT post-edition effort. However, as one such study (O'Brien, 2006) notes, it is unfortunately not sufficient to just "*apply the rules [i.e. guidelines] and press Translate. We need to analyze the effect that rules are having on different language pairs and MT systems, and we need to tune our rule sets and texts accordingly*".

In the software system presented here, SORT (*SOurce Rewriting Tool*), we build on the basic insight that formulation of the source needs to be geared to the specific MT model being used, and propose the following approach. First, we assume that the original source text in English (say) is not necessarily under the user's control, but may be given to her. While she is a fluent English speaker, she does not know at all the target language, but uses an MT system; crucially, this system is able *to provide estimates of the quality of its translations* (Specia et al., 2009). SORT then automatically produces a number of rewritings of each English sentence, translates them with the MT system, and displays to the user those rewritings for which the translation quality estimates are higher than the estimate for the original source. The user then interactively selects one such rewriting per sentence, checking that it does not distort the original meaning, and finally the translations of these

[1] http://www.bing.com/translator, accessed on 4/4/2013.

[2] http://www.asd-ste100.org

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 85–90,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

reformulations are made available.

One advantage of this framework is that the proposed rewritings are implicitly "aware" of the underlying strengths and limitations of the specific MT model. A good *quality estimation*[3] component, for instance, will feel more confident about the translation of an unambiguous word like *weapon* than about that of an ambiguous one such as *arm*, or about the translation of a known term in its domain than about a term not seen during training.

Such a tool is especially relevant for business situations where post-edition costs are very high, for instance because of lack of people both expert in the domain and competent in the target language. Post-edition must be reserved for the most difficult cases, while pre-edition may be easier to organize. While the setup cannot fully guarantee the accuracy of all translations, it can reduce the number of sentences that need to go through post-edition and the overall cost of this task.

2 The rewriting tool

In this section we describe SORT, our implementation of the aforementioned rewriting approach. While the entire process can in principle be fully automated, we focus here on an interactive process where the user views and approves suggested rewritings. The details of the rewriting methods and of the quality estimation used in the current implementation are described in Sections 3 and 4.

Figure 1 presents the system's interface, which is accessed as a web application. With this interface, the user uploads the document that needs to be translated. The translation confidence of each sentence is computed and displayed next to it. The confidence scores are color-coded to enable quickly focusing on the sentences that require more attention. Green denotes sentences for which the translation confidence is high, and are thus expected to produce good translations. Red marks sentences that are estimated to be poorly translated, and all those in between are marked with an orange label.

We attempt to suggest rewritings only for sentences that are estimated to be not so well translated. When we are able to propose rewriting(s) with higher translation confidence than the original, a magnifying glass icon is displayed next to the sentence. Clicking it displays, on the right side of the screen, an ordered list of the more confident rewritings, along with their corresponding confidence estimations. The first sentence on the list is always the original one, to let it be edited, and to make it easier to view the difference between the original and the rewritings. An example is shown on the right side of Figure 1, where we see a rewriting suggestion for the fourth sentence in the document. Here, the suggestion is simply to replace the word *captured* with the word *caught*, a rewriting that is estimated to improve the translation of the sentence.

The user can select one of the suggestions or choose to edit either the original or one of the rewritings. The current sentence which is being examined is marked with a different color and the alternative under focus is marked with a small icon (the bidirectional arrows). The differences between the alternatives and the original are highlighted. After the user's confirmation (with the check mark icon), the display of the document on the left-hand side is updated based on her selection, including the updated confidence estimation. At any time, the user (if she speaks the target language) can click on the cogwheel icon and view the translation of the source or of its rewritten version. When done, the user can save the edited text or its translation. Moses Release 1.0 of an English-Spanish Europarl-trained model[4] was used in this work to obtain English-Spanish translations.

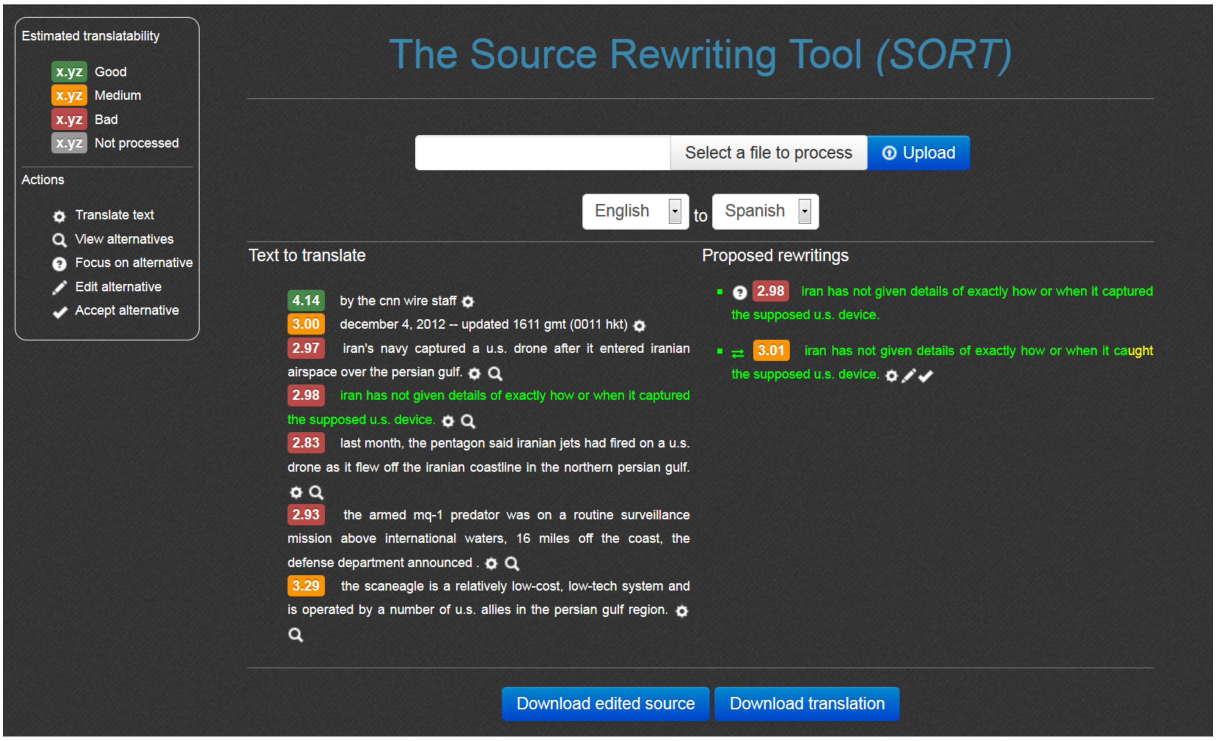
2.1 System and software architecture

SORT is implemented as a web application, using an MVC (Model View Controller) software architecture. The *Model* part is formed by Java classes representing the application state (user input, selected text lines, associated rewriting propositions and scores). The *Controller* consists of several servlet components handling each user interaction with the backend server (file uploads, SMT tools calls via XML-RPC or use of the embedded Java library that handles the actual rewritings). Finally, the *View* is built with standard web technologies: HTML5, JavaScript (AJAX) and CSS style sheets. The application was developed and deployed on Linux (CentOS release 6.4), with a Java Runtime 6 (Java HotSpot 64-Bit Server VM), within a Tomcat 7.0 Application Server, and tested with Firefox as the web client both on Linux and Windows 7.

Figure 2 shows the system architecture of SORT,

[3] Also known as *confidence estimation*.

[4] http://www.statmt.org/moses/RELEASE-1.0/model/



Estimated translatability

- x.yz Good
- x.yz Medium
- x.yz Bad
- x.yz Not processed

Actions

- ⚙ Translate text
- 🔍 View alternatives
- ? Focus on alternative
- ✎ Edit alternative
- ✓ Accept alternative

The Source Rewriting Tool *(SORT)*

Select a file to process | Upload

English to Spanish

Text to translate

4.14 by the cnn wire staff ⚙

3.00 december 4, 2012 -- updated 1611 gmt (0011 hkt) ⚙

2.97 iran's navy captured a u.s. drone after it entered iranian airspace over the persian gulf. ⚙ 🔍

2.98 iran has not given details of exactly how or when it captured the supposed u.s. device. ⚙ 🔍

2.83 last month, the pentagon said iranian jets had fired on a u.s. drone as it flew off the iranian coastline in the northern persian gulf. ⚙ 🔍

2.93 the armed mq-1 predator was on a routine surveillance mission above international waters, 16 miles off the coast, the defense department announced . ⚙ 🔍

3.29 the scaneagle is a relatively low-cost, low-tech system and is operated by a number of u.s. allies in the persian gulf region. ⚙ 🔍

Proposed rewritings

- ? 2.98 iran has not given details of exactly how or when it captured the supposed u.s. device.
- ⇄ 3.01 iran has not given details of exactly how or when it caught the supposed u.s. device. ⚙ ✎ ✓

Download edited source | Download translation

Figure 1: SORT's interface

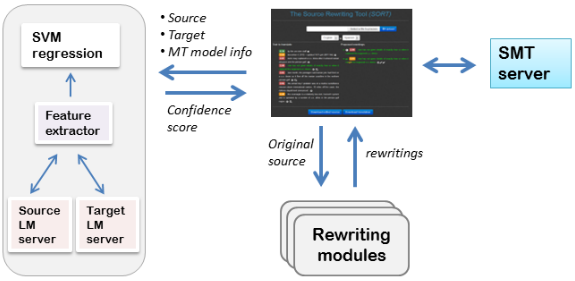


Figure 2: SORT's system architecture. For simplicity, only partial input-output details are shown.

with some details of the current implementation. The entire process is performed via a client-server architecture in order to provide responsiveness, as required in an interactive system. The user communicates with the system through the interface shown in Figure 1. When a document is loaded, its sentences are translated in parallel by an SMT Moses server (Koehn et al., 2007). Then, the source and the target are sent to the confidence estimator, and the translation model information is also made available to it. The confidence estimator extracts features from that input and returns a confidence score. Specifically, the language model features are computed with two SRILM servers (Stolcke, 2002), one for the source language and one for the target language. Rewritings are produced by the rewriting modules (see Section 3 for the implemented rewriting methods). For each rewriting, the same process of translation and confidence estimation is performed. Translations are cached during the session; thus, when the user wishes to view a translation or download the translations of the entire document, the response is immediate.

3 Source rewriting

Various methods can be used to rewrite a source text. In what follows we describe two rewriting methods, based on *Text Simplification* techniques, which we implemented and integrated in the current version of SORT. Simplification operations include the replacement of words by simpler ones, removal of complicated syntactic structures, shortening of sentences etc. (Feng, 2008). Our assumption is that simpler sentences are more likely to yield higher quality translations. Clearly, this is not always the case; yet, we leave this decision to the confidence estimation component.

Sentence-level simplification (Specia, 2010) has proposed to model text simplification as a Statistical Machine Translation (SMT) task where the goal is to translate sentences to their simplified version in the *same* language. In this approach, a simplification model is learnt from a parallel corpus of texts and their simplified versions. Applying this method, we train an SMT model from English to Simple English, based on the PWKP parallel corpus generated from Wikipedia (Zhu et al., 2010);[5] we use only alignments involving a single sentence on each side. This results in a phrase table containing many entries where source and target phrases are identical, but also phrase-pairs that are mapping complex phrases to their simplified counterparts, such as the following:

- *due to its location on → because it was on*
- *primarily dry and secondarily cold → both cold and dry*
- *the high mountainous alps → the alps*

Also, the language model is trained with Simple English sentences to encourage the generation of simpler texts. Given a source text, it is translated to its simpler version, and its n-best translations are assessed by the confidence estimation component.

Lexical simplification One of the primary operations for text-simplification is lexical substitution (Table 2 in (Specia, 2010)). Hence, in addition to rewriting a full sentence using the previous technique, we implemented a second method, addressing lexical simplification directly, and only modifying local aspects of the source sentence. The approach here is to extract relevant synonyms from our trained SMT model of English to Simplified English, and use them as substitutions to simplify new sentences. We extract all single token mappings from the phrase table of the trained model, removing punctuations, numbers and stop-words. We check whether their lemmas were synonyms in WordNet (Fellbaum, 1998) (with all possible parts-of-speech as this information was not available in the SMT model). Only those are left as valid substitution pairs. When a match of an English word is found in the source sentence it is replaced with its simpler synonym to generate an alternative for the source. For example, using this rewriting method for the source sentence "*Why the Galileo research program* **superseded rival** *programs,*" three rewritings of the sentence are generated when *rival* is substituted by *competitor* or *superseded* by *replaced*, and when both substitutions occur together.

In the current version of SORT, both sentence-level and lexical simplification methods are used in conjunction to suggest rewritings for sentences with low confidence scores.

4 Confidence estimation

Our confidence estimator is based on the system and data provided for the 2012 *Quality estimation shared task* (Callison-Burch et al., 2012). In this task, participants were required to estimate the quality of automated translations. Their estimates were compared to human scores of the translation which referred to the suitability of the translation for post-editing. The scores ranged from 1 to 5, where 1 corresponded to translation that practically needs to be done from scratch, and 5 to translations that requires little to no editing.

The task's training set consisted of approximately 1800 source sentences in English, their Moses translations to Spanish and the scores given to the translations by the three judges. With this data we trained an SVM regression model using SVM^{light} (Joachims, 1999). Features were extracted with the task's feature-extraction baseline module. Two types of features are used in this module (i) *black-box* features, which do not assume access to the translation system, such as the length of the source and the target, number of punctuation marks and language model probabilities, and (ii) *glass-box* features, which are extracted from the translation model, such as the average number of translations per source word (Specia et al., 2009).

5 Initial evaluation and analysis

We performed an initial evaluation of our approach in an English to Spanish translation setting, using the 2008 News Commentary data.[6] First, two annotators who speak English but not Spanish used SORT to rewrite an English text. They reviewed the proposed rewritings for 960 sentences and were instructed to "trust the judgment" of the confidence estimator; that is, reviewing the suggestions from the most to the least confident one, they accepted the first rewriting that was fluent and preserved the meaning of the source document as a whole. 440 pairs of the original sentence and the selected alternative were then both translated to Spanish and were presented as competitors to

[5] Downloaded from: http://www.ukp.tu-darmstadt.de/data/sentence-simplification

[6] Available at http://www.statmt.org

three native Spanish speakers. The sentences were placed within their context in the original document, taken from the Spanish side of the corpus. The order of presentation of the two competitors was random. In this evaluation, the translation of the original was preferred 20.6% of the cases, the rewriting 30.4% of them, and for 49% of the sentences, no clear winner was chosen.[7] Among the two rewriting methods, the sentence-level method more often resulted in preferred translations.

These results suggest that rewriting is estimated to improve translation quality. However, the amount of preferred original translations indicates that the confidence estimator is not always discriminative enough: by construction, for every rewriting that is displayed, the confidence component estimates the translation of the original to be less accurate than that of the rewriting; yet, this is not always reflected in the preferences of the evaluators. On a different dimension than translation quality, the large number of cases with no clear winner, and the analysis we conducted, indicate that the user's cognitive effort would be decreased if we only displayed those rewritings associated with a substantial improvement in confidence; due to the nature of our methods, frequently, identical or near-identical translations were generated, with only marginal differences in confidence, e.g., when two source synonyms were translated to the same target word. Also, often a wrong synonym was suggested as a replacement for a word (e.g. *Christmas air* for *Christmas atmosphere*). This was somewhat surprising as we had expected the language model features of the confidence estimator to help removing these cases. While they were filtered by the English-speaking users, and thus did not present a problem for translation, they created unnecessary workload. Putting more emphasis on context features in the confidence estimation or explicitly verifying context-suitability of a lexical substitutions could help addressing this issue.

6 Related work

Some related approaches focus on the authoring process and control *a priori* the range of possible texts, either by interactively enforcing lexical and syntactic constraints on the source that simplify the operations of a rule-based translation system (Carbonell et al., 1997), or by semantically guiding a monolingual author in the generation of multilingual texts (Power and Scott, 1998; Dymetman et al., 2000). A recent approach (Venkatapathy and Mirkin, 2012) proposes an authoring tool that consults the MT system itself to propose phrases that should be used during composition to obtain better translations. All these methods address the authoring of the source text from scratch. This is inherently different from the objective of our work where an existing text is modified to improve its translatability. Moving away from authoring approaches, (Choumane et al., 2005) propose an interactive system where the author helps a rule-based translation system disambiguate a source text inside a structured document editor. The techniques are generic and are not automatically adapted to a specific MT system or model. Closer to our approach of modifying the source text, one approach is to paraphrase the source or to generate sentences entailed by it (Callison-Burch et al., 2006; Mirkin et al., 2009; Marton et al., 2009; Aziz et al., 2010). These works, however, focus on handling out-of-vocabulary (OOV) words, do not assess the translatability of the source sentence and are not interactive.[8] The MonoTrans2 project (Hu et al., 2011) proposes monolingual-based editing for translation. Monolingual speakers of the source and target language collaborate to improve the translation. Unlike our approach, here both the feedback for poorly translated sentences and the actual modification of the source is done by humans. This contrasts with the automatic handling (albeit less accurate) of both these tasks in our work.

7 Conclusions and future work

We introduced a system for rewriting texts for translation under the control of a confidence estimator. While we focused on an interactive mode, where a monolingual user is asked to check the quality of the source reformulations, in an extension of this approach, the quality of the reformulations could also be assessed automatically, removing the interactive aspects at the cost of an increased risk of rewriting errors. For future work we wish to add more powerful rewriting techniques that are able to explore a larger space of possible reformulations, but compensate this ex-

[7] One should consider these figures with caution, as the numbers may be too small to be statistically meaningful.

[8] Another way to use paraphrases for improved translation has been proposed by (Max, 2010) who uses paraphrasing of the source text to increase the number of training examples for the SMT system.

panded space by robust filtering methods. Based on an evaluation of the quality of the generated alternatives as well as on user selection decisions, we may be able to learn a quality estimator for the rewriting operations themselves. Such methods could be useful both in an interactive mode, to minimize the effort of the monolingual source user, as well as in an automatic mode, to avoid misinterpretation. In this work we used an available baseline feature extraction module for confidence estimation. A better estimator could benefit our system significantly, as we argued above. Lastly, we wish to further improve the user interface of the tool, based on feedback from actual users.

References

[Aziz et al.2010] Wilker Aziz, Marc Dymetman, Shachar Mirkin, Lucia Specia, Nicola Cancedda, and Ido Dagan. 2010. Learning an expert from human annotations in statistical machine translation: the case of out-of-vocabularywords. In *Proceedings of EAMT*.

[Callison-Burch et al.2006] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of HLT-NAACL*.

[Callison-Burch et al.2012] Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of WMT*.

[Carbonell et al.1997] Jaime G Carbonell, Sharlene L Gallup, Timothy J Harris, James W Higdon, Dennis A Hill, David C Hudson, David Nasjleti, Mervin L Rennich, Peggy M Andersen, Michael M Bauer, et al. 1997. Integrated authoring and translation system. US Patent 5,677,835.

[Choumane et al.2005] Ali Choumane, Hervé Blanchon, and Cécile Roisin. 2005. Integrating translation services within a structured editor. In *Proceedings of the ACM symposium on Document engineering*. ACM.

[Dymetman et al.2000] Marc Dymetman, Veronika Lux, and Aarne Ranta. 2000. Xml and multilingual document authoring: Convergent trends. In *Proceedings of COLING*.

[Fellbaum1998] Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

[Feng2008] Lijun Feng. 2008. Text simplification: A survey. Technical report, CUNY.

[Hu et al.2011] Chang Hu, Philip Resnik, Yakov Kronrod, Vladimir Eidelman, Olivia Buzek, and Benjamin B. Bederson. 2011. The value of monolingual crowdsourcing in a real-world translation scenario: simulation using haitian creole emergency sms messages. In *Proceedings of WMT*.

[Joachims1999] T. Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press.

[Koehn et al.2007] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL, Demo and Poster Sessions*.

[Marton et al.2009] Yuval Marton, Chris Callison-Burch, and Philip Resnik. 2009. Improved statistical machine translation using monolingually-derived paraphrases. In *Proceedings of EMNLP*.

[Max2010] Aurélien Max. 2010. Example-based paraphrasing for improved phrase-based statistical machine translation. In *Proceedings of EMNLP*.

[Mirkin et al.2009] Shachar Mirkin, Lucia Specia, Nicola Cancedda, Ido Dagan, Marc Dymetman, and Idan Szpektor. 2009. Source-language entailment modeling for translating unknown terms. In *Proceedings of ACL-IJCNLP*.

[O'Brien2006] Sharon O'Brien. 2006. Controlled Language and Post-Editing. *Multilingual*, 17(7):17–19.

[Power and Scott1998] Richard Power and Donia Scott. 1998. Multilingual authoring using feedback texts. In *Proceedings of ACL*.

[Specia et al.2009] Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of EAMT*.

[Specia2010] Lucia Specia. 2010. Translating from complex to simplified sentences. In *Proceedings of PROPOR*.

[Stolcke2002] Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *INTERSPEECH*.

[Venkatapathy and Mirkin2012] Sriram Venkatapathy and Shachar Mirkin. 2012. An SMT-driven authoring tool. In *Proceedings of COLING 2012: Demonstration Papers*.

[Zhu et al.2010] Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of COLING*.