Reassessing the proper place of man and machine in translation: a pre-translation scenario

Julia Ive, Aurélien Max, François Yvon

Reassessing the Proper Place of Man and Machine in Translation: A Pre-Translation Scenario*

Julia Ive Aurélien Max
François Yvon

LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91 403 Orsay, France

first.last@limsi.fr

Abstract

Traditionally, human-machine interaction to reach an improved Machine Translation (MT) output takes place *ex-post* and consists in correcting this output. In this work, we investigate other modes of intervention in the MT process. We propose a Pre-Edition (PRE) protocol that involves: (a) the detection of MT translation difficulties; (b) the resolution of those difficulties by a human translator, who provides their translations (pre-translation); (c) the integration of the obtained information prior to the automatic translation. This approach can meet individual interaction preferences of certain translators and can be particularly useful for production environments, where more control over output quality is needed. Early resolution of translation difficulties can prevent downstream errors, thus improving the final translation quality for "free". We show that translation difficulty can be reliably predicted for English for various source units. We demonstrate that the pre-translation information can be successfully exploited by an MT system and that the indirect effects are genuine, accounting for around 16% of the total improvement. We also provide a study of the human effort involved in the resolution process.

*Preprint of a paper published as: Julia Ive, Aurélien Max and François Yvon (2018). "Reassessing the proper place of man and machine in translation: a pre-translation scenario". *Machine Translation* journal. https://doi.org/10.1007/s10590-018-9223-9.

1 Introduction

In the recent years, Machine Translation (MT) has made a lot of progress and continues to improve. It is successfully used in many situations, including professional translation environments and production scenarios. However, the translation process is very complex and is difficult to formalize. It requires vaster knowledge than can be captured by machines, even when provided with very large quantities of translated texts. This means that for the majority of setups, MT requires at least some human intervention to reach publishable quality.

Traditionally, human-machine interaction during the MT process is reduced to the following procedure: the MT output serves as a draft that is then manually corrected (Computer-Assisted Translation, CAT). This process is usually referred to as Post-Edition (PE). A substantial part of the current research in CAT is dedicated to studying PE in order to make this process less demanding in terms of human effort. This research is performed mainly within the *Adaptive MT* approach (Koehn et al, 2013; Federico et al, 2013; Denkowski et al, 2014), where an MT system constantly adapts itself to the flow of the human PE feedback.

In this work, we study an alternative mode of collaboration, where the human intervention happens prior to machine translation. Our interaction protocol involves the following three steps: (1) automatic detection of fragments of the source text that could be problematic for the MT system; (2) resolution of these difficulties by a human expert, who provides the system with the expected translation of these segments; (3) exploitation of this information by the MT system. Such an *ex-ante* human intervention in MT is referred to as Pre-Edition (PRE).

Steps (1)-(3) reproduce the typical behavior of a human translator: he or she will first analyze the source text, detect its parts that will be difficult to translate, then consult external sources of information to resolve difficulties before translating. This suggests that our protocol may meet individual preferences of some translators and can be proposed as an alternative or as a complement to Post-Edition (Kay, 1997).

There are other arguments in favor of PRE: it gives the human expert more control over the MT process, as it guarantees that some erroneous lexical choices or false senses will not appear in the target text. Furthermore, when difficult segments are repeated across a long document, their correct translation will be entered only once, where PE would imply multiple corrections of the same segment.[1] Finally, it is expected that constraining the translation process with human suggestions will also improve the machine output in the neighborhood of these

[1] A problem that no longer occurs when PE is complemented with adaptive learning. However, in such scenarios the human is usually not asked explicitly to provide generalized translations for all the corresponding contexts. Therefore, the corrections of recurrent translations may still be required.

correct segments, resulting in indirect improvements of the MT that happen, as it were, for free.

To better assess the real potential of PRE, we seek to answer in this work the following questions: (a) Can translation difficulties be reliably identified in step (1) ? Should difficulties be detected at the level of words or phrases ? (b) Can the human translations provided for difficult-to-translate units be successfully exploited by an MT system ? How significant are indirect improvements ? What is their nature ? (c) How realistic is the pre-translation protocol in terms of the human effort involved ? A last question, which may be worth asking in multi-source translation scenarios, finally is (d) Are some source difficulties common to several target languages ? Or are they specific to each language pair ?

We start by describing our notion of translation difficulty in Section 2. We consider several types of segmentations and experiment with multiple sets of features to identify how to best predict translation difficulties. Our results are detailled in Section 3, where we also analyze source-side translation difficulties for multiple language pairs sharing the same source text. All these experiments suggest that difficult segments can be quite reliably detected in the source text. We then introduce our protocol for resolving translation difficulties in Section 4, and leverage our automatic predictions to perform an extrinsic evaluation of this resolution protocol. Even though we use simulated, rather than actual, human assistance, our results suggest that pre-translation can be effective and that the indirect improvements are genuine. In Section 5, we finally present our estimation of the human effort involved in pre-translation. We conclude by discussing related work in Section 6 and summarizing our main findings in Section 7.

2 Methodology

The notion of subsentential translation difficulty as a system-related measure was first introduced by Mohit and Hwa (2007). In our attempt to detect difficulties, we largely follow their specification of the problem and their resolution protocol. Their formulation makes the task quite similar to Quality Estimation (QE): where QE detects difficulties or potential errors at the phrase or word level on the target side, we try to perform this detection on the source side. Based on this analogy, we extend the work of Mohit and Hwa (2007) with state-of-the-art methods borrowed from the QE literature.

2.1 Problem Statement and Notations

Following Mohit and Hwa (2007), we cast source translation difficulty prediction as a classification problem, where words or phrases are tagged as either difficult-

to-translate (DT) or easy-to-translate (ET).

To generate training segmentations and labelings for the classifier, we start with pairs of parallel sentences $(\mathbf{f}, \mathbf{e})$, assuming that a Part-of-Speech (POS) tagging and a shallow parse of **f** are also available.

The translation difficulties we want to predict are system-dependent. We thus assume access to an MT system that generates some "draft" output $\mathbf{e}^1$ (the 1-best translation hypothesis) for a source sentence **f** with reference translation $\hat{\mathbf{e}}$. We also assume an alignment $\mathbf{a}^1$ between the words f_i of the source sentence and the words e^1_j in the 1-best hypothesis. This word alignment also enables to derive the 1-best translation $e^1_r \cdots e^1_g = e^1_{[r:g]}$ of arbitrary segments $f_{[k:t]}$ of the source sentence.

We believe that these assumptions are realistic and that our methodology could be applied to any MT system. Statistical MT decoders generate a phrase alignment that can readily be turned into a word alignment. For other architectures, such alignment could be obtained *ex-post*: either by realigning the output translation or, for neural MT decoders, by using the values of the attention layer (Bahdanau et al, 2014) as a proxy to deterministic alignments.

We describe below the preprocessing procedure used to label the training data as well as the various sets of features considered and classification methods used in our experiments.

2.2 Generating Gold Annotations

Generating labels Knowing both $\hat{\mathbf{e}}$ and the alignments $\mathbf{a}^1$ between **f** and $\mathbf{e}^1$, we label the training data using the word alignment $\mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$ computed by Translation Edit Rate (TER) (Snover et al, 2006). In this alignment, each word-to-word connection is labeled with a type of post-editing operation: match (M), shift (SH), substitution (S) or deletion (D) applied to e^1_j to obtain $\hat{e}_m$. From the resulting 3-way alignment $\mathbf{f} \rightarrow \mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$, these post-editing labels can be projected back onto each token f_i as follows:

- if f_i only aligns with words labeled as *M*, f_i is labeled as ET;
- in all other cases, f_i is marked as DT.

These word-level labels are extended into segment-level labels as follows. Assuming a segmentation π of **f**, we label each segment $f_{[k:t]}$ as DT if at least 50% of $f_k \cdots f_t$ are labeled as DT; the remaining segments are labeled as ET.

This annotation procedure is illustrated in Figures 1 and 2.

Defining Segments We will contrast 3 strategies to define the source segmentation π: the segmentation into individual words (WORD-SEG); the segmentation

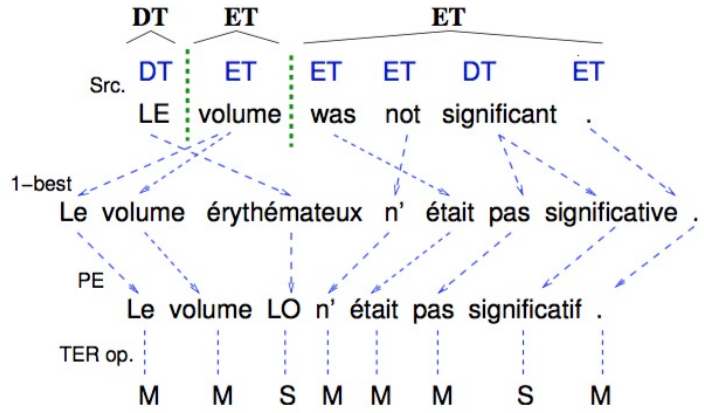


Figure 1: Labeling the sentence "LE volume was not significant". Segment boundaries are marked with dashed lines.

induced by the MT phrase-based decoder (MT-SEG); and a syntactically-motivated segmentation obtained by shallow parsing (SYNT-SEG).

Discussion We have chosen to keep the automatic labeling process simple and unambiguous, keeping in mind that the resulting labels will not be entirely correct. A first source of error is the noise in the two alignment processes necessary to compute the word labels: in our experiments, the use of a phrase-based system will ensure that at least the source-target alignment is correct.

Another defect of our procedure concerns the labeling of source words that are not aligned: for these words we had to make a rather arbitrary choice and label them as `DT`, where the alternative label could also have been justified.

Third, our labeling scheme makes no distinction between translation difficulties and *reordering difficulties*. Such difficulties can be associated with words marked with a shift (SH) operation. However, those words are badly defined and cannot be reliably detected (note that they account for less than 5% of all words in our experiments).

The policy of labeling segments according to the percentage of correctly translated words is also rather naive as it does not take the semantic content of words into account. For instance, a segment made of a determiner and a noun will be marked as `DT` if only one of its words is `DT`; this label may be fine when the noun is `DT`, much more debatable if the error is on the determiner.

Concerning the segmentation types, we believe that SYNT-SEG has more potential to "correctly" handle difficulties in labeling segments. WORD-SEG is prone to tokenization errors and does not always operate with minimal sense-bearing units. MT-SEG segments are idiosyncratic, whereas SYNT-SEG chunks are con-

sistent. Moreover, using SYNT-SEG ensures that human translators will provide translations for grammatical phrases, which is probably easier than doing so for random chunks of words. Figure 2 illustrates two labelings of a sentence using MT-SEG and SYNT-SEG.

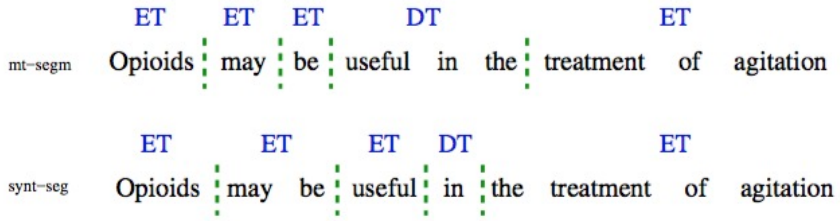


Figure 2: Labeling the sentence "Opioids may be useful in the treatment of agitation" segmented with MT-SEG and SYNT-SEG.

2.3 Features

The prediction of difficulty labels relies on a feature-based representation of the input sentence. Our main source of inspiration for designing these word- and phrase-level features is the annual QE shared task of the Workshop on Machine Translation (WMT) (Bojar et al, 2017).

Note that in our case, feature extraction uses both source and target side information. For QE, the majority of features is typically extracted from MT output. In our setting, we instead extract most features from MT *input*.

At the *word level*, we distinguish 27 *black-box* (system-independent) features and 1 *glass-box* (system-dependent) feature. The former include the following groups of word-level features: (a) 3 *basic features* (`bs`): the word f_i, its lemma and its POS tag; (b) 20 *standard features* (`st`): this is the baseline set of the WMT'16 word-level QE task (Bojar et al, 2016); (c) 3 *syntactic features* (`snt`): the shallow parsing tag, the input dependency label of f_i and the depth of f_i in the dependency tree, measured as the distance from the root; (d) for the MEDICAL domain, we add one extra *term feature* (`trm`): a binary feature indicating whether a word is a term or a part of a compound term. The term mapping was performed with the `Metamap` tool for medical texts (Aronson and Lang, 2010). We also extract the number of possible translations of f_i, as defined by the lexical translation probability model of a system, as a glass-box feature (`gl`).

At the *phrase level*, we extract 72 **black-box** features and 1 **glass-box** feature for each segment $f_{[k:t]}$: (a) 8 *basic features*: the sequence of words, their hypothesis translation, lemmas and POS, plus the left (f_{k-2}, f_{k-1}) and right (f_{t+1}, f_{t+2}) contexts; (b) 59 *standard phrase-level features*: phrase-level features used in the

WMT'16 QE task, excluding the majority of POS features; to these, we add the target POS tag sequence of the aligned $e_{[r:g]}$; (c) 4 *syntactic features*: the constituency label covering the longest span of $f_{[k:t]}$, the percentage of words whose syntactic heads are outside the boundaries of $f_{[k:t]}$, and the maximum and minimum depth of a $f_r, k \leq r \leq t$ in the dependency tree; (d) 1 *term feature*: the percentage of $f_r, k \leq r \leq t$ marked as terms, computed as described for the word-level feature. The only glass-box feature is the percentage of OOV words in $f_{[k:t]}$.

2.4 Classification Algorithms

Our baseline classifier is Support Vector Machines (SVMs) (Cortes and Vapnik, 1995), as in (Mohit and Hwa, 2007). We also experiment with Conditional Random Fields (CRFs) (Lafferty et al, 2001), Random Forests (RFs) (Breiman, 2001) and Feedforward Neural Networks (FFNNs). CRFs are a state-of-the-art solution in many sequence labeling tasks since they take label dependencies into account. Given that translations of words are interdependent, it is expected that word-level labels will also influence each other, making CRFs thus a natural pick for the word level prediction task. Phrases are supposed to partly capture dependencies between the words they contain, their translations are less interdependent; this suggests other algorithms such as RFs and FFNNs that handle each instance independently may be more effective.

3 Detection of Translation Difficulties

In this section, we will evaluate our methodology for detecting translation difficulties on two tasks: first on a English-to-French translation task in the MEDICAL domain, then on a multi-target translation task with English as the source language, in the United Nation (UN) domain. For the MEDICAL domain, we report the results of an intrinsic evaluation of our classifiers contrasting 3 types of segmentation, each corresponding to the development of a dedicated classifier: 1 for the word-level prediction (WORD-SEG) and 2 for phrase-level predictions (MT-SEG and SYNT-SEG, cf. section 3.1). For the UN domain, we focus only on SYNT-SEG, as it is the most promising segmentation strategy (cf. the discussion of section 2.2); being independent of the MT system, it also makes the comparison between various target languages much easier and will allow us to study study difficulties that are common to several target languages (cf. section 3.2).

3.1 Experiments in the MEDICAL Domain

3.1.1 Data

For this set of experiments, we used the data provided with the WMT'14 medical task (*train-MT*) to train a phrase-based MT system.[2]

We built the classifier training data (*train-classif*), as well as the development and test sets from an in-domain Cochrane English-French corpus of medical review abstracts (Ive et al, 2016). Reference translations for this corpus are post-edited MT translations. The data were created in a narrowly-specialized production setting, where post-editing was performed by domain specialists and the resulting work was reviewed by a professional translator. The development and test sets were used both for MT and classification evaluation (cf. Table 1). The classifier training data was annotated as described in section 2.2 (cf. Table 2). POS tagging and shallow parsing were performed respectively using the `Stanford POS Tagger` (Toutanova et al, 2003) and the `OpenNLP` toolkit.[3]

set	lines	#tok.(en)	#tok.(fr)
train-MT	4.47M	66M	76M
train-classif	15K	344K	430K
dev	832	19K	26K
test	831	21K	26K

Table 1: Basic corpus statistics for MEDICAL. # denotes count.

strategy	#	$\bar{l}$	DT
WORD-SEG	344K	1	42%
MT-SEG	206K	1.7	50%
SYNT-SEG	210K	1.6	44%

Table 2: Statistics for the annotated training data (*train-classif*) for MEDICAL ($\bar{l}$ denotes average segment length, # denotes count).

Figure 3 displays the distribution of DT and ET segments according to the percentage of DT words they contain, for both SYNT-SEG and MT-SEG. This figure shows that the vast majority of segments are unambiguously labelled as ET or DT; on average only 9% of all the SYNT-SEG and MT-SEG segments contain exactly 50% of DT words, which is the borderline case. Thus, we consider that our labeling procedure is only slightly biased by the approximations we use to define DT segments.

[2] http://statmt.org/wmt14/medical-task

[3] http://opennlp.apache.org/index.html.

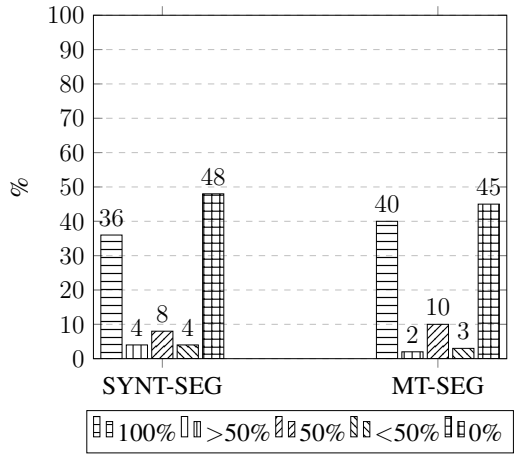


Figure 3: Distribution of source segments in the test set according to the percentage of DT words they contain: MEDICAL domain.

3.1.2 System Building

We built an English-French phrase-based MT (PBMT) system with Moses (Koehn et al, 2007); word alignments were computed using fast_align (Dyer et al, 2013). We trained a 6-gram language model with modified Kneser-Ney smoothing on the French part of the MT training data using the SRILM toolkit (Stolcke, 2002). The MT was tuned with kb-mira and 300-best lists (Cherry and Foster, 2012). This baseline system achieved 26.48 BLEU (Papineni et al, 2002) and 0.59 TER on our test data.

3.1.3 Detecting Translation Difficulties

Our experiments used the RF[4] and SVM[5] implementations available in Scikit-learn (Pedregosa et al, 2011), as well as the implementation of CRFs[6] available in Wapiti (Lavergne et al, 2010). We built a Feedforward Neural Network with 2 hidden layers. For the word and phrase features representing source and target sequences, as well their contexts, we used pre-trained 300-dimensional word em-

[4]Grid search with 5-fold cross-validation was used to tune the following parameters towards F-score: the optimizing criterion, the number of estimators, the maximum depth and the minimum number of leaf samples. All other parameters are those provided by default.

[5]We use a Radial Basis Function (RBF) kernel. Grid search with 5-fold cross-validation was used to tune γ and C towards F-score. All other parameters are those provided by default.

[6]We used the l-bfgs algorithm as the optimization algorithm. All other parameters are those provided by default. All the hyperparameters were tuned on the development set.

beddings (Mikolov et al, 2013).[7] All the layers besides the output layer used the `relu` function as the activation function. The output layer uses the `sigmoid` function as the activation function. The model is trained to minimize the binary cross-entropy loss using the Adam optimizer (Kingma and Ba, 2014). We built our network using the `Keras` toolkit. [8] All the hyperparameters were tuned on the development set. The best performance was achieved after 10 epochs.

To handle instability of RFs and FFNNs, each classification experiment was run 10 times. We provide averaged results. Finally, we used the random classifier implementation available in `Scikit-learn` to create a baseline.[9]

The standard word-level and phrase-level sets of features were extracted using `Marmot` (Logacheva et al, 2016) and `Quest++` (Specia et al, 2013). Lemmatization was performed using the `TreeTagger` (Schmid, 1995).

Hereinafter, we report evaluation results using standard metrics: the F-score per class (F_{DT} and F_{ET}), as well as the macro-averaged F-score (F_{mcr}).

	Random		SVMs		CRFs		RFs		FFNNs	
	F_{DT}	F_{ET}	F_{DT}	F_{ET}	F_{DT}	F_{ET}	F_{DT}	F_{ET}	F_{DT}	F_{ET}
WORD-SEG	0.43	0.57	0.58	0.68	0.73	0.79	0.72	0.80	0.71	0.78
MT-SEG	0.51	0.49	0.68	0.56	0.75	0.74	**0.78**	**0.77**	0.76	0.75
SYNT-SEG	0.39	0.51	0.53	0.65	0.72	0.76	**0.74**	**0.79**	0.72	0.76

Table 3: Performance of the automatic detection of translation difficulties.

Experiments with various classification algorithms are reported in Table 3. RFs, FFNNs and CRFs obtain similar performance, systematically outperforming SVMs: for instance, for SYNT-SEG, the prediction quality for SVMs is about 0.16 points below the averaged F_{mcr} of 0.75 achieved with CRFs, RFs and FFNNs.

For the phrase-level segmentations, RFs are slightly better than CRFs and FFNNs, yielding an improvement of 0.03 in F_{mcr} for MT-SEG. RFs do not only improve performance, they are also faster and easier to train, since they take segments, instead of sentences, as training examples. This means in particular that

[7] The first hidden layer contained the quantity of units equal to the total quantity of features (the embedding dimension was taken into account, e.g., 1822 hidden units for WORD-SEG), the second layer used twice as less units. The following features served as inputs to embedding layers: word f_i, its left and right context f_{i-1} and f_{i+1}, its aligned word e^1_j, its left and right context e^1_{j-1} and e^1_{j+1} for WORD-SEG; sequence $f_{[k:t]}$, its aligned translation $e_{[r:g]}$ and its the left (f_{k-2}, f_{k-1}) and right (f_{t+1}, f_{t+2}) contexts for phrase-level segmentations. The length of a segment sequence was limited to 10 words, masking was used for shorter segments.

[8] `https://github.com/fchollet/keras`

[9] We used the `DummyClassifier` with default parameters.

the data can be more easily balanced when necessary. For all these reasons, we will use RFs in all our subsequent experiments and for all segmentation types.

	WORD-SEG		MT-SEG		SYNT-SEG	
set	$F_{\texttt{DT}}$	$F_{\texttt{ET}}$	$F_{\texttt{DT}}$	$F_{\texttt{ET}}$	$F_{\texttt{DT}}$	$F_{\texttt{ET}}$
word	-	-	0.76	0.77	0.73	0.79
all	0.72	0.80	0.78	0.77	0.74	0.79
-gl	0.71	0.80	0.78	0.77	0.74	0.79
-trm	0.72	0.80	0.78	0.77	0.74	0.79
-snt	0.72	0.79	0.77	0.77	0.75	0.79
-st	**0.61**	**0.74**	**0.69**	**0.71**	**0.63**	**0.74**
-bs	0.71	0.79	0.77	0.77	0.74	0.79

Table 4: Feature ablation experiments: MEDICAL domain (*word* denotes results for DT and ET labels predicted for WORD-SEG, then projected to the phrase level).

Overall, our evaluations showed that translation difficulties can be reliably identified for all the segmentation strategies (average $F_{mcr} = 0.77$, cf. Table 4). The prediction performance at the word level is similar to the average prediction performance at the phrase level: an average $\Delta F_{mcr} = 0.01$ is observed when we project DT and ET labels predicted for words (i.e., WORD-SEG) to the phrase level.

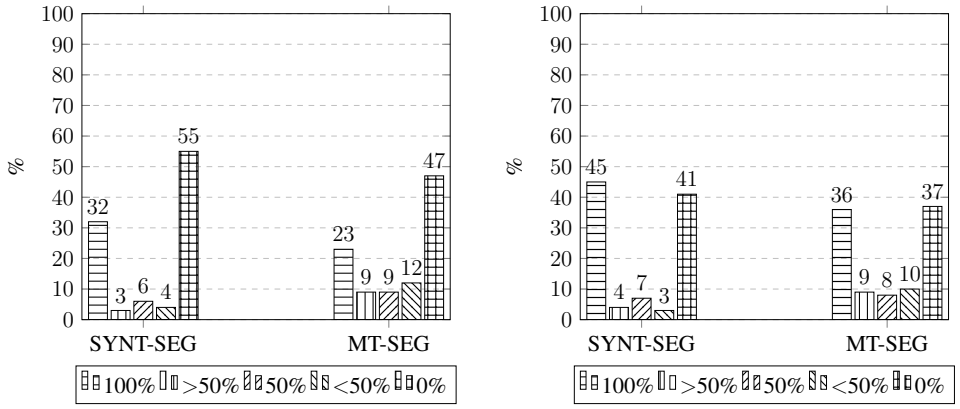
According to the feature ablation experiments reported in Table 4, the set of standard features turned out to be the most helpful for all the segmentation strategies (e.g., $\Delta F_{mcr} = 0.08$ for MT-SEG). The other groups of features were not useful in our setting. The set of features used in further MEDICAL experiments includes the basic and standard features for all the segmentation strategies. For this domain, these features are sufficient to reach an average performance of $F_{mcr} = 0.77$.

	# tok.					
set	lines	EN	AR	ES	FR	RU
MT	5.7M	164M	171M	189M	193M	149M
train	15K	430K	449K	494K	507K	391K
dev	1K	29K	30K	33K	34K	26K
test	1K	29K	30K	33K	34K	26K

Table 5: Basic corpus statistics for UN. # denotes count.

	#	$\bar{l}$	% DT			
			AR	ES	FR	RU
WORD-SEG	420K	1	47	33	38	52
SYNT-SEG	261K	1.6	42	30	35	47
BLEU			38.7	50.3	45.1	36.8
TER			0.51	0.39	0.45	0.54

Table 6: Statistics for the annotated training data (*train-classif*) for UN ($\bar{l}$ is the average segment length, # denotes count).



(a) FR (b) RU

Table 7: Distribution of source segments in the test set according to the percentage of DT words they contain : UN.

3.2 Experiments in the UN Domain

We now assess our methodology for difficulty detection in the United Nation (UN) domain with multiple target languages (Arabic, French, Russian, and Spanish) and try to determine whether this detection can benefit from having access to multiple target languages. The main motivation to study translation difficulty detection in a multilingual context is that this setting opens a range of new perspectives in terms of human effort reduction: for instance, common difficulties can be identified and possibly resolved once and for many languages; segments that are difficult-to-translate for one language can be automatically resolved using translations into other languages, where those segments may be easy-to-translate, etc.

Before such perspectives are explored, a first step consists in evaluating how much difficulties are common to several target language pairs. We will also experiment with features extracted from translations into other languages. Note that we will only consider SYNT-SEG segmentations, as it makes cross-language comparisons much easier (cf. the discussion in Section 2.2).

3.2.1 Data

We used the Arabic (AR), English (EN), French (FR), Russian (RU), and Spanish (ES) parts of the multilingual *MultiUN* parallel corpus (Eisele and Chen, 2010; Tiedemann, 2012), making sure that exactly the same set of source (English) sentences is used in all systems. We applied in-house scripts for cleaning and removing duplicate lines. We used the `TreeTagger` tool (Schmid, 1995) for RU and the `Stanford Tokenizer`[10] tool for all other languages. For AR, as is standard practice, we split complex words in smaller parts using the `Stanford Word Segmenter` (Monroe et al, 2014).

Similarly to the previous experiments, the resulting data were separated into MT and classifier training, development and test sets. The latter two were used both for MT and classification evaluation (cf. Table 5). The classifier training data were prepared as in Section 3.1.1; the corresponding statistics appear in the top part of Table 6.

Observations regarding the percentage of predicted `DT` words in segments produced by SYNT-SEG and MT-SEG confirm our intuition that SYNT-SEG better models translation difficulties. For instance, for RU and FR SYNT-SEG segments are less ambiguous: we observe on average 9% more segments unambiguously labeled as `DT` for SYNT-SEG than for MT-SEG (see Figure 7).

[10] `http://nlp.stanford.edu/software/tokenizer.shtml`

3.2.2 Translation System Building

We built English-{Arabic, French, Russian, Spanish} PBMT systems using the sampling strategy described in (Gong et al, 2014) in order to obtain compact, yet competitive, PTs for our large datasets.

Word alignments were computed using `fast_align` (Dyer et al, 2013). We used 4-gram language models trained with modified Kneser-Ney smoothing on the target parts of the MT training data using the `SRILM` toolkit (Stolcke, 2002) and `kb-mira` with 300-best lists for tuning. The performance of these systems is in Table 6 (bottom part). As expected, we can see that translation quality inversely correlates with the proportion of difficulties in each language pair: `DT` words are indeed much more frequent in RU and AR than in ES and FR.

3.2.3 Features

For the UN setting, we experiment with different feature sets with a special focus on the separation of source and target features. In this configuration we will be able to investigate how much translation difficulty can be attributed to the source language, and for how much difficulty the target language is responsible.

We extracted the following set of 189 **black-box** features per language pair: (a) 33 *source* features (SRC-feat): the group of basic features, excluding the hypothesis translation, 22 source features from the group of standard features, the group of syntactic features; (b) 39 *target* features per language pair (e.g., ES, FR, etc.): the hypothesis translation $e_{[r:g]}$, 37 standard target features, and the glass-box feature.

3.2.4 Analysis of Source Difficulties

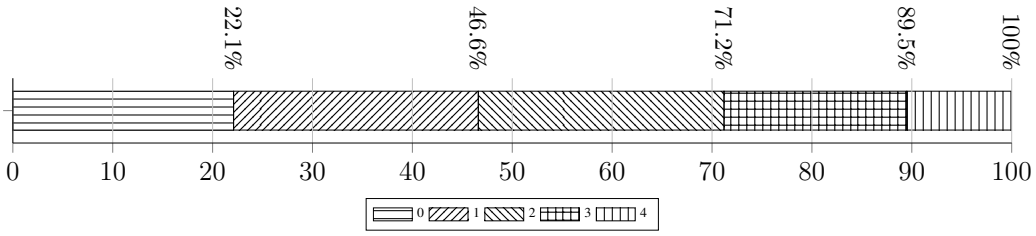


Figure 4: Proportions of difficulties that are shared across languages. About 20% of the word occurrences are easy in all languages; near 25% of the words are difficult in only one language; and a little bit more than 11% are difficult in all four languages.

POS	dep	f_i	chunk
IN	case	the	VP
NN	det	to	NP
DT	amod	of	PP
JJ	nmod	in	O
VBN	root	a	ADVP

Table 8: Most frequent POS, input dependency labels, source words and syntactic chunk labels for words that are difficult in all languages (AR-ES-FR-RU).

lang	AR	ES	FR	RU
AR		21	23	31
ES	21		20	22
FR	23	20		25
RU	31	22	25	

Figure 5: Percentage of common DT words per language pair in the training set (computed over the total number of words).

We now provide quantitative and qualitative observations regarding the annotated difficulties for the four language pairs both at the word and phrase levels by looking at the classifier training corpus.

DT word occurrences common to all four target languages make only around 10.5% of all the source words. Other significant groups of source difficulties are the difficulties of the systems of the "worst" quality (e.g., 9.5% of the source words for RU, 9.4% – common for AR and RU) (see Figure 4).

Note that the distribution for SYNT-SEG is similar. These distributions suggest that the detected difficulties depend more on the language pair than solely on the source language.

These distributions also do not show any significant similarities in translation difficulties even for related target languages (for instance for WORD-SEG, difficulties common to ES and FR make around 20% of all the English occurrences, which is less than the percentage of the difficulties common to both AR and RU (31%), see Figure 5). Another illustration of this is that around 50% of the DT segments for FR are ET for ES (cf. example in Table 10).

Looking more closely at shared difficulties (Tables 8 and 9), we see that highly ambiguous English prepositions ("of", "in", "to", in constructions with the case dependency label), make a substantial part of these recurring difficulties. Other frequent difficulties include the translation of English determiners, as well as of some frequent ambiguous nouns with very general meaning (e.g., "order", "development", "place", with frequencies belonging to the highest quartile in the MT

training corpus).[11]

These observations lead us conclude that only a small quantity of difficulties can be resolved once and for many languages; multi-source and pivoting approaches, where translations into different languages will help each other, open a broader perspective of MT quality improvement in a multi-target setting.

AR-ES-FR-RU	AR-RU	ES-FR
order	situation	implementation
development	programme	progress
place	development	level
action	report	area
work	security	number

Table 9: Examples of the most frequent DT nouns (NN) for different language-pair groups.

EN	the oversight and monitoring of **such places** is adequate
ES ET	la supervisiòn y el seguimiento **de esos lugares** sea adecuada 'the supervision and monitoring **of such sites** is appropriate'
FR DT	la responsabilité du contrôle **et tel lieu** est suffisant 'the responsibility for control **and such place** is sufficient'

Table 10: Example of translation difficulties that are not common for ES and FR.

3.2.5 Detecting Translation Difficulties

We follow the same protocol for training and testing translation difficulty classifiers as in Section 3.1.3.[12] Our results, reported in Table 11 for the Random Forest classifier, are somewhat below (about 0.09 points in F measure) in terms of predicting DT labels to what is observed in the MEDICAL domain. Nonetheless, they show that translation difficulties can be reliably identified for all the translation directions (average $F_{mcr} = 0.70$).

[11] Translations of such nouns tend to vary greatly depending on the context, even when natural translation variability is taken into account. Some of them are also homonymous to verbs, which contributes to their translation difficulty.

[12] We artificially balanced the quantity of examples in both classes for the language pairs where we found an unbalanced proportion of ET and DT (EN-ES, EN-FR) by removing the least frequent examples of ET. This resulted in a reduction of around 34% of the initial training data and a prediction improvement of about 0.07 in F_{DT}.

EN-AR			EN-ES			EN-FR			EN-RU		
feat.	F_{DT}	F_{ET}	feat.	F_{DT}	F_{ET}	feat.	F_{DT}	F_{ET}	feat.	F_{DT}	F_{ET}
SRC-feat	0.590	.72	SRC-feat	0.570	.75	SRC-feat	0.580	.70	SRC-feat	0.690	.66
AR	**0.640**	**.77**	ES	**0.600**	**.78**	FR	**0.610**	**.74**	RU	0.69	**0.75**
word	0.640	.72	*word*	0.600	.74	*word*	0.620	.71	*word*	0.720	.70
all	0.650	.77	**all**	0.610	.77	**all**	0.630	.73	**all**	0.700	.74
-ES	0.650	.77	-AR	0.610	.77	-AR	0.620	.73	-AR	0.700	.74
-FR	0.650	.77	-FR	0.610	.77	-ES	0.620	.73	-ES	0.710	.74
-RU	0.650	.77	-RU	0.610	.77	-RU	0.620	.73	-FR	0.710	.74
-FR-RU	0.650	.64	-FR-RU	0.610	.77	-ES-RU	0.620	.73	-ES-FR	0.710	.74
-ES-RU	0.650	.64	-AR-RU	0.610	.77	-AR-RU	0.620	.74	-AR-FR	0.700	.75
-ES-FR	0.660	.64	-AR-FR	0.610	.77	-AR-ES	0.620	.74	-AR-ES	0.710	.75

Table 11: Feature ablation experiments: SYNT-SEG, UN domain (*word* denotes results for the predicted word-level labels projected to the phrase level).

The set of source features is enough to reach an average performance of $F_{\text{DT}} = 0.61$. In general, adding target features improves prediction accuracy (by $\Delta F_{mcr} = 0.04$) on average for all the language pairs. Adding the features for other target languages does not improve prediction quality, which we believe can be explained by the unsystematic nature of difficulties.

4 Resolution of Translation Difficulties

In the previous sections, we have shown how to detect translation difficulties prior to translation with a relatively fair accuracy, based on an analysis of the source text and its automatic translation. These results were replicated for several language pairs and two domains. The next question is to evaluate how *useful* these results might be for downstream tasks, or for actual translators. To this end, we propose to simulate a human-machine interaction involving three steps: (a) automatic difficulty detection, (b) pre-translation of a portion of potentially difficult segments, (c) constrained machine translation, taking into account the suggestions of step (b). Results of this extrinsic evaluation suggest that the detection of MT difficulties works as expected: fixing segments automatically flagged as difficult yields much larger improvements than fixing the easy ones. Our study enables us to measure the strength of indirect improvements (i.e. improvements of the MT

that indirectly result from having some parts pre-translated).

4.1 Simulating a Human Interaction with Pre-Edition

In order to perform an extrinsic evaluation of our method for detecting translation difficulties, we integrate its results into a resolution protocol involving pre-translation.

We limit ourselves to a rather artificial scenario where difficulties are detected and resolved on a per-sentence level, which will make analysis and comparison with post-editing easier.[13] Performing pre-edition at the level of documents would arguably be more natural for a human translator, and likely more rewarding – we will return to this discussion in Section 5.

Our protocol aims to replicate the following processing and interaction steps (see Figure 6):

1. generate a baseline translation using MT;
2. automatically identify `DT` segments in the input (cf. Section 2);
3. ask the user to provide translations for a certain number of the `DT` segments displayed;[14]
4. generate the final MT, using the translations provided by the human translator as constraints during decoding.

Note that the simulation of user input (step 3) requires automatically obtaining reference translations for (some) `DT` segments in our test set. Here again, we use the word alignments $\mathbf{f} \rightarrow \mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$ produced by TER.[15]

Such a pre-translation scenario may fit personal preferences of certain translators when they interact with the machine (Kay, 1997). Indeed, it loosely mimicks the activity of professional translators: first, analyze the source text, looking for parts that will be difficult for him/her to translate; then consult any available external source of information; finally translate, taking the obtained information into account.

[13] In all the experiments, we will translate complete texts rather than isolated random sentences. Our reference translations are thus likely to be somewhat "normalized", i.e. to contain less translation variety than random sentences.

[14] In a more elaborate version, the user could select these translations from the variants proposed by an MT system (Cheng et al, 2016), or/and from the cache of past translations of `DT` segments, thus potentially saving many keystrokes.

[15] For this experiment, unaligned words in the reference $\hat{e}$ are aligned (recursively) to the same word(s) as their syntactic heads. Dependencies were identified with the help of the `Stanford Parser` toolkit.

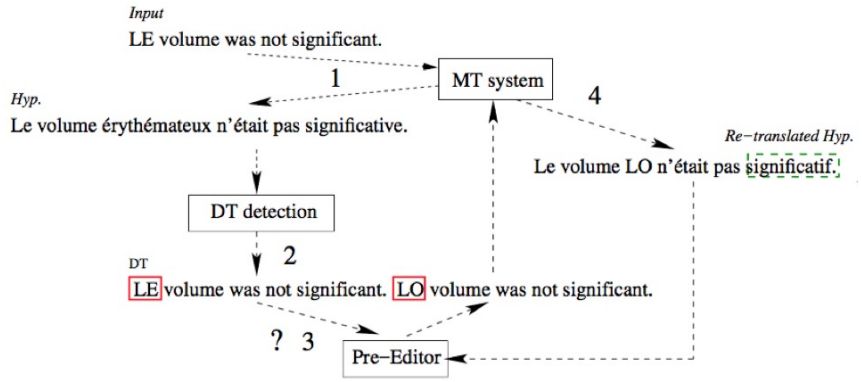


Figure 6: Improving translation through our protocol for difficulty resolution. The red boxed text is pre-translated, the green dashed boxed text is improved for free.

To measure the effectiveness of our DT detection scheme, we study the relationship between the number of pre-translated segments (in step 3) and the overall translation quality. In each sentence, DT words or segments are ranked by their decreasing posterior class probability, meaning that segments that are most likely to be difficult are pre-translated first: we thus expect to see a sharp decrease of translation errors after just a few of these difficult parts are correctly translated. Each experiment assumes that a fixed number p of DT segments is provided, where p varies between 1 and $max_{\texttt{DT}}$, the maximum number of DT segments in a test sentence.

Pre-translation is implemented in our phrase-based translation systems using the `exclusive xml-mode` of the `Moses` decoder.[16] To ensure comparability of the 3 segmentation strategies, for each experiment, we report in our graphs the per sentence *averages* of pre-translated words.[17]

To measure the improvement in translation quality, we mostly used TER, which is the obvious candidate to measure the residual PE effort that would be necessary to produce an entirely correct translation (here, to reproduce the reference translation). Note that the difference in TER between baseline outputs of step 1 and the improved outputs of a system using PRE (step 4) is due to (a) more matches, directly resulting from generating correct pre-translations; (b) indirect "contextual" improvements in the neighborhood of these good translations. To evaluate such indirect effects, for each experiment, we also compute the TER

[16]Note that similar technical solutions also exist for other MT architecture, see e.g. (Chatterjee et al, 2017; Hokamp and Liu, 2017) for constrained decoding in NMT.

[17]We round word averages to nearest integer, which may result in several TER values per each rounded value. We report averaged TER values for such cases.

score of *pseudo post-edits*, where in the initial MT hypotheses we replaced translations of the DT segments (obtained from the $\mathbf{f} \rightarrow \mathbf{e}^1$ alignments, as produced by the decoder) with human suggestions. No re-translation was performed (see Figure 7). TER differences between pre-edition and pseudo post-edition precisely correspond to these indirect improvements.

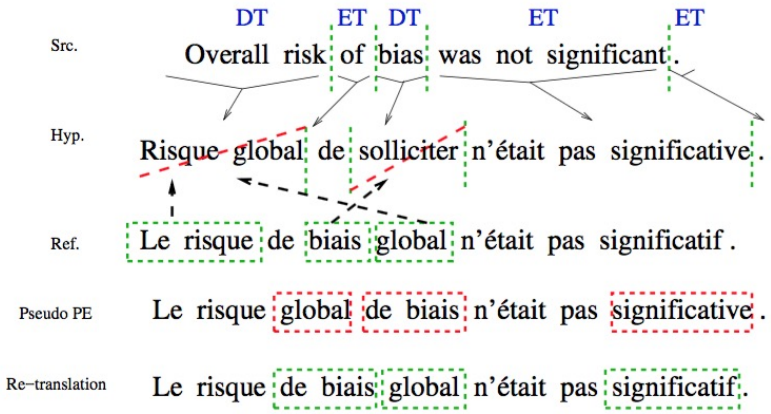


Figure 7: Pseudo PE for the sentence "Overall risk of bias was low." - We simulate pre/post-edition of two DT segments: "overall risk" and "bias".

In the experiments reported below, automatic DT detection is also contrasted with two extreme situations: in an oracle setting, we reproduce the same measurements using *reference difficulty labels* (computed as in Section 2.2) to evaluate the difference of our method with a fully correct DT detection. When DT labels are predicted the posterior probability computed by the classifier can be used to determine the order of pre-translation (e.g., words more likely to be DT will be pre-translated first). When we use reference labels in an oracle setting instead of the predicted ones, this order can be random. Our other contrast study corresponds to the case where segments are pre-translated based on their posterior probability of being ET, thus simulating an extremely poor DT detection method (the worst possible classifier providing zero DT recall and classifying all the ET as DT).

We also study scenarios corresponding to the automatic resolution of the detected DT segments and show that the quality of pre-translations obtained automatically can be sufficient to obtain beneficial PRE results.

4.2 Simultated Pre-Translation in the MEDICAL Domain

Results of the extrinsic evaluation for MEDICAL are plotted in Figures 8a, 8b and 8c, one for each segmentation strategy.

Our results show that in all cases, pre-translating all the DT segments[18] results in a massive improvement in TER. For MT-SEG, for instance, the corresponding gain is about 0.35 absolute, and we see similar improvements in BLEU (31.4 points).[19] The results obtained in the oracle settings are only slightly better (ΔTER $= 0.41$, ΔBLEU $= 37.4$). As expected, the maximum average improvement achieved after constraining the translation of all the ET segments is much lower (ΔTER $= 0.12$, ΔBLEU $= 12.6$, and likewise in the oracle condition with ΔTER $= 0.06$, ΔBLEU $= 8.4$), since we mostly fix translations that are already correct. These series of results confirm that our DT prediction methodology is not only accurate, but also only mildly sensitive to erroneous labels.

The increase in performance w.r.t. to the amount of pre-edition can be summarized by the slope of the regression function: for instance, for WORD-SEG we estimate that each pre-edited word improves the translation by about 0.026 TER point.

Contrasting segmentations, in the oracle setting the MT-SEG strategy yields the best final improvement in translation quality, with a final TER score of 0.18 (BLEU $= 63.9$).

SYNT-SEG offers an almost equally effective alternative to MT-SEG (the marginal improvement per additional word is 0.002 TER).

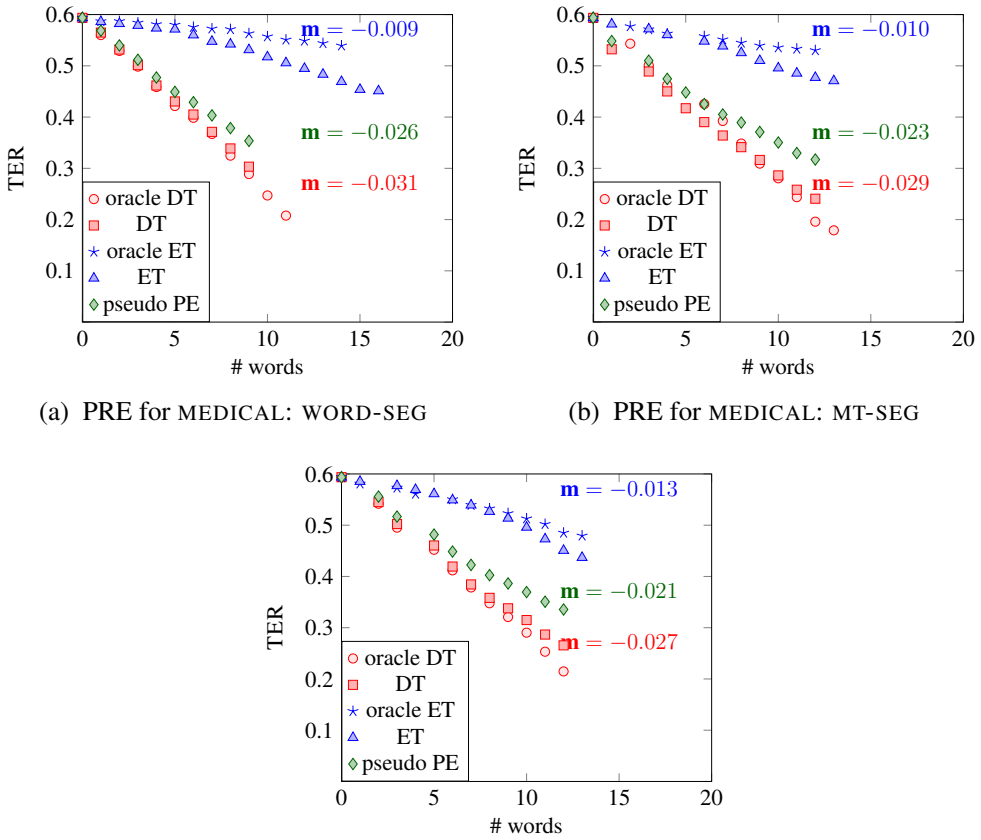
Note that even after pre-translating all the DT segments in the oracle setting translations are still not perfect: the residual improvement needed to go from an average TER $= 0.20$ to TER $= 0.0$ can be explained by residual reordering and omission errors. Indeed, for the oracle MT-SEG, we found that 36% and 33% of the residual edit operations were respectively insertion and shift operations.

Our figures also report the improvements obtained by replacing the DT segments with their correct translation *a posteriori* (pseudo post-edition on the plots). By contrasting these with the pre-editing simulation, we can assess the significance of *indirect improvements*, corresponding to additional matches in the MT for segments that have not been pre-edited: our experiments show that they account for about 16-23% of the total improvement. For instance, the WORD-SEG condition yields an indirect extra-return of about 0.005 TER per additional word, which is more than the half of the return observed when pre-translating ET segments.

To get a closer insight into the nature of these indirect positive changes, we computed the percentage of correctly translated words for each POS in the various settings (pre-edition, pseudo post-edition) - ignoring the words that have been

[18] Recall that DT segments represent approximately 50% of a sentence, cf. Section 3.1.3.

[19] ΔTER is computed as the absolute difference between the initial MT quality score (a point with a pair of coordinates (0,n)) and a quality score resulting from a pre-translation experiment (any other point). A maximum quality gain is computed as the absolute difference of the y-coordinates values for points (0,n) and (max(x),m).



(a) PRE for MEDICAL: WORD-SEG

(b) PRE for MEDICAL: MT-SEG

(c) PRE for MEDICAL: SYNT-SEG

Figure 8: PE effort reduction. Each point on Figures (a)-(c) corresponds to providing the correct translation for t words in each sentence; for each condition (automatic or oracle DT labels, automatic or oracle ET labels, pseudo-post edition), we observe an average improvement in TER. We also report on the graph the slope for the non-oracle condition, which corresponds to the marginal improvement in TER for each extra pre-edited word. # denotes count; m – the slope.

input by the user. These correctly translated words were detected using the word alignments $\mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$ produced by TER.

We found that the pre-edition scenario mostly indirectly influences the translation of adverbs (the percentage of correct translation increases by +5.7% absolute), of verbs (+3.9%), adjectives (+1.5%) and nouns (+1.1%). The intrinsic properties of PBMT (i.e. translation by composition) also creates situations where pre-translation will have negative influence, especially on the translation of auxiliary words (e.g., pronouns, prepositions etc.), which is, however, largely compensated by the positive influence on main POS (cf. Table 12).

POS	#	Δ
ADJ	677	**1.5**
ADV	247	**5.7**
CONJ	1308	0.5
DET	591	0.3
N	2710	**1.1**
PRP	200	-1
PREP	1702	-0.2
PUNCT	2370	-1
V	676	**3.9**

Table 12: Measuring the contextual influence on a per POS basis for the MEDICAL system SYNT-SEG: # denotes, for each POS, the number of words which were not pre-translated; Δ denotes the changes in percentage of "correctly" translated POS after resolving all the DT segments.

Finally, a number of difficulties can be resolved with the help of automatic resolution. Indeed, translations of DT segments can be extracted from parallel/comparable corpora. However the following questions should be asked: will such pre-translations be useful ? And if yes how sensitive is the quality of PRE with respect to the quality of those automatic pre-translations ?

To answer these questions, we take advantage of supplementary resources for the English-French medical task in the form of a narrowly-specialized Cochrane English-French corpus. This corpus contains $\approx$ 31K sentences, where the translations are post-edited versions of Google Translate translations (Ive et al, 2016). This corpus is distinct from the dataset used in classification experiments. Note that this corpus is too small to train an MT system and can only be used as a complementary corpus. We used this data to obtain targeted help as follows: we first automatically detect DT SYNT-SEG segments in the MEDICAL test set, and search for possible pre-translations in a phrase table (PT) extracted from this additional corpus. We found matches for around 60% of the initially detected DT

segments. We pre-translated those segments using the `exclusive xml-mode` of the `Moses` decoder. In each experiment we used the first, second or third most probable translation option from the PT (according to $\phi(\bar{e}|\bar{f})$). In an oracle setting, for the same 60% of the `DT` segments, we used pre-translations extracted from the references.

The results of these experiments are in Table 13, where we see that automatic resolution already improves TER by 0.07 points, as compared to the ΔTER =0.16 improvement obtained with oracle pre-translations.

We can thus conclude that automatic resolution can be beneficial and should be attempted as a preliminary stage to human resolution. It is however clear that the gains in translation performance are very sensitive to the quality of the automatic pre-translations: every time we choose pre-translations of the next best quality, the overall gain reduces by around 50% (e.g., ΔTER =0.07 with the first PT option vs. ΔTER =0.03 with the second PT option).

setup	TER
MT	0.59
all `DT`	
ref.	0.21
60 % of `DT`	
ref.	0.43
1st PT option	0.52
2nd PT option	0.56
3rd PT option	0.58

Table 13: Sensitivity of translation quality to the quality of PRE (PT options are chosen according to the values of $\phi(\bar{e}|\bar{f})$, sorted in the descending order; ref. denotes experiments with reference pre-translations; MT – initial MT quality).

4.3 Experiments in the UN Domain

An extrinsic evaluation of `DT` detection was also performed with our UN data, following exactly the same protocol as above. The `Stanford Arabic Parser` tool (Green and Manning, 2010) and the `MaltParser` tool (Sharoff and Nivre, 2011) for RU were used to obtain reference translations. Results of those experiments are plotted in Figures 9a, 9b, 9c and 9d. A first observation is that for all languages, the number of `ET` words is much larger than the number of `DT` words.

For all the language pairs, pre-translating all the `DT` segments results in a massive improvement in translation quality, which is consistent with our MEDICAL

experiments and confirms the efficiency of our methodology. For EN-RU, for instance, TER improves by 0.21 absolute, and so does BLEU (16.78 points). The difference with the oracle condition is marked, suggesting that it would be here worthwhile to improve DT detection. Again, the maximum improvement achieved after constraining the translation of a comparable number of ET segments remains much lower (e.g., again for EN-RU ΔTER $= 0.14$, ΔBLEU $= 10.15$).

Indirect improvements remain significant, accounting for about 20% of the total improvement for French and Spanish, in line with our previous observations.

For the other two languages, they are more limited and only correspond to approximately 6% of the total improvement. This may be partly attributed to the morphological complexity of the Arabic and Russian languages.

Finally, the influence per POS for EN-FR is again the strongest for nouns (see Table 12). A positive influence is observed for the translation of determiners for all the language pairs, especially for EN-AR and EN-RU.

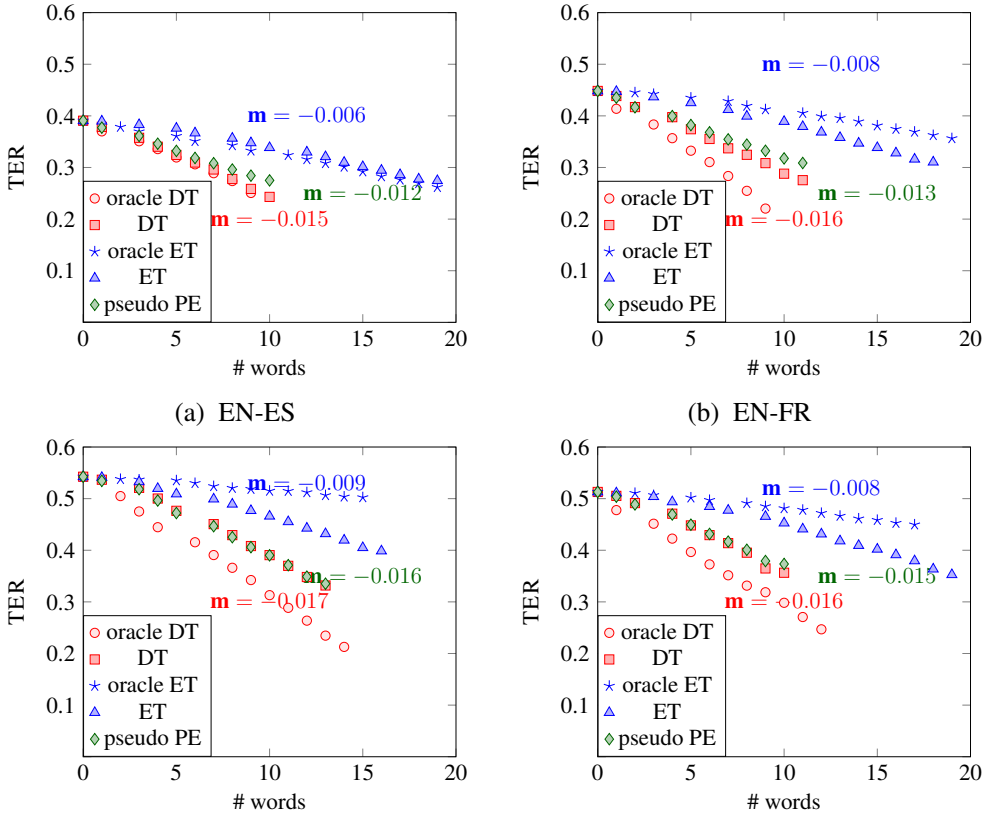
5 Comparing Pre-Edition and Post-Edition

In this section, we investigate how realistic the pre-translation protocol is in terms of the human effort involved. We first present differences between the post- and the pre-editing processes. We then compare (simulated) pre-translation effort with (simulated) post-edition effort, suggesting that pre-edition might also be viable in terms of actual human cost.

5.1 Post-edition

Post-edition (PE) is the process of having a human operator edit and revise the output of an automatic translation system. Assuming that the initial translation is sufficiently good, PE has the potential to greatly speed-up the production of high-quality translations in comparison to a human translation; it is also often acknowledged to yield translations that can be more consistent than human translation. PE is often used in combination with adaptive MT, where the system instantly learns from the editor's input, thereby continuously improving the MT through online updates.

PRE mostly differs from PE in that it aims to fix the MT system before any translation is actually generated: the effort required resembles what human translators actually do when preparing for new translations. As PRE is meant to improve MT output, it can be applied separately or in conjunction with PE, with the effect to reduce the final PE effort. Targeted versions of both processes are meant to solicit human assistance only where needed and to reduce human cognitive effort of analyzing poor MT.



(a) EN-ES (b) EN-FR

(c) EN-RU (d) EN-AR

Figure 9: PE effort reduction (Each point on Figures (a)-(d) corresponds to providing the correct translation for t words in each sentence; for each condition (automatic or oracle DT labels, automatic or oracle ET labels, pseudo-post edition), we observe an average improvement in TER. We also report on the graph the slope for the non-oracle condition, which corresponds the marginal improvement in TER for each extra pre-edited word. # denotes count; m – the slope.

5.2 Assessing Human Effort

Measuring TER scores when a certain quantity of DT segments is pre-translated (as we did above) tells us only part of the story, since this might well be an unrealistic effort for the human pre-editors. In this section, we compute the improvements in TER as a function of the number of (simulated) keystrokes (characters) a human would have to type to provide the reference translation. We reproduce the plots of Section 4.2 with a different x-axis and report per sentence averages of keystrokes.

We compare this effort to the one involved in (a) conventional PE, as well as (b) PE+online adaptation, where the hypotheses are generated by a system performing online updates to its models after each post-edited sentence. To implement this comparison, we also simulated the latter procedure, using again the $\mathbf{e}^1 \rightarrow \hat{\mathbf{e}}$ alignments produced by TER. In each experiment, we "correct" a certain number of words c in each test sentence. c is incremented from 1 to max_{TER}, the maximum quantity of words in a sentence that should be corrected to obtain a reference as prescribed by TER. "Correction" involves the following operations: we replace substituted words with their reference translations, remove deleted words and insert the corresponding reference tokens. We again measure the number of typed-in characters as a proxy to the human effort. We use the Levenshtein edit distance (Levenshtein, 1966) for substituted words, otherwise the length of deleted or inserted words is used. More costly operations (in terms of input characters) are applied first.

To compare PRE to PE+online adaptation, we took advantage of the Moses implementation of adaptive PBMT (Germann, 2015). For the online adaptation of feature weights, we re-implemented the solution proposed by Mathur et al (2013) in the current version of Moses.[20] In a nutshell, this method stores the training data in efficient data structures so that models can be re-estimated for each new input taking previous feedback into account. Those updates are accompanied by online updates of model weights that are meant to increase the score of hypotheses that are close to post-edited translations.

5.3 Results

A MEDICAL PBMT system implementing online updates was created as described in Section 3.1.2. To simulate user feedback we used the same development and test sets, as well as MGIZA (Gao and Vogel, 2008) $\mathbf{f} \rightarrow \hat{\mathbf{e}}$ alignments. Parameters for the online adaptation of feature weights were tuned on the development set using the *Simplex Algorithm* (Nelder and Mead, 1965). The quality improvement

[20] As of October 2016

quality for this system is of $\Delta\text{TER} = 0.16$, $\Delta\text{BLEU} = 15.63$ as compared to the static system.

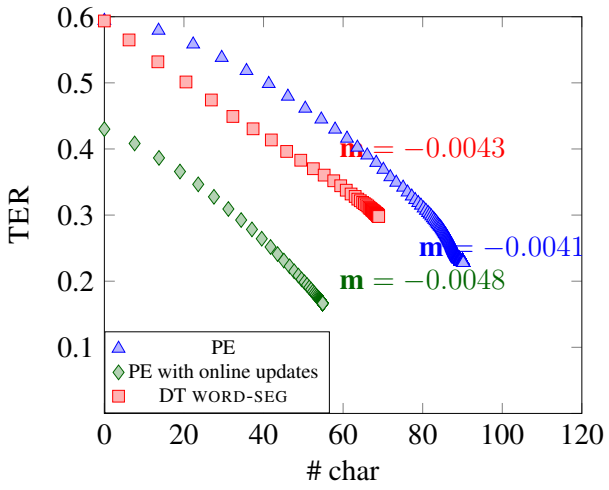


Figure 10: Human effort reduction after PRE WORD-SEG, PE of the initial or online-updated MT output (**m** denotes the slope) for MEDICAL. # denotes count; ; m – the slope.

Figure 10 plots the human effort for PRE WORD-SEG and both kinds of PE. Each point on those plots denotes an experiment, where we simulated correction or input (in the case of pre-translation) of a certain number of words in each test sentence. For instance, if a user types in around 20 characters per sentence to correct MT output, the quality of the static MT output will improve by $\Delta\text{TER} = 0.04$ (blue triangle), the quality of the adaptive MT output will improve by $\Delta\text{TER} = 0.06$ (green diamond). And if a user types in around 20 characters to pre-translate a certain number of DT words per test sentence, the quality of the output will improve by $\Delta\text{TER} = 0.09$ (red square).

According to those results, the human effort reduction for the static PE is similar to the one of PRE: PRE WORD-SEG yields an additional decrease of $\Delta\text{TER} = 0.0002$ per character as compared to PE. The PE with online updates is only slightly more effective: an additional decrease of $\Delta\text{TER} = 0.0005$ per character as compared to PRE WORD-SEG.

Note that our automatic measurements of the human PE effort are quite optimistic, as in real-life settings we cannot expect post-editors to perfectly optimize their keystrokes. Furthermore, the simulation of PE experiments does not take into account the cognitive effort needed to localize the required correction(s). We also put the PE with online updates in a very favorable condition by updating the models with the alignments statically produced by MGIZA. We tend to consider that our PRE scenario remains competitive in terms of the human effort involved,

and believe that the comparative merit of both approaches (PRE and PE) will only be resolved through experiments involving human pre- and post-editors.

6 Related Work

The task of automatically detecting translation difficulty in the source text was introduced by Mohit and Hwa (2007).

They cast the task of detecting difficult segments as a binary classification task, considering only segments that are syntactically motivated, also with additional length requirements: Mohit and Hwa (2007) consider only parse tree constituents, whose string span is between 25% and 75% of the full sentence length. The notion of DT segments used in this early work significantly differs from ours: a phrase is marked as difficult-to-translate if the removal of its translation from an MT hypothesis would yield a positive impact on the resulting document-level BLEU score (assuming the reference would also be simplified accordingly). Mohit and Hwa (2007) consider rather long idiosyncratic syntactic units (with an average length of 8.8 (detokenized) Arabic words for DT). We do not use their approach as baseline in our work since their definition of DT is crucially different from ours and the DT segments they obtain can hardly be resolvable in practice. Segments of such length can hardly be generalizable over their occurrences, and we do not consider that they would be suitable for human resolution. They would also be also difficult to mine in the case of automatic resolution. Using this definition of DT units, these authors found that SVM-based classifiers performed quite reliably for their Arabic-English data, with an average accuracy of 72%.

We borrow from this work the assumption that translation difficulties are directly related to translation quality, marking phrases as DT or ET depending on whether they are "correctly" translated. For this purpose, we use the automatic hypothesis $\rightarrow$ reference alignment as produced by TER to detect "incorrectly" translated segments. This procedure does not require to artificially alter source or target sentences. Like Mohit and Hwa (2007), we also consider syntactic segments, as well as two additional segmentation strategies. However, we do not impose any length constraints on segments, and thus eventually consider shorter phrases than these authors (around 1.6 words on average for SYNT-SEG), which probably makes our detection task somewhat harder than theirs.

Another closely related work is that of Cheng et al (2016). The authors propose a Pick-Revise approach to Interactive MT (PRIMT). In PRIMT, a user interacts with an MT system in the following way: observing an initial automatic translation, the user picks a difficult source segment and provides a better translation (revision) for this segment, which is then used to produce a hopefully improved MT. This process can be iterated multiple times. The authors discuss

ways to automate these two steps, yielding an interaction protocol, which can be viewed as an online version of our simulation. Automatic picking here is also viewed as a binary task, using the following definition of difficulty: a word is `difficult-to-translate` if constraining its translation significantly improves the quality of a re-translated sentence (measured in BLEU) as compared to improvements obtained for other phrases in the same sentence. Note that this definition favours phrases whose correct translation have larger indirect effects. The authors experiment with Maximum Entropy classification models, SVMs and Feedforward Neural Networks. The last yield the best performance in their experiments.

Our work also draws much of its inspiration from recent developments in automatic Quality Estimation (QE), and the design of our classification protocol (features, classification algorithms) is similar to what is typically done in QE; this is especially the case for phrase-level settings (Logacheva and Specia, 2015). Contrarily to QE, our source-oriented context enables us to perform multi-target experiments: we thus consider our approach as more practical, as well as a nice testbed for studying source-language and language-pairs effects. Our main new contribution relative to phrase-level QE is again the study of syntactically-motivated segmentations and multilingual features.

More generally, the idea of targeted PRE goes back to the Human-Assisted MT systems of the end of the previous century (Kay, 1973; Tomita, 1985; Brown and Nirenburg, 1990; Blanchon, 1992), which were focused on resolving local ambiguities through *ex-ante* human intervention. Modern adaptive MT (Mathur et al, 2014; Denkowski et al, 2014; Wuebker et al, 2015) learns from human corrections to automatic translation to improve subsequent output. However, in a real-life setting, checking the long term validity of such updates is often impossible. In fact, in case of erroneous feedback or of too frequent updates, adaptive MT even risks to be harmful for output quality. This issue is usually addressed by *ex-post* Active Learning (AL) strategies (González-Rubio et al, 2012; Alabau et al, 2014; Du et al, 2015), which select which part of human feedback will be used for updates to make them more efficient. Our resolution protocol can be viewed as an *ex-ante* AL solution to get supplementary human information that will be used for targeted adaptation of MT.

7 Conclusions

In this paper, we have studied how human-machine interaction can be used to improve MT output. We have proposed a scenario that involves the detection of MT translation difficulties and their resolution by pre-translation. Solving difficulties *ex-ante*, rather than *ex-post* in a PE setting gives the translator a very transparent

and predictable way to alter the MT behavior, and could meet individual interaction preferences of some professionals.

Our main conclusions are the following: (a) DT segments can be reliably identified at different segmentation levels, even with relatively simple system-independent sets of features; (b) asking a human to pre-translate these segments can be useful to improve MT quality; the indirect effects of pre-edition are genuine (up to 20% of the total improvement in translation quality); they can be both positive and negative; (c) our PRE scenario is realistic in terms of the human effort involved; (d) translation difficulties depend on the language pair, rather than solely on the source language.

In the future, we plan to reproduce our study for other types of MT, including neural MT and other language pairs. Developing realistic scenarios for detecting and pre-translating DT segments at the document level in a multilingual setting is one of our priority. The main challenge of the document-level resolution is to choose a set of DT segments that should be (a) likely to improve significantly the translation in multiple places (including direct and indirect effects), and (b) whose translation should be as context-independent as possible, so that the same pre-translation can be reused multiple times. In this respect, an interesting perspective is proposed by Bandit Learning (Bubeck and Cesa-Bianchi, 2012), which seeks to maximize long term reward (here automatic translation quality) based on actual human activity, without any knowledge of internals of the system's behavior.

Another line of improvements to our methodology is to go beyond the formal evaluation of translation quality as proposed by naive reference-based metrics such as BLEU or TER. This would allow us to introduce semantic-based labelings of ET and DT, with the view that DT segments should be defined as segments whose translation yields a significant information loss.

Last, we also wish to make translation difficulty resolution automatic. Indeed, once DT and ET segments are detected in the input, it is possible to search for translations of DT segments using additional sources of information, for instance non-parallel corpora, or using multi-source and pivot system. The latter resolution strategy, implemented through system combination, could help to locally improve MT without jeopardizing the quality of already "correctly" translated segments.

References

Alabau V, González-Rubio J, Ortiz-Martínez D, Sanchis Trilles G, Casacuberta F, García-Martínez M, Mesa-Lao B, Petersen DC, Dragsted B, Carl M (2014) Integrating online and active learning in a computer-assisted translation workbench. Association for Machine Translation in the Americas, Proc. the 11th AMTA Workshop on Interactive and Adaptive Machine Translation, pp 1–9

Aronson AR, Lang FM (2010) An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 17(3):229–236

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. CoRR abs/1409.0473

Blanchon H (1992) A solution for the problem of interactive disambiguation. Coling Organizing Committee, Proc. COLING, pp 1234–1238

Bojar O, Chatterjee R, Federmann C, Graham Y, Haddow B, Huck M, Jimeno Yepes A, Koehn P, Logacheva V, Monz C, Negri M, Neveol A, Neves M, Popel M, Post M, Rubino R, Scarton C, Specia L, Turchi M, Verspoor K, Zampieri M (2016) Findings of the 2016 workshop on statistical machine translation. Association for Computational Linguistics, Berlin, Germany, Proc. WMT, pp 131–198

Bojar O, Chatterjee R, Federmann C, Graham Y, Haddow B, Huang S, Huck M, Koehn P, Liu Q, Logacheva V, Monz C, Negri M, Post M, Rubino R, Specia L, Turchi M (2017) Findings of the 2017 conference on machine translation (wmt17). Association for Computational Linguistics, Copenhagen, Denmark, Proc. WMT, pp 169–214

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Brown RD, Nirenburg S (1990) Human-computer interaction for semantic disambiguation. Association for Computational Linguistics, Proc. COLING, pp 42–47

Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. Found and Trends in Mach Learn 5(1):1–122

Chatterjee R, Negri M, Turchi M, Federico M, Specia L, Blain F (2017) Guiding neural machine translation decoding with external knowledge. Association for Computational Linguistics, Copenhagen, Denmark, Proc. WMT, pp 157–168

Cheng S, Huang S, Chen H, Dai XY, Chen J (2016) PRIMT: A pick-revise framework for interactive machine translation. Association for Computational Linguistics, Proc. NAACL-HLT, pp 1240–1249

Cherry C, Foster G (2012) Batch tuning strategies for statistical machine translation. Association for Computational Linguistics, Proc. NAACL-HLT, pp 427–436

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297

Denkowski M, Dyer C, Lavie A (2014) Learning from post-editing: Online model adaptation for statistical machine translation. Association for Computational Linguistics, Gothenburg, Sweden, Proc. EACL, pp 395–404

Du J, Srivastava A, Way A, Maldonado-Guerra A, Lewis D (2015) An empirical study of segment prioritization for incrementally retrained post-editing-based SMT. Association for Machine Translation in the Americas, Proc. MTS, pp 172–186

Dyer C, Chahuneau V, Smith NA (2013) A simple, fast, and effective reparameterization of IBM model 2. Association for Computational Linguistics, Proc. NAACL, pp 644–648

Eisele A, Chen Y (2010) MultiUN: A multilingual corpus from United Nation documents. European Language Resources Association, Proc. LREC, pp 2868–2872

Federico M, Koehn P, Schwenk H, Trombetti M (2013) Matecat: Machine translation enhanced computer assisted translation. International Association for Machine Translation, European Association for Machine Translation, Proc. MTS, p 425

Gao Q, Vogel S (2008) Parallel implementations of word alignment tool. Association for Computational Linguistics, Proc. SETQA-NLP, pp 49–57

Germann U (2015) Sampling phrase tables for the Moses statistical machine translation system. Prague Bull of Math Linguist 104(1):39–50

Gong L, Max A, Yvon F (2014) Incremental development of statistical machine translation systems. Proc. IWSLT, pp 214–222

González-Rubio J, Ortiz-Martínez D, Casacuberta F (2012) Active learning for interactive machine translation. Association for Computational Linguistics, Proc. EACL, pp 245–254

Green S, Manning CD (2010) Better Arabic parsing: Baselines, evaluations, and analysis. Coling 2010 Organizing Committee, Proc. COLING, pp 394–402

Hokamp C, Liu Q (2017) Lexically constrained decoding for sequence generation using grid beam search. Association for Computational Linguistics, Proc. ACL, pp 1535–1546, DOI 10.18653/v1/P17-1141

Ive J, Max A, Yvon F, Ravaud P (2016) Diagnosing high-quality statistical machine translation using traces of post-edition operations. In: Proc. of the LREC

2016 Workshop: Translation evaluation – From fragmented tools and data sets to an integrated ecosystem, European Language Resources Association, pp 55–62

Kay M (1973) The MIND system. In: Rustin R (ed) Natural language processing: Courant Computer Science Symposium 8: December 20-21, 1971, Algorithmics Press, pp 155–188

Kay M (1997) The proper place of men and machines in language translation. Mach Transl 12(1):3–23

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. CoRR abs/1412.6980

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, Herbst E (2007) Association for Computational Linguistics, Prague, Czech Republic, Proc. ACL, pp 177–180

Koehn P, Carl M, Casacuberta F, Marcos E (2013) CASMACAT: Cognitive analysis and statistical methods for advanced computer aided translation. International Association for Machine Translation, European Association for Machine Translation, Proc. MTS, p 411

Lafferty JD, McCallum A, Pereira FCN (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Morgan Kaufmann Publishers Inc., Proc. ICML, pp 282–289

Lavergne T, Cappé O, Yvon F (2010) Practical very large scale CRFs. Association for Computational Linguistics, Proc. ACL, pp 504–513

Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions and reversals. Sov Phys Doklady 10(8):707–710

Logacheva V, Specia L (2015) Phrase-level quality estimation for machine translation. Proc. IWSLT, pp 143–150

Logacheva V, Hokamp C, Specia L (2016) MARMOT: A toolkit for translation quality estimation at the word level. European Language Resources Association, Portorož, Slovenia, Proc. LREC, pp 3671–3674

Mathur P, Mauro C, Federico M (2013) Online learning approaches in computer assisted translation. Association for Computational Linguistics, Sofia, Bulgaria, Proc. WMT, pp 301–308

Mathur P, Cettolo M, Federico M, de Souza JGC (2014) Online multi-user adaptive statistical machine translation. Association for Machine Translation in the Americas, Proc. AMTA, pp 152–166

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. CoRR abs/1301.3781

Mohit B, Hwa R (2007) Localization of difficult-to-translate phrases. Association for Computational Linguistics, Proc. StatMT Workshop, pp 248–255

Monroe W, Green S, Manning CD (2014) Word segmentation of informal Arabic with domain adaptation. Association for Computational Linguistics, Baltimore, Maryland, Proc. ACL, pp 206–211

Nelder JA, Mead R (1965) A simplex method for function minimization. Comput J 7(4):308–313

Papineni K, Roukos S, Ward T, Zhu WJ (2002) Bleu: a method for automatic evaluation of machine translation. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, Proc. ACL, pp 311–318

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830

Schmid H (1995) Improvements in part-of-speech tagging with an application to German. Association for Computational Linguistics, Proc. ACL-SIGDAT, pp 47–50

Sharoff S, Nivre J (2011) The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. Izdatel'stvo RGGU, Proc. Dialogue, pp 657–670

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. Association for Machine Translation in the Americas, Proc. AMTA, pp 223–231

Specia L, Shah K, de Souza JG, Cohn T (2013) QuEst - a translation quality estimation framework. Association for Computational Linguistics, Proc. ACL, pp 79–84

Stolcke A (2002) SRILM - an extensible language modeling toolkit. Denver, Colorado, Proc. ICSLP, pp 901–904

Tiedemann J (2012) Parallel data, tools and interfaces in OPUS. European Language Resources Association, Proc. LREC, pp 2214–2218

Tomita M (1985) Feasibility study of personal/interactive machine translation systems. In: Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Proc. Conf on Theor and Methodol Issues in Mach Transl of Nat Lang, pp 289–297

Toutanova K, Klein D, Manning CD, Singer Y (2003) Feature-rich part-of-speech tagging with a cyclic dependency network. Association for Computational Linguistics, Proc. NAACL-HLT, pp 173–180

Wuebker J, Green S, DeNero J (2015) Hierarchical incremental adaptation for statistical machine translation. Association for Computational Linguistics, Proc. EMNLP, pp 1059–1065