Diagnosing High-Quality Statistical Machine Translation Using Traces of Post-Edition Operations

Julia Ive, Aurélien Max, François Yvon, Philippe Ravaud

Diagnosing High-Quality Statistical Machine Translation Using Traces of Post-Edition Operations

Julia Ive[1,2], **Aurélien Max**[1], **François Yvon**[1], **Philippe Ravaud**[2]
(1) LIMSI, CNRS, Univ Paris-Sud, Université Paris-Saclay, 91 403 Orsay, France,
(2) Cochrane France, INSERM U1153, 75181 Paris, France
{julia.ive,amax,yvon}@limsi.fr, philippe.ravaud@htd.aphp.fr

Abstract

This paper proposes a fine-grained flexible analysis methodology to reveal the residual difficulties of a high-quality Statistical Machine Translation (SMT) system. This proposal is motivated by the fact that the traditional automated metrics are not enough informative to indicate the nature and reasons of those residual difficulties. Their resolution is however a key point towards improving the high-quality output. The novelty of our approach consists in diagnosing Machine Translation (MT) performance by making a connection between errors, the characteristics of source sentences and some internal parameters of the system, using traces of Post-Edition (PE) operations as well as Quality Estimation (QE) techniques. Our methodology is illustrated on a SMT system adapted to the medical domain, based on a high quality English-French parallel corpus of Cochrane systematic review abstracts. Our experimental results show that the main difficulties that the system faces are in the domains of term precision and source language syntactic and stylistic peculiarities. We furthermore provide general information regarding the corpus structure and its specificities, including internal stylistic varieties characteristic of this sub-genre.

Keywords: MT evaluation, high-quality SMT, post-edition

1. Introduction

Nowadays, narrowly-specialized MT systems are able to produce very high quality translations, as measured by automated metrics. In most cases, though, the final output still requires manual improvements to reach a publishable quality. However, standard automated metrics such as (H)BLEU (Papineni et al., 2002), (H)METEOR (Denkowski and Lavie, 2014) or (H)TER (Snover et al., 2006)[1] provide little clues regarding the remaining errors, and are of little help to suggest fixes or improvements.

The same can be said of automated error analysis techniques, which are often based on similar principles (Popovic and Ney, 2011; Bojar, 2011): In particular, they often consider the system as a black-box and tend to ignore the characteristics of the source text.

In this study, we propose an alternative fine-grained methodology that helps indicate translation difficulties in connection to the peculiarities of the source document, and also provide some hints as to the reasons of those difficulties in relation to the original corpus and the internal scoring procedures. Such a methodology proves especially useful in the context of high-quality MT, which requires more targeted and sophisticated solutions for further improvement.

Our approach is illustrated using a medical SMT system built from a corpus of Cochrane medical systematic review abstracts. An English-French parallel corpus of such abstracts, including human and post-edited automatic translations, will be described.

The rest of this paper is organized as follows: in Section 2., we will present the main characteristics of the Cochrane corpus used. In Section 3., we will describe the challenges of the medical translation task in the context of the Cochrane Collaboration, before introducing our MT system analysis methodology in Section 4. We will finally present the results of the analysis applied to the Cochrane SMT system in Section 5., and conclude and discuss further prospects for MT evaluation and diagnosis in Section 6.

[1]Hereinafter, 'H' will be added to refer to the automated metrics applied to the references created by post-editing the evaluated MT output.

2. The Cochrane Bilingual Parallel Corpus

Cochrane France is part of the international non-profit Cochrane Collaboration[2] whose main mission is to globally spread high-quality evidence-based research in medicine. To this end, the Cochrane Collaboration publishes high-standard research reviews in English and selective translation of their abstracts into (as of now) 12 languages including French, Spanish, Japanese, and traditional Chinese. The review abstracts are publicly available online [3]. Full research reviews are openly accessible only for the low-income and middle-income countries.

Each Cochrane review abstract is made up of the following parts: (a) a plain language summary (PLS, 40% of the abstract, written in popular scientific style), focused on patient comprehension; (b) a scientific abstract (ABS, 60% of the open access abstract, written in scientific technical style), targeting medical experts.

The English-French Cochrane parallel corpus used in this study consists of the following:[4]

- **Cochrane Reference Corpus**: a high-quality corpus consisting of review abstracts translated by agencies and reviewed by domain professionals over a three-year period (2011-2013).

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

[3]http://www.cochranelibrary.com

[4]The corpus consisting of source text, machine translation output and PE output is available at http://www.translatecochrane.fr/corpus.

- **Cochrane Post-editing (PE) Corpus**: a lower quality corpus consisting of machine-translated review abstracts post-edited mainly by volunteer domain professionals over a 6-month period (Oct. 2013-May 2014). The MT was performed by different versions of the Cochrane SMT.
- **Cochrane Google Post-editing (PE) Corpus**: a lower quality corpus consisting of machine-translated review abstracts by the Google online system[5] post-edited by both professional translators and volunteer domain professionals over a 1-year period (Aug. 2014-Sep. 2015).

Table 1 provides statistics about each part of the corpus.

Corpus	# Lines	# Tokens, en (src)	# Tokens, fr (trg)
Cochrane Reference	≈ 130 K	≈ 2.9 M	≈ 3.6 M
Cochrane PE	≈ 21 K	≈ 500 K	≈ 600 K
Cochrane Google PE	≈ 31 K	≈ 740 K	≈ 890 K

Table 1: Corpora sizes

3. Automatic Translation of Cochrane Systematic Review Abstracts: Challenges and Solutions

The translation of English medical texts, in particular that of Cochrane systematic review abstracts, presents a series of challenges regarding:

1. the translation of the terminology and the professional jargon (e.g. abbreviations);
2. the translation of complex syntactic structures and compounds;
3. the adaptation to variations within the scientific style (this is particularly important in the Cochrane context, where different language styles are in use in the PLS and ABS sections).

We manually inspected the paraphrase tables extracted from PLS and ABS parts of the Cochrane Reference and PE Corpora to reveal the following stylistic differences between the registers (Denkowski and Lavie, 2014; Bannard and Callison-Burch, 2005):

1. terminology register (e.g., Source: "*cycling*", ABS: "*cyclisme*" 'cycling'[6], PLS: "*vélo*" 'bicycle'; Source: "*surgical fixation*", ABS: "*ostéosynthèse chirurgicale*" 'surgical osteosynthesis', PLS: "*fixation chirurgicale*" 'surgical fixation');
2. professional jargon (e.g., Source: "*once-daily*", ABS: "*une administration quotidienne*" 'a daily administration', PLS: "*une fois par jour*" 'once a day'; Source: "*viral*", ABS: "*viral*" 'viral', PLS: "*par des virus*" 'by viruses');
3. selective translation of names (e.g., Source: "*Cochrane Library*", ABS: "*Cochrane Library*", PLS: "*Bibliothèque Cochrane*" 'Cochrane Library'; Source: "*Cochrane Review*", ABS: "*Cochrane Review*", PLS: "*revue Cochrane*" 'Cochrane review');
4. general language (e.g., Source: "*to*", ABS: "*afin de*" 'so that', PLS: "*pour*" 'to'; Source: "*flexible*", ABS: "*flexible*" 'flexible', PLS: "*souple*" 'soft').

The use of domain adaptation techniques, as well as more *ad-hoc* solutions, can help to obtain a better performance in medical MT (Costa-jussà et al., 2012; Wang et al., 2014; Boguraev et al., 2015). In any case, high-quality translation in specialized domains requires training data that closely match the test data.

The Cochrane SMT system for translating the systematic review abstracts is an example of such a narrowly-specialized system. In its current form, our system uses the Moses toolkit (Koehn et al., 2007). The Cochrane Reference corpus is used to train the main model (phrase table and reordering model `msd-bidirectional-fe`). Cochrane PE and additional corpora (WMT'14 medical task parallel data [7]) models (same components as for the main model) were used only for n-grams (up to $n = 4$) when no translation is found by the first model. The monolingual parts of the corpora mentioned above, as well as general domain data (WMT'13 news data [8]) were used to train the corresponding language models.

The system was tuned using post-edited data, which is in line with the final quality requirements of producing comprehensible texts with minimum corrections to the MT output.

An automatic evaluation of this system was performed using a test set comprising 713 sentences for the PLS part and 949 sentences for the ABS part. Those sentences were extracted from the corresponding machine-translated and post-edited review abstracts.

Results, presented in Table 2, reveal a high level of translation performance according to the automatic metrics used, with a slightly better performance for the ABS section.

We also report a comparison with translations produced by the online Google system [9], as well as with the translations of the target test set produced by a lower performance system trained only on the WMT'14 medical task parallel data (WMT'14 SMT). This system uses the language models built with the monolingual parts of the WMT'14 medical data and WMT'13 news data. It was tuned using the same post-edited Cochrane data as the Cochrane SMT.

The linear lattice BLEU oracle (LB-4g) was used to estimate the system potential (Sokolov et al., 2012). The atypically low oracle improvements in terms of the automatic metrics scores (+6 H-BLEU, +4 H-METEOR) suggest that the system produces translations that are close to the best translations it can produce given its training data.

Analysis of the HTERp traces confirmed the system performance differences for the PLS and ABS parts (see Table 3). For our experiments, we used the HTERpA configuration (Snover et al., 2009), optimized for human adequacy judgments, with the following components for processing French: the Snowball stemmer (Porter, 2001), and a paraphrase table extracted from the concatenation of

[5] `https://translate.google.com`

[6] Hereinafter, literal translations are provided by the first author.

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

[8] `http://statmt.org/wmt13/translation-task.html`

[9] the version publicly available in Sep. 2015

Metric	Cochrane SMT			WMT'14 SMT			Google SMT		
	ALL	PLS	ABS	ALL	PLS	ABS	ALL	PLS	ABS
H-BLEU↑	57	55	58	29	30	28	49	50	48
Oracle H-BLEU↑	63	62	64	40	41	39	NA	NA	NA
H-METEOR↑	73	72	74	56	55	56	67	67	66
Oracle H-METEOR↑	77	75	78	59	59	58	NA	NA	NA
H-TER↓	30	32	28	58	54	62	36	37	35
Oracle H-TER↓	30	32	28	55	50	59	NA	NA	NA

Table 2: Automatic evaluation results

the Cochrane Reference and PE corpora (Denkowski and Lavie, 2014; Bannard and Callison-Burch, 2005).

	PLS	**ABS**
HTERp Score ↓	25	25
# Hyp. Tokens	18534	31872
# Ref. Tokens	18502	32438
Operation	% Hyp. Tokens Edited	
Shift	4	5
Match	74	78
Stem match	3	3
Paraphrase	7	6
Substitution	8	7
Deletion	8	6
Edition	% Ref. Tokens Edited	
Insertion	7	7

Table 3: Number of hypothesis/reference tokens (words) aligned by an HTERp operation or a match

The post-edition operations performed to the output translation tend to be non-repetitive: only about 11% of edited tokens/pairs of tokens per operation are unique, but the most frequent post-edition operations (see Table 4) do not exceed 11% of all the changes per operation.

	PLS		ABS	
Operation	Tokens	%	Tokens	%
Stem Match	de → des	11	de → des	11
Paraphrase	les pansements → pansements à base	1	de la même fratrie → frères et sœurs	1
Substitution	les → des	2	, → ;	8
Deletion	de	6	les	5
Insertion	,	4	de	4

Table 4: Most frequent token changes per operation

As shown in Table 5, the most common Part-of-Speech (POS) substitution patterns reveal frequent modifications to nouns (NC) and to POS's that cooccur with them (DET, P, ADJ), potentially forming terms and terminological constructions, as well as grammatical changes to verbs (V (gram)) (Toutanova et al., 2003; Schmid, 1995).

PLS		ABS	
Pattern	%	Pattern	%
P → P	10	P → P	9
NC → NC	7	NC → NC	8
DET → DET	7	PUNC → PUNC	8
DET → P	5	DET → P	6
P → DET	4	DET → DET	4
V → V (gram)	3	ADJ → ADJ	4
ADJ → ADJ	3	P → DET	4
ADJ → NC	3	ADJ → NC	3
VPP → VPP	2	V → V (gram)	2
V → V	2	NC → P	2

Table 5: Most common POS substitution patterns

Such unusually high translation quality scores do not allow us, however, to dispense with a final post-edition step before publication. Also, improving the system to reduce the post-editor burden remains an important goal. To this end, a fine-grained performance analysis is needed to detect the remaining translation difficulties and to guide future improvements to the system. Further, while analyzing the high-performance MT, we will talk about "residual" errors and difficulties.

4. Diagnosing MT Performance

Since most human evaluation procedures are very costly, MT quality is traditionally measured using reference-based automatic metrics that compute a similarity score between the machine output and one or several human translations (or post-editions) (e.g., (H)BLEU (Papineni et al., 2002), (H)TER (Snover et al., 2006), (H)TER-plus (Snover et al., 2009), (H)METEOR (Denkowski and Lavie, 2014) etc.), which are based on an automatic alignment between words from the machine translation and words from the reference translation. Such alignments are often taken as the basis for an automated error analysis (e.g., (Popovic and Ney, 2011; Berka et al., 2012)). These methods, however, regard the system as a black-box and analyze only its output without any connection to the source text or to the system's specificities.

The trend to take more insight into system internals is observed for Quality Estimation (QE) of MT (Specia et al., 2010; Specia and Giménez, 2010), where most approaches based on Machine Learning techniques take into account both the output, its alignment to the source text, and additional systems scores (Wisniewski et al., 2014; Specia et al., 2015). Irvine *et al.* (2013) go one step further, trying to investigate the interconnection between the source, target and system-dependent characteristics in an attempt to detect domain adaptation errors. An approach of analyzing MT performance in a contrastive manner per linguistic phenomena (e.g., POS) is proposed by Max *et al.* (2010).

Inspired by these latter studies, we propose a new methodology for diagnosing MT performance that should help us to answer the following questions: Which kind of **translation difficulties does a system face**? Are those difficulties related to a greater extent to the **initial corpus quality or to the system scoring procedure**? To the best of our knowledge, this is the first attempt to analyze high-quality SMT by associating residual errors, detected during PE, with source characteristics and system parameters.

Taking into account the observations presented in Table 5, we decided to focus on the translation quality of certain syntactic constituents and POS, in particular noun phrases, as potential complex terminological structures, verbs and nouns (Klein and Manning, 2003).

We extracted the following groups of unique source n-grams (units): the ones corresponding to longest noun phrases (NP), then from the rest of the sentence we extracted units corresponding to the neighboring/single verbs (V) and nouns (N). The residual sentence spans of varying length, not covered so far, were put in a separate group (Rest). A sketch of our protocol is provided in Figure 1.

Further, we distinguished the following subordinate groups: the units that are present in the system's phrase table (PT) and also present in the 1-best hypothesis segmentation in

$>= 80\%$ of their occurrences (k_{1-best}); the ones that are present in PT but are absent from the 1-best segmentation in $>= 80\%$ of their occurrences (k_{pres}); and the units that are absent from the PT (k_{abs}).

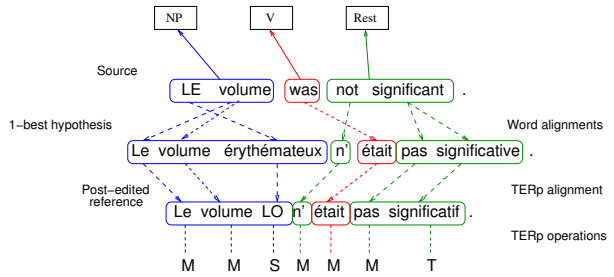


Figure 1: Illustration of our analysis strategy

Using the output word alignments, as well the hypothesis → post-edited reference alignments produced by HTERp, we compute for each unit (k_i) the averaged translation quality statistics for all its occurrences (t_j), by comparing the aligned hypothesis segment (h_s) to its aligned reference segment (r_m). Hypothesis → oracle hypothesis and oracle hypothesis → post-edited reference HTERp alignments were used to calculate the averaged oracle translation quality statistics. More precisely, we estimate the following parameters:

1. unit frequency (fr);
2. unit length in words ($\#\, w_k$);
3. average per occurrence t_j percentage of the unit hypothesis segment words w_h aligned to reference segment words w_r with each TERp operation or a match (e.g., match (M), substitution (S), stem match (T), paraphrase (P) etc.), and correspondingly for the oracle hypothesis segment (M_O, S_O etc.):

$$M = \frac{\#\, M_{w_h}}{\#\, w_h} \tag{1}$$

To trace the connection between the system performance and source peculiarities, we calculate the unit *term rate*:

$$\text{term rate} = \frac{\#\, w_k^t}{\#\, w_k} \tag{2}$$

where w_k^t is the words of a unit marked as terms or parts of complex terms.

The term mapping was performed with the Metamap tool for medical texts (UMLS, 2009). Metamap searches were parametrized to avoid mapping to general concepts. A corpus statistics filter was used to further exclude highly frequent words.

Our methodology extends the approach described in (Irvine et al., 2013) and associates target errors with occurrences in the original training corpus. We do so by computing the prior translation entropy (H_{prior}) of the distribution of the phrase translation probabilities $p(\bar{t}|\bar{s})$ of all the possible target bi-phrases $\bar{t}$ with $\bar{s}$ equal to the unit, taken from the PT with lemmatized $\bar{t}$:

$$H_{prior} = -\sum_{k=1}^{n} p_k(\bar{t}|\bar{s}) \log p_k(\bar{t}|\bar{s}) \tag{3}$$

We attempt to correlate the errors with the scoring procedure by measuring the presence of the reference translation in the oracle hypothesis. We extend the analysis of this correlation by computing the average posterior entropy (H_{post}) of the normalized distribution of the 1-gram path posterior probabilities $P(u|\varepsilon)$, composing a unit.

$$H_{post} = -\sum_{k=1}^{n} P_k(u|\varepsilon) \log P_k(u|\varepsilon) \tag{4}$$

We calculate 1-gram posterior probabilities $P(u|\varepsilon)$ from the estimation of path posterior probabilities as defined in (de Gispert et al., 2013):

$$P(u|\varepsilon) = \frac{\sum_{E \in \varepsilon_u} \exp(\alpha H(E,F))}{\sum_{E' \in \varepsilon} \exp(\alpha H(E',F))} \tag{5}$$

where ε is the space of translation hypotheses (a $10K$-best list was chosen), and $H(E,F)$ is the score assigned by the model to the sentence pair (E,F).

The probabilities of the target bi-phrases $\bar{t}$ and path posterior probabilities of 1-grams sharing the same lemma were added.

5. Evaluation Results

The proposed methodology was applied to the test set presented in Section 3. to analyze the functioning of the Cochrane SMT, as well as the functioning of the less competitive WMT'14 SMT. Examples of the test set sentences demonstrating the translation challenge are provided in Table 6.

During our analysis of residual translation difficulties of the Cochrane SMT, we attempted to find answers to the following questions:

1. What are the "worst" translated unit groups for the high-performance system?

We took the average percentage of matches per group M as an indicator of translation quality (see Figure 2a). We explored the group characteristics by analyzing their general statistics (see Table 7) and the *term rate* (see Figure 2c).

From Figure 2a we can see that the system faces difficulties translating the units of the V group (lowest average $M \approx 53\%$), although the majority of those units are known to the model (97%, *1-best+Pres*, see Table 7).

For the NP group, Figure 2a shows the "worst" translation quality of the units that are absent from the PT (M=74%, *Abs*), which need to be translated by composition.

Figure 2c detects the high term concentration for the N group units (average *term rate*=30%). Thus, the "worst" translated units of the N group (M=24%, *Abs*) are mainly terms unknown to the model. The high rate of N units that are present in the 1-best segmentation (25%, *1-best*, see Table 7) suggests frequent term translation inconsistency due to lack of context information.

PLS	
Source	A lack of growth and poor nutrition are common in children with chronic diseases like cystic fibrosis and paediatric cancer.
Cochrane SMT	Un manque de la croissance et une mauvaise nutrition sont fréquents chez les enfants atteints de maladies chroniques comme la mucoviscidose et le cancer pédiatrique. 'A lack of growth and bad nutrition are common in children suffering from chronic diseases like cystic fibrosis and paediatric cancer.'
Oracle	Un manque de la croissance et une mauvaise nutrition sont fréquents chez les enfants atteints de maladies chroniques comme la mucoviscidose et les cancers. chez les enfants 'A lack of growth and bad nutrition are common in children suffering from chronic diseases like cystic fibrosis and cancers. in children'
PE	Une croissance réduite et une mauvaise nutrition sont fréquentes chez les enfants atteints de maladies chroniques comme la mucoviscidose et les cancers pédiatriques. 'A reduced growth and bad nutrition are common in children suffering from chronic diseases like cystic fibrosis and the paediatric cancers.'
ABS	
Source	Poor growth and nutritional status are common in children with chronic diseases.
Cochrane SMT	Une mauvaise croissance et le statut nutritionnel sont fréquents chez les enfants atteints de maladies chroniques. 'A bad growth and the nutritional status are common in children suffering from chronic diseases.'
Oracle	Une mauvaise croissance et le statut nutritionnel sont fréquents chez l'enfant de 'A bad growth and the nutritional status are common in the child of'
PE	Une croissance réduite et un mauvais statut nutritionnel sont fréquents chez l'enfant atteint de maladie chronique. 'A reduced growth and a bad nutritional status are common in the child suffering from a chronic disease.'

Table 6: Examples of PLS and ABS test set sentences

The same difficulties are observed for the less competitive WMT'14 SMT: the V group units are the "worst" translated (lowest average $M \approx 36\%$); translation of the NP group units absent from PT is of a low quality (M=61%, *Abs*); translation of the term N units present in the 1-best segmentation is often inconsistent (M=44%, *1-best*, *term rate*=34%, see Figure 3a, Figure 3c).

NP total : 3528						
	Cochrane SMT			WMT'14 SMT		
	1-best	Pres	Abs	1-best	Pres	Abs
%	10	27	63	9	10	81
$\# w_k$	2	3	10	2	2	9
fr	1	1	1	1	1	1
N total : 336						
	Cochrane SMT			WMT'14 SMT		
	1-best	Pres	Abs	1-best	Pres	Abs
%	25	71	4	41	47	12
$\# w_k$	1	1	1	1	1	1
fr	1	2	1	1	3	1
V total : 982						
	Cochrane SMT			WMT'14 SMT		
	1-best	Pres	Abs	1-best	Pres	Abs
%	18	79	3	32	62	6
$\# w_k$	1	1	2	1	1	2
fr	1	3	1	1	4	1
Rest total : 931						
	Cochrane SMT			WMT'14 SMT		
	1-best	Pres	Abs	1-best	Pres	Abs
%	13	75	12	21	57	22
$\# w_k$	2	2	2	1	1	2
fr	1	6	1	1	8	1

Table 7: General statistics per unit group

2. To which extent the high-performance system scoring procedure is responsible for the residual errors?

To answer this question we analyzed the per-group differences between system hypotheses and oracle hypotheses match percentage values ΔM (see Figures 2a, 2b).

Additionally, to evaluate the scoring procedure we studied the correlation between the low/high match percentage zones (see Figure 2a) and the prior/posterior entropy values (see Figures 4a, 4b). E.g., we can see that the present in the PT (*1-best+Pres*) N group units with the high match percentage (average $M \approx 73\%$) and the V group units with the low match percentage (average $M \approx 57\%$) both correspond to the same average prior entropy value ($H_{prior} \approx 2$), as well as to the absence of significant difference between the average posterior entropy values ($H_{post} \approx 0.4$ and $H_{post} = 0.3$ correspondingly).

With the average ΔM of about 5%, we can conclude that in the majority of cases the system is unable to produce "correct" translations. The absence of correlation between the match percentage and prior/posterior entropy values confirms that the scoring procedure is not responsible for most of the errors.

In comparison, the scoring procedure of the WMT'14 SMT can be improved more efficiently. The oracle changes to the WMT'14 SMT output (ΔM of about 4%) are more significant since they are performed for more units. From Table 7 and Figures 3a, 3b, we see that the translation of 41% of the *1-best* N group units is improved with ΔM=1% (compare to 25% of N *1-best* units with ΔM=1% for the Cochrane SMT, see Figures 2a, 2b).

For the WMT'14 SMT we should also notice the presence of a more distinct correlation between the translation quality indicator and entropy values: e.g., the high posterior entropy value ($H_{post} = 0.5$) for the *1-best* N units corresponds to the low match percentage (M=44%, see Figures 4c, 3a).

3. What is the nature of the per-group residual errors?

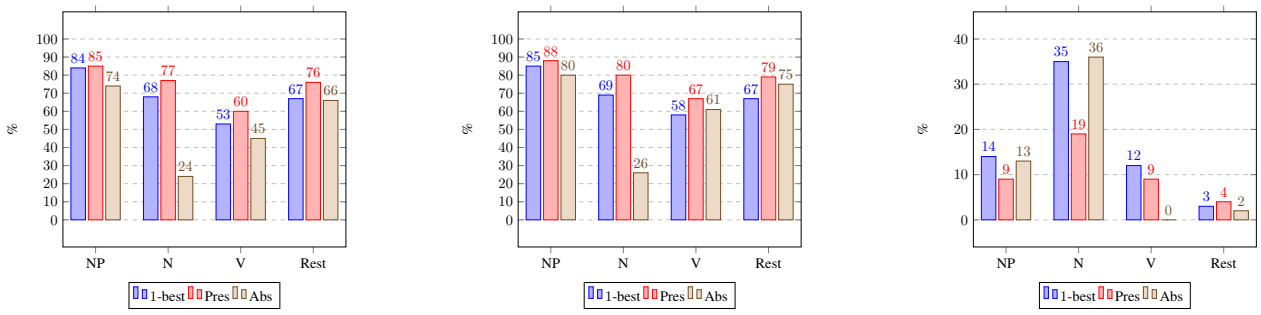
The manual analysis of the "worst" ($M <= 20\%$) and "best" ($M >= 80\%$) translated unit occurrences for the Cochrane SMT within the target groups provides some insight as to the nature of the residual errors (see Table 8).

Confirming our previous observations, the remaining errors of the N and NP groups concern mainly terms unknown to the model (out-of-vocabulary (OOV)), as well as errors in term and professional jargon precision (e.g., Source: "*cardiotoxicity*", MT: "*cardiotoxicité*" 'cardiotoxicity', PE: "*toxicité cardiaque*" 'cardiac toxicity', absent from the oracle hypothesis; Source: "*IDA*", MT: "*une anémie ferriprive*" 'iron deficiency anemia', PE: "*l'IDA*" 'IDA', absent from the oracle hypothesis).

In the NP group we often face complex terminological constructions translated by composition (e.g., Source: "*people with functioning kidney transplants*", MT: "*les personnes atteintes de fonctionnement de greffes de rein*" 'people suffering from functioning of kidney transplants', PE: "*des receveurs de greffe rénale fonctionnelle*" 'functional renal transplant recipients', absent from the oracle hypothesis).

The residual translation errors related to the V group are mostly caused by the specificities of the source language:

1. source syntactic/stylistic peculiarities (very often expletive constructions), requiring restructuring on the target language side (see Table 9);
2. tense and modality (e.g., Source: "*may reduce*", MT: "*peut réduire*", Oracle: "*peut réduire*" 'can reduce', PE: "*pourrait réduire*" 'could reduce').

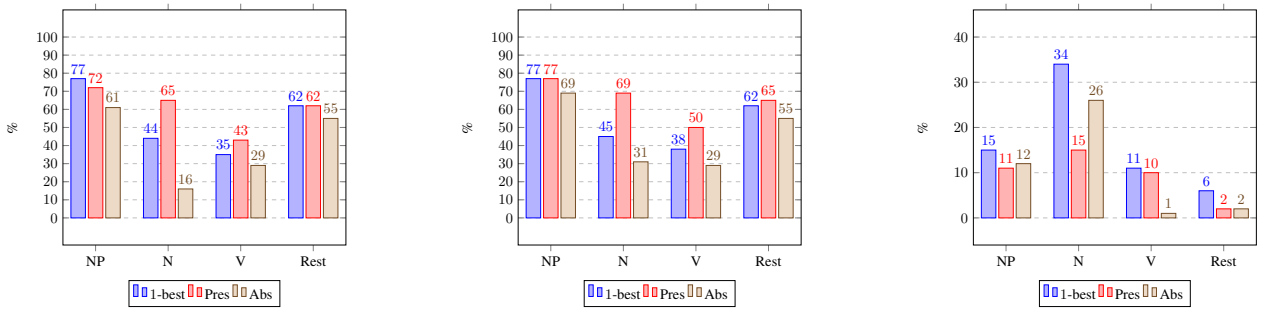


(a) Average percentage of matches per unit group

(b) Average percentage of oracle matches per unit group

(c) Average term rate per unit group

Figure 2: Translation quality statistics for Cochrane SMT



(a) Average percentage of matches per unit group

(b) Average percentage of oracle matches per unit group

(c) Average term rate per unit group

Figure 3: Translation quality statistics for WMT'14 SMT

	NP				N				V			
	PLS		ABS		PLS		ABS		PLS		ABS	
	Worst	Best	Worst	Best	Worst	Best	Worst	Best	Worst	Best	Worst	Best
$\#\ total$	1641		2495		304		365		1206		1604	
%	5	57	4	64	21	66	16	73	33	58	31	61
$\#\ w_k$	2	5	3	7	1	1	1	1	1	1	1	1
fr	3	2	2	2	2	4	3	7	15	23	26	30
M,%	0	100	0	100	0	100	0	100	0	100	0	100
M_O,%	34	100	30	100	27	100	25	100	18	100	16	100
$term\ rate$,%	11	10	15	10	44	18	28	20	6	6	14	10
H_{post}	0.5	0.3	0.4	0.2	0.3	0.3	0.4	0.2	0.4	0.4	0.4	0.4

Table 8: Statistics about the "worst" and "best" translated unit occurrences

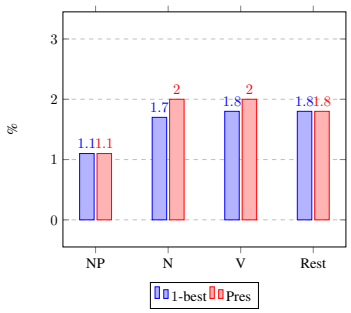
We should also notice an increased quantity of paraphrasing corrections performed to the V group (e.g., Source: "*we **searched** all databases*", MT: "*nous avons **effectué des recherches** dans toutes les bases de données*" 'we have performed searches in all the databases', PE: "*nous avons **interrogé** toutes les bases de données*" 'we have questioned all the databases', oracle output corresponds to MT). Those rephrasings have a negative impact on the automatic evaluation metrics. The semantic and stylistic necessity of those changes need further investigations.

In comparison, stylistic changes within NP and N groups are quite rare (e.g., Source PLS: "*the Canadian Institutes of Health Research*", MT: "*la Canadian Institutes of Health Research*" 'the Canadian Institutes of Health Research', Oracle: "*la Canadian Institutes de recherche en santé de recherche*" 'The Canadian Institutes of research in health of research', PE: "*les instituts de recherche en santé du Canada*" 'the institutes of research in health of Canada').

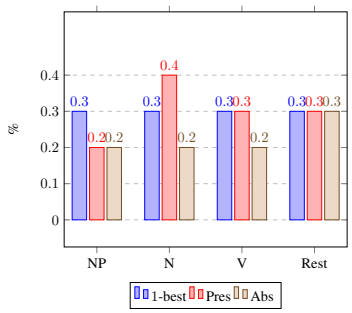
4. Which kinds of residual errors could be potentially resolved by the high-performance system given its training data?

We also performed a manual analysis of the oracle improvements to the "worst" translated unit occurrences within the target groups (ΔM of about 25%, see Table 8). They mostly concern:

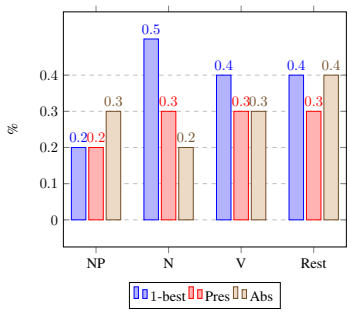
1. grammatical errors (change of article or preposition for the N and NP groups, e.g., Source: "*with taxanes*", MT: "*avec taxane*" 'with taxane', PE: "*avec les taxanes*" 'with the taxanes', oracle output corresponds to PE; tense changes for the V group, e.g., Source: "*were excluded*", MT: "*ont été exclues*" 'have been excluded', PE: "*étaient exclues*" 'were excluded', oracle output corresponds to PE);
2. certain reformulations (e.g., Source: "*the trial ... **showed** a clear benefit*", MT: "*l'essai ... **a montré** un bénéfice clair*" 'the trial ... **has shown** a clear evidence', PE: "*l'essai ... **a mis en évidence** un bénéfice clair*" 'the trial ... **has highlighted** a clear evidence', oracle output corresponds to PE);
3. some terminological precision errors, including terminological construction translated by composition (e.g., Source: "*alternative treatments*", MT: "*d'autres traitements*" 'other treatements', PE: "*des traitements alternatifs*" 'alternative treatements', oracle output corresponds to PE; Source: "*wound management properties*", MT: "*la prise en charge de la plaie propriétés*" 'the management of the wound any properties', PE: "*les propriétés*" 'the properties', oracle output corresponds to PE);



(a) Average translation prior entropy per unit group (Cochrane SMT)



(b) Average translation posterior entropy per unit group (Cochrane SMT)



(c) Average translation posterior entropy per unit group (WMT'14 SMT)

Figure 4: Entropy Estimations

Source	However, the evidence for survival improvement is still lacking.
MT	Cependant, les preuves d'amélioration de la survie est encore manquantes. 'However, the proofs of the improvement of survival is still missing.'
Oracle	Cependant, les preuves d'amélioration de la survie, il manque toujours de la. 'However, the proofs of the improvement of survival, it misses still the.'
PE	Cependant, il manque toujours de données probantes sur l'amélioration de la survie. 'However, it still misses the proving data on the improvement of survival.'

Table 9: Sentence restructuring example

4. minor (rarely major) reformulations and restructurings (e.g., Source: "*a one-day training course on* ***how to resuscitate*** *newborn babies*", MT: "*un schéma d'évolution de formation* ***sur la façon de réanimer*** *des nouveau-nés*" 'a scheme of development of training on the way to resuscitate newborns', Oracle: "*un schéma d'évolution de formation* ***sur la réanimation*** *des nouveau-nés*" 'a scheme of development of training on the resuscitation of newborns', PE: "*une formation d'un jour* ***sur la réanimation*** *des nouveau-nés*" 'a training of one day on the resuscitation of newborns').

As a summary, we can enumerate the following main translation difficulties faced by our Cochrane MT system:

1. term and professional jargon translation precision;
2. translation of complex terminological constructions;
3. translation of source-specific syntactic/stylistic constructions requiring target-side reformulation;
4. translation of verbs (grammatical/stylistic variant).

We tend to relate those difficulties to the nature of the medical translation task, since they are not specific to the high-performance system. They are caused by the original corpus limitations (absence of the "correct" translation in the training data), as well as to the limitations of SMT in general. Those limitations include the inability to resolve structural differences between languages or to take the more distant context into account.

The indicated issues can be partially solved by *ad hoc* solutions (fine-tuning of the system parameters to improve scoring, model separation to resolve stylistic differences, rewriting of source sentences, etc.), though their final resolution requires professional human knowledge.

6. Conclusion

In this article, we have introduced a fine-grained analysis methodology for high-quality narrow-domain SMT, which are typical situations where automatic error metrics prove not informative enough to guide the improvement of systems. Such levels of high performance, however, require adapted solutions.

The novelty of the proposed approach consists in diagnosing high-performance MT by finding an interconnection between residual errors, source phenomena and system parameters, such as original corpus quality and system scoring procedure, and using post-editing traces and Quality Estimation techniques. Thus, this approach provides some necessary hints to better detect translation difficulties and identify their reasons.

It can be used as an effective means to explore a system's potential with the perspective of improving it further.

We have demonstrated the usefulness of such an analysis on the example of the high-quality medical Cochrane SMT system. We found that its residual errors most significantly concern terminology and professional jargon, which are caused by the original corpus limitations, as shown by oracle estimations. The other main difficulty is the syntactic and stylistic peculiarities of the source language, often requiring reformulations on the target side. Those difficulties are related to the nature of the medical translation task and are not specific to the high-performance MT, as confirmed by our comparative study.

The described analysis procedure can be further extended by introducing an algorithm that will make a decision on the translation difficulty of a text given a system. This final decision can be provided as a difficulty score.

We also presented a high-quality English-French parallel corpus of Cochrane systematic review abstracts, which can be used for a variety of NLP tasks. We provided a description of the corpus (human translated and PE parts), as well as the translation challenges related to the genre of medical reviews with its internal stylistic variety (popular scientific vs. scientific style).

Acknowledgments

The work of the first author is supported by a CIFRE grant from the French ANRT.

7. Bibliographical References

Bannard, C. and Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Berka, J., Bojar, O., Fishel, M., Popović, M., and Zeman, D. (2012). Automatic MT error analysis: Hjerson helping Addicter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2158–2163, Istanbul, Turkey, May.

Boguraev, B., Manandise, E., and Segal, B. (2015). The bare necessities: Increasing lexical coverage for multiword domain terms with less lexical data. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 60–64, Denver, Colorado, June.

Bojar, O. (2011). Analyzing error types in English-Czech machine translation. In *The Prague Bulletin of Mathematical Linguistics*, April.

Costa-jussà, M. R., Farrús, M., and Pons, J. S. (2012). Machine translation in medicine. a quality analysis of statistical machine translation in the medical domain. In *Proceedings of the 1st Virtual International Conference on Advanced Research in Scientific Areas*, pages 1995–1998, December.

de Gispert, A., Blackwood, G. W., Iglesias, G., and Byrne, W. (2013). N-gram posterior probability confidence measures for statistical machine translation: an empirical study. *Machine Translation*, 27(2):85–114.

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Irvine, A., Morgan, J., Carpuat, M., III, H. D., and Munteanu, D. (2013). Measuring machine translation errors in new domains.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June.

Max, A., Crego, J. M., and Yvon, F. (2010). Contrastive lexical evaluation of machine translation. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. pages 311–318.

Popovic, M. and Ney, H. (2011). Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37:657–688.

Porter, M. F. (2001). Snowball: A language for stemming algorithms.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50.

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

Snover, M., Madnani, N., Dorr, B. J., and Schwartz, R. (2009). Fluency, adequacy, or HTER?: Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 259–268, Stroudsburg, PA, USA.

Sokolov, A., Wisniewski, G., and Yvon, F. (2012). Computing lattice BLEU oracle scores for machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 120–129, Avignon, France, April. Association for Computational Linguistics.

Specia, L. and Giménez, J. (2010). Combining confidence estimation and reference-based metrics for segment-level MT evaluation. In *Ninth Conference of the Association for Machine Translation in the Americas*, AMTA, Denver, Colorado.

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

Specia, L., Paetzold, G., and Scarton, C. (2015). Multilevel translation quality prediction with QuEst++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China, July.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

UMLS. (2009). UMLS reference manual. *Multidisciplinary Information Retrieval*.

Wang, L., Lu, Y., Wong, D. F., Chao, L. S., Wang, Y., and Oliveira, F. (2014). Combining domain adaptation approaches for medical text translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 254–259, Baltimore, Maryland, USA, June.

Wisniewski, G., Pécheux, N., Allauzen, A., and Yvon, F. (2014). LIMSI submission for WMT'14 QE task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 348–354, Baltimore, Maryland, USA, June.