Neutralizing the Effect of Translation Shifts on Automatic Machine Translation Evaluation

First Author, Second Author, and Third Author

Affiliation / Address line
e-mail@domain
Affiliation / Address line
e-mail@domain
Affiliation / Address line
e-mail@domain

Abstract. State-of-the-art automatic Machine Translation [MT] evaluation is based on the idea that the closer MT output is to Human Translation [HT], the higher its quality. Thus, automatic evaluation is typically approached by measuring some sort of similarity between machine and human translations. Most widely used evaluation systems calculate similarity at surface level, for example, by computing the number of shared word n-grams. The correlation between automatic and manual evaluation scores at sentence level is still not satisfactory. One of the main reasons is that metrics underscore acceptable candidate translations due to their inability to tackle lexical and syntactic variation between possible translation options. Acceptable differences between candidate and reference translations are frequently due to what is called "optional translation shifts". It is common practice in HT to paraphrase what could be viewed as close version of the source text in order to adapt it to target language use. When a reference translation contains such changes, using it as the only point of comparison is less informative, as the differences are not indicative of MT errors. To alleviate this problem, we design a paraphrase generation system based on a set of rules that model prototypical optional shifts that may have been applied by human translators. Applying the rules to the available human reference, the system generates additional translation references in a principled and controlled way. We show how using linguistic rules for the generation of additional references neutralizes the negative effect of optional translation shifts on n-gram-based MT evaluation.

Keywords: Translation shifts, Machine Translation Evaluation, Paraphrase Generation

1 Introduction

One of the most important observations in the field of translation studies is that a translated text can differ from the original at any linguistic level – lexical, syntactic, discourse – and still be considered perfectly acceptable. The departures from theoretical formal correspondence between source and target language units for the sake of textual equivalence are denominated translation shifts [1]. It is one of the key concepts in translation theory. Apart from the obvious transformations necessary for grammatical well-formedness, it is common practice in translation to introduce optional changes to the way information is presented in the source text. Although such changes are not strictly necessary, they are part and parcel of Human Translation [HT], as professional translators are expected to adapt the original to the norms and conventions of target language use depending on the text genre, text type, register, means of communication, etc.

The distinctive properties of translated texts have been extensively studied both in the field of translation theory and in computational linguistics. Surprisingly, they have rarely been discussed in the field of automatic MT evaluation, although the vast majority of evaluation systems are actually based on the degree of similarity between MT and HT. As similarity is normally calculated at surface level, the performance of the metrics depends on the availability of a heterogeneous set of human reference translations. In practice, however, only one reference is available and its characteristics can strongly affect the results of automatic evaluation. As an illustration, consider Table 1, which shows an example of English-Spanish MT evaluated manually and automatically (back translation to English is given in square brackets and relevant constructions are marked in bold).[1] Manual evaluation is scaled from 1 to 4 and automatic evaluation score is produced by state-of-the-art BLEU evaluation system [2] (BLEU scores range from 0 to 1).

Table 1. Example of passive/active alternation in reference translation

Source	All these activities should be monitored and supported by parliament.		
Reference	El parlamento debería controlar y apoyar todas estas actividades. [The parliament **should control and support** all these activities]	**Human**	**BLEU**
Candidate	Todas estas actividades debería ser controladas y apoyado por el parlamento. [All these activities **should be controlled and supported** by the parliament]	4	0,1783

[1] Here and in what follows examples are extracted from English-Spanish MT evaluation data set used in the present work (see Section 5 for the description).

Here the English analytic passive construction is transformed into an active clause in the reference, whereas in MT no such changes are introduced and source structure is preserved. However, MT obtains maximum score in manual evaluation. Clearly, human evaluators do not penalize the absence of optional changes if the sentence is well-formed and delivers the contents of the original. By contrast, the score produced by BLEU, which is based on n-gram matching, is extremely low because of the small number of shared word sequences between candidate and reference translations. If along with the available reference, other translation options preserving the source analytic passive construction were provided, BLEU could do a much better job in approaching human assessment.

To analyze the actual impact of optional translation shifts on automatic MT evaluation, we developed a paraphrase generation system, which is based on a set of hand-crafted transformation rules that "undo" optional shifts in HT. The system is designed for English-Spanish translation. We focus on the syntactic aspect of linguistic variation as lexical issues have been already addressed in the literature (see, for example, [3]). For evaluation, we performed a detailed manual annotation of structural changes in an English-Spanish parallel corpus and measured the proportion of cases where using additional references produced by our system improves automatic evaluation score.

The rest of this paper is organized as follows. In Section 2 we briefly describe the background of our work. Section 3 introduces the related work. Section 4 describes the paraphrase generation system. In Section 5 experiments and results are presented. Finally, in Section 6 we give the conclusions and discuss future work.

2 Background

Translation process is conditioned by the tension between two prototypical expectations: that of maximal similarity between source and translated texts and that of naturalness of the translated text in the target language. In terms of the distance between source and target texts, researches distinguish between analogous, equivalent and contextually appropriate translation [4]. Analogous translation involves similarity in form, as the translation retains as many forms of the original as possible. Equivalent translation gives priority to the semantic content, retaining the propositional meaning of the original as closely as possible. Finally, contextually appropriate translation optimizes discourse relevance and text processing conditions taking into account broad linguistic and extra-linguistic context without caring much for adherence to structure or lexis of the source.

The latter type of translation is the most common in practice. Translators normally "shift" away from the original and paraphrase what could be viewed as its close version. Optional shifts occur when formally similar structures have different semantic and/or pragmatic values in the languages involved [1]. Even in typologically related languages, where formally similar structures are available in many cases, not only we find different lexical and grammatical devices but also different uses of analogous

lexical and grammatical means guided by language-specific principles of language use including stylistic issues and discourse processing conditions.

Here it should be noted that linguistic variation between possible translations of the same sentence is not only given by the presence or absence of optional translation shifts. It may occur when no formally equivalent construction is actually available in the target language and obligatory changes may be performed in various ways. Furthermore, alternative translation options can contain marked uses of language. For example, phrases occupying unmarked position in the source sentence may be topicalized in translation involving a change with respect to the original word order, due to the differences in discourse processing and information structure preferences in the source and target languages. Note that in MT it is improbable to find such changes, as the unmarked options are normally the most frequent ones.

As we aim to generate translation alternatives in target language, we studied available formal descriptions of linguistic paraphrase [5,6] as well as translation shifts classifications [7,8,9]. Based on these works, we developed the following typology[2] of translation phenomena, which was used for the design of transformation rules and for the evaluation of our paraphrase generation system (see Section 5).

1. Changes in grammatical features (finiteness, mood, modality, tense, aspect, etc.)
2. Changes in grammatical category (pronominalization, nominalization, manner adverbial → predicative adjective, etc.)
3. Diathesis changes (passive construction → active construction, personal clause → impersonal clause)
4. Level and function changes (phrase → clause, main clause → subordinate clause, temporal clause → conditional clause, locative-possessive alternation, etc.)[3]
5. Word order change (subject-predicate inversion, clitic climbing, changes in the position of adverbial modifiers, etc.)
6. Changes in the number of constituents (ellipsis, additions of content words, deletions of content words)

As mentioned earlier, some changes are mandatory, as they are related to systemic differences between the languages involved, while others are not strictly required to produce a faithful and grammatically well-formed translation. In the present work we are interested in the differences between MT and HT induced by the presence of optional shifts in the latter, therefore only those were considered for the classification.

3 Related Work

Research in MT evaluation has demonstrated that the performance of the metrics improves significantly if a heterogeneous set of references is provided. In practice,

[2] Individual changes that are entailed by other operations are not annotated separately. For instance, inversion of arguments induced by diathesis alternation is not considered in the category of word order changes.

[3] Changes pertaining to this category have not been implemented, so we do not consider them in the following discussion.

however, only one human reference is available for evaluation, and attempts have been made to generate additional references automatically [10]. For this purpose, data-driven methods are normally used [11]. Data-driven approaches have the advantage that information is automatically extracted from the data. However, they are not suitable for dealing with long-distance structural changes. More importantly, they do not allow inspecting the type of MT-HT differences that have been neutralized by using additional references.

A more similar approach to ours is developed by [9] who propose a linguistic framework for formally codifying close translation properties. The authors state that close translation is the limit of MT performance. This is arguable because modern statistical MT actually tries to model translation decisions in context and can do better than close translation. We put into practice the idea presented in [9] by designing a system that generates close translation options automatically. We consider, however, that both shifted and close translation alternatives should be used for evaluation to be able to make a more fine-grained comparison of the quality of translations produced by different systems.

4 System Description

Our paraphrase generation system is intended to enhance MT evaluation with additional automatically generated references. The rationale behind the selection of particular paraphrase rules was their relevance in the context of English-Spanish translation. We defined sets of target language constructions that are approximately semantically equivalent in Spanish but are given different uses in source and target languages or no formal equivalent is available on either side. Thus, we expect that these constructions will be involved in prototypical structural changes in HT and are to be transformed in order to generate close translation options.

Table 2 presents the transformation rules we implemented. The rules are put in relation to the structural shifts typology presented in the previous section.

Table 2. Transformation rules for paraphrase generation

Translation shifts	Transformation rules
Grammatical Features	Past Simple ↔ Present Perfect Present Simple ↔ Present Perfect Present Simple ↔ Past Imperfective Simple Verb Form → Progressive construction Simple Future ↔ Periphrastic Future Recent Past Periphrasis → Present Perfect + 'recently' Habitual Aspect Periphrasis → Simple Verb Form + 'normally' Repetitive Aspect Periphrasis → Simple Verb Form + 'again'
Grammatical Category	Nominalization ↔ Denominalization Prepositional Phrase → Adverbial Modifier Copulative Clause → Adverbial Modifier
Diathesis	Active → Analytic Passive Synthetic Passive → Analytic Passive Personal → Impersonal
Word Order	Post-verbal Subject → Pre-verbal Subject Pre-verbal Adverbial ↔ Post-verbal Adverbial VP-External Adverbial → VP-Internal Adverbial Sentence-initial Adverbial ↔ Post-verbal Adverbial Sentence-initial Detached PP ↔ Post-verbal Detached PP Pre-nominal Adjectival Modifier → Post-nominal Adjectival Modifier Clitics before VP ↔ Clitics after VP
Addition / Deletion	Personal Pronouns [+Subject function] Repeated Prepositional Heads in Coordinated PPs

For cases where the direction of optional change in HT cannot be predicted, the rules are applied in both directions (marked with bi-directional arrows in Table 2). For example, the English verb forms in *-ing* with nominal function may be translated by nouns or infinitives in Spanish and in this case we cannot say that one option is closer to the source sentence than the other.

It should be noted that some of the constructions in Table 2 may be considered equivalent only given a specific linguistic context (for instance, tense alternations). We do not use any source-side information and thus applying such rules may result in

paraphrases that change the contents of the original. However, these are not supposed to affect the results because the paraphrases are to be used together with the true human reference. Thus, in case human translator has changed a source structure that is preserved in MT, using the relevant paraphrase increases automatic evaluation score. If it is not the case, additional reference is not supposed to have any effect on the evaluation.

The system operates on dependency trees in CONLL format and returns full transformed sentences. At the analysis phrase, the structures to be transformed are identified by means of regular expression matching. In addition to syntactic information, grammatical dictionary of Spanish Resource Grammar [12] with verb frame information is used to introduce lexical restrictions for rule application.

If the conditions are matched and no restrictions are found, at the generation phase the system reconstructs the sentence with relevant changes using information extracted from the parses of the input sentences and morphological dictionary look-up in order to generate the appropriate word forms. From one input sentence, the system generates a set of paraphrases (as many as there are rules applied).

Note that not all of the relevant operations that have been described in Section 2 can be efficiently modelled in this way. One problem is that in cases of deletion operations in HT where content is left implicit, we lack information to reproduce it. Furthermore, no quality language processing tools are yet available for analyzing certain complex phenomena. For example, in case of pronominalization shift, when full noun phrase is substituted by pronoun, which frequently happens in translation in order to avoid repetition, we lack high-quality co-reference resolution tools to reconstruct the original full noun phrase.

5 Experiments and Evaluation

The aim of the evaluation was twofold. In the first place, we wanted to assess the performance of the paraphrase generation system intrinsically. In the second place, we aimed to test the impact that translation phenomena discussed above have on MT evaluation. That is to say, we wanted to see how, in case optional transformations occur in HT, using additional references generated by the system affects automatic evaluation score.

For that purpose, in the first place, a parallel corpus annotated with translation shifts was necessary. There are parallel corpora in which translation shifts are annotated for some languages [8], [9], but no such resource is available for English-Spanish translation. In the second place, data set with manual evaluation scores for MT is required. We decided to carry out manual annotation and classification of translation shifts on sentences extracted from [13] MT evaluation data set. The data set consists of 4,000 source sentences in English, their corresponding reference translations to Spanish randomly extracted from Europarl [14], the translations of four statistical MT systems and manual evaluation scores. MT systems were trained with data from the same domain. Manual evaluation scores were provided by professional

translators using post-editing criterion.[4] It should be noted that Europarl is especially relevant for our work because it is the most widely used in MT development and evaluation and thus it is important to discuss the characteristics of the reference translations that are part of the corpus.

We randomly selected 290 sentences from the data set. Optional structural shifts in reference translations were annotated and classified manually using the typology presented in Section 2. The sentences were processed using MaltParser dependency parser [15] with Spanish models[5] and paraphrase generation system was applied to HTs to produce a separate reference set for each group of rules. MTs were automatically evaluated with BLEU in a single reference baseline scenario and in a multi-reference scenario, with automatically generated paraphrases. Note that when multiple references are provided BLEU takes into consideration n-gram matches between candidate translation and each of the references, which allows assessing the impact of different types of translation phenomena on evaluation.

We compared the resulting BLEU scores and calculated precision and recall based on the following principles. The purpose of using additional references was to increase BLEU scores for cases when a translation shift occurs in HT while MT contains the corresponding structure that is formally equivalent to the source. Thus, for each group of rules we considered that the application was successful if using the respective set of paraphrases increases BLEU score and the corresponding translation shift occurs in HT (true positives).

By contrast, rule application was considered unsuccessful when no translation shift of certain type occurs in HT and applying the corresponding set of rules increases the evaluation score (false positives), or when there is an optional change in the reference and applying the corresponding set of rules does not increase BLEU score (false negatives).

As mentioned earlier, our system currently covers a limited set of translation phenomena. Therefore, in order to assess the performance of the system per se, we counted recall separately for all the annotated translation shifts vs. translation shifts modelled by the rules. Table 3 presents precision and recall for each group of rules as well as the frequency of the translation phenomena involved.

---

[4] They were asked to indicate the amount of editing needed to make the MT ready for publishing, on a four-point scale: 1 - requires complete retranslation; 2 - a lot of post editing needed; 3 - a little post editing needed; 4 - fit for purpose.

[5] Available at http://www.iula.upf.edu/recurs01_mpars_uk.htm

Table 3. Precision and Recall for rule application and Frequency of translation shifts

Rule sets	P	R (all)	R (modelled)	Freq
Grammatical Features	0.76	0.43	0.60	104
Grammatical Category	0.70	0.30	0.61	77
Diathesis	0.59	0.20	0.43	66
Word order	0.79	0.40	0.72	151
Addition / Deletion	0.61	0.23	0.56	82
Total	**0.69**	**0.32**	**0.58**	**480**

The results must be interpreted as follows. The overall precision indicates that in 70% of cases of rule application the system successfully reconstructs the close translation option and using it as additional reference increases BLEU score. As expected, the recall is low in case all translation phenomena are considered, and much higher if calculated only for the phenomena covered by the rules. Thus, intrinsically the system shows good performance in the cases it is designed to deal with. The overall number of translation shifts is high as there are an average of 1.7 optional changes per sentence in the reference, confirming the idea that such changes are indeed common practice in HT. An example of successful rule application is given in Table 4[6].

Table 4. Example of category change in reference translation

Source	this event , on the eve of the lahti meeting , is clearly of particularly crucial significance to us .
MT	este acontecimiento , en vísperas de la reunión lahti , es claramente de especialmente crucial importancia para nosotros . [this event , on the eve of the lahti meeting , is clearly of particularly crucial significance for us .] Human evaluation = 4 BLEU with HRT = 0.2477 BLEU with ART = 0.2610
HRT	está claro que este acontecimiento , en vísperas del encuentro de lahti , reviste para nosotros una especial trascendencia . [**it is clear that** this event , on the eve of the lahti meeting , represents for us a special importance .]
ART	claramente , este acontecimiento , en vísperas del encuentro de lahti , reviste para nosotros una especial trascendencia . [**clearly** , this event , on the eve of the lahti meeting , represents for us a special importance .]

In this example clause-level adverbial modifier is changed into predicative adjective in HT (with corresponding changes in sentence structure). This transformation is common in English-Spanish translation, as translators are advised to avoid excessive use of manner adverbials in *-mente* (-ly) considered a calque from English where they are more frequent. MT preserves the structural organization of the original, which

[6] HRT stands for Human Reference Translation and ART stands for Automatically-generated Reference Translation.

results in a sentence that is stylistically flawed, but is perfectly acceptable according to human evaluation score. The paraphrase generated by our system successfully neutralizes this shift in HT, and using it increases BLEU score. Note, however, that the increase is small as the system is not able to predict the exact position of the adverbial.

As far as specific groups of rules are concerned, the lowest results are for diathesis changes. In this group the most frequent transformation is reflexive passive → analytic passive. The resulting paraphrases are irrelevant, as they do not increase BLEU score because the corresponding shift frequently occurs in MTs. This is understandable given the nature of statistical MT. Since the change only involves local context and is consistently present in English-Spanish translations, it is expected to be found in high quality MT.

By contrast, word order changes are more challenging for statistical systems. For this reason, the group of rules that neutralize the optional changes affecting word order obtained the highest precision and recall. As an illustration, consider the example shown in Table 5.

Table 5. Example of diathesis change and subject-predicate inversion in human reference

Source	appropriate arrangements have been made for consultation with the member states .
MT	los preparativos apropiados se han hecho para su consulta con los estados miembros . [**appropriate arrangements MPASS[7] have made** for the consultation to the member states] Human evaluation = 4 BLEU with HRT = 0.3013 BLEU with ART1= 0.3013 BLEU with ART2 = 0.4683
HRT	se han realizado los preparativos apropiados para la consulta a los estados miembros . [**MPASS have made appropriate arrangements** for the consultation to the member states]
ART1	han sido realizados los preparativos apropiados para la consulta a los estados miembros . [**have been made appropriate arrangements** for the consultation to the member states]
ART2	los preparativos apropiados se han realizado para la consulta a los estados miembros . [**appropriate arrangements MPASS have made** for the consultation to the member states]

Here the reference contains two optional changes: the transformation from analytic passive to reflexive passive and subject-predicate inversion. The first paraphrase delivers the close version with analytic passive construction. The second paraphrase reconstructs the word order of the source sentence neutralizing the subject-predicate inversion present in HT. In the case of word order, the rule is applied successfully as

[7] MPASS stands for the Spanish passive marker "se".

it increases BLEU score. In the case of diathesis transformation, the shift occurs in both HT and MT and thus the transformation performed by our system is not relevant.

Another source of errors is that, contrary to our assumption, not using source-side information does introduce noise. This is the case, for example, when the transformation involves adding a function word that happens to be present in MT but does not form part of the same syntactic construction.

Finally, both precision and recall are affected by parser errors. For instance, order changes cannot be addressed in cases where the parser fails to identify the head of the element to be moved. Parser errors are especially harmful for rule-based approach as the patterns have to be defined in detail and the conditions need to be exactly satisfied for the rules to apply.

6 Conclusions and Future Work

Translation theory aims at explaining and predicting translators' behaviour. It is thus natural to use it in the field of MT. In present work, we bring together the research accomplished in the field of MT evaluation and theoretical notions from translation studies.

HT deviates from the source text in many ways making HT-MT comparison less informative for reference-based automatic MT evaluation. To show how this problem can be solved, we developed a rule-based paraphrase generation system for Spanish that produces additional translation options for English-Spanish automatic MT evaluation. We demonstrated that optional structural shifts have negative effect on the performance of evaluation systems, which can be neutralized by using additional references that contain close translation options.

The results show that different translation phenomena have different impact on evaluation scores. The relevance of the paraphrases produced by our system depends on the corpus (underlying HT strategy) and the type of MT. The test set used in the present work contains only statistical systems trained with data from the same domain. Therefore, considerable number of optional shifts that are regularly present in the reference is also found in MT. An idea worth investigating is that using the information on the type of reference (in our case, human or automatically generated) that MT is more similar to, quality levels can be defined and used to rate and describe the characteristics of a given system, making more fine-grained distinctions of MT quality.

The results are encouraging but certainly leave large room for improvement. We plan to augment the set of rules to perform a large scale evaluation of MT systems based on different strategies and assess the effect using the paraphrases have on the correlation with human judgments. Also, to alleviate the shortcomings of rule-based paraphrase generation, a hybrid approach in which some of the relevant operations are learnt automatically may be used.

References

1. Szymańska, I.: Mosaics. A Construction-Grammar-based approach to translation. Warszawa: Semper (2011)
2. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. RC22176 (Technical Report), IBM T.J. Watson Research Center (2001)
3. Denkowski, M., Lavie, A.: Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In: Proceedings of the EACL 2014 Workshop on Statistical Machine Translation (2014)
4. Doherty, M.: Language processing in discourse: a key to felicitous translation. London: Routledge (2002)
5. Barrón-Cedeño, A., Vila, M., Martí M., Rosso, P: Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Computational Linguistics 39(4), 917–947 (2013)
6. Bhagat, R., Hovy, E.: What is a paraphrase? Computational Linguistics 39(3), 463–472 (2013)
7. van Leuven-Zwart, K. M: Translation and original: Similarities and dissimilarities. Target 1(2), 151–181 (1989)
8. Cyrus, L.: Building a Resource for Studying Translation Shifts. In: Proceedings of the 5th International Conference on Linguistic Resources and Evaluation, 1240–1245 (2006)
9. Ahrenberg, L.: Codified Close Translation as a Standard for MT. In: Proceedings of the 10th Annual Conference of the European Association for Machine Translation, 13–22 (2005)
10. Owczarzak, K., Groves, D., Genabith, J. V., Way, A: Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, 148–155 (2006)
11. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: Proceedings of ACL (2005)
12. Marimon, M.: The Spanish DELPH-IN grammar. Language Resources and Evaluation 47(2), 371–397 (2013)
13. Specia, L., Turchi, M., Cancedda, N., Dymetman, M., Cristianini, N.: Estimating the Sentence-Level Quality of Machine Translation Systems. In: 13th Conference of the European Association for Machine Translation, 28–37 (2009)
14. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. MT Summit (2005)
15. Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E.: MaltParser: A language-independent system for data-driven dependency parsing. Natural Language Engineering 13(2), 95–135 (2007)