A Machine Translationness Typology for MT Evaluations

Joaquim Moré López and Salvador Climent Roca
Universitat Oberta de Catalunya (UOC)
jmore@uoc.edu; scliment@uoc.edu

Abstract. In this article we present machine translationness (MTness, henceforth) as an approach for MT evaluations, in contrast to the notion of human likeness which is the basic criteria for state-of-the-art automatic evaluations. Our approach is based on the detection of phenomena that contribute to the MTness of a translation and, consequently, its poor fluency. We present here a typology of MTness based on an empirical study in which human and machine translations were Turing tested. We also discuss the types of the typology that should be focused on in MT evaluations, the subjective quality of the human likeness assumption, and the advantages of the MTness approach, in terms of time and financial costs.

Keywords: machine translationness, human likeness, evaluation .

1 Introduction

Automatic MT evaluations are generally based on the assumption that everyone would agree in considering a translation made by a professional translator as a good translation (Human Translation Goodness, HTG henceforth). This is the basic idea behind metrics based on the Assumption of Reference Proximity (ARP) such as BLEU [1], evaluations which consist in classifying automatically translations as human (good) or machine (bad) [2] and the approach which combines different metrics in order to give a result which correlates the closest to the 'human likeness' of MT translations [3].

However, contrary to MT evaluation, little is said about evaluating human translations (HT), which should be a prior step if we want to use them in MT evaluations; their good quality is generally taken for granted, but sometimes some HT translations prove to be worse than their MT counterparts and human translators also use MT systems and present MT output as original [4]. Thus, we present an assumption we consider will be more widely – if not universally – agreed upon: translations that sound like machine translations are bad.

Actually, this assumption is implied in the methodology based on HTG: what sounds human-like is good, thus what sounds machine-like is bad. Therefore, if we focus not on the human-likeness but on the machine-likeness of a translation, that is to say, its machine-translationness (MTness), we could carry out fast evaluations on the fly that would convey a significant idea about the quality of a system and its main drawbacks ([5] and [6]). Depending on the purpose of the evaluation, focusing on examples of MTness is enough and saves the time and money-consuming task of preparing human translation references or collecting huge training corpora of machine

and human translations for the human-machine classifier. Likewise, the cost of evaluating the human translations would also be saved. Finally, the evaluation based on detecting MTness would use a more reliable criterion. For this reason, it is important to establish a typology of MTness features.

This paper presents an MTness typology from the output of 2 Catalan-Spanish and 2 English-Spanish MT systems. We also introduced Catalan-Spanish human translations in order to check whether MTness is a feature attributable to machine translations only. The typology is not the result of an aprioristic approach based on known MT error types but the result of an empirical study of how people with different learning backgrounds and different reading skills appreciate MTness.

The paper is organised as follows: in section 2 we briefly describe the methodology of the empirical study. In section 3 we present the typology with a brief description of each type. In section 4 we discuss the overlapping of some of these types in machine and human translations, according to the informants. Finally, in section 5 we present the conclusions and future lines of research.

2 Methodology of the Empirical Study

2.1 Description of the Experiment and its Goal

The empirical study consisted in performing what can be considered a Turing test. 100 people read a number of machine and human translations. For each translation they had to state whether it was a machine or a human translation, and when it was considered a machine translation they had to underline the pieces they found machine-like.

Our main concern was to typify the linguistic phenomena responsible for MTness that would be detected by several people, regardless of their reading skills and learning background. For this reason, each translation was read by three people with different cultural backgrounds, individually and in isolation. The typology was built by analysing the segments of the same translation that at least two people underlined.

The experiment was performed without any computational support in order not to restrict the informants to those with technological skills.

2.2 The Informants

The informants were people living in Catalonia, literate, over 16 years of age and, in order to avoid bias in the results because of expertise, they were not language experts and were not familiar with computational linguistics. Each set of translations was evaluated by three people of different ages and levels of reading comprehension. These levels were established according to their studies, as we assumed that the higher the level of studies, the higher the reading skills they must have in order to understand textual complexity and abstract contents. Although gender was not expected to be significant, we decided to balance the number of men and women (50

each).

2.3 Translation Corpus

The informants evaluated the output of two Catalan-Spanish and two English-Spanish MT systems, and human translations from Catalan into Spanish.

The translations were single sentences with no contextual relation to the previous and the following sentence. We decided to decontextualise the translations because MT systems translate sentence by sentence with no contextual information. We wanted the informants to face the same situation as the MT systems, with no context to fill in their comprehension gaps. Contextual interpretation distinguishes humans from machines so we did not want this human capacity to condition the informant when underlining examples of MTness.

We chose translations that were comprehensible for the informants, regardless of their learning and professional background. Sentences from news and tourism magazines were collected for the corpus. Moreover, in order to check how the knowledge of a domain may influence the detection of MTness, we mixed in sentences from articles about computers, economics, speeches from the Europarl corpus, and provisions and acts published in the official gazette of the Catalan government. Nonetheless, these sentences were written to be understood by the general public.

The sentences evaluated were the translations of 250 sentences in Catalan and 250 sentences in English. Both the Catalan and the English sentences were translated into Spanish by a rule-based system and by a statistical-based system. We decided to do so because, besides creating the MTness typology, we were also interested in finding out whether a rule-based or a statistical-based system produced more MTness than the other. Thus, the number of translations performed by the different systems amounted to 1000. These translations were replicated three times because we wanted each of them to be evaluated by three people (c.f. 2.2), in order to know the degree of agreement between informants when underlining segments with MTness. We also added 750 human translations of the Catalan sentences, which was equivalent to 25% of the total corpus.

These volumes were established according to the representativeness of the data collected and the viability of the experiment. We had to take into account that a large number of translations for each informant would have caused fatigue and attention span problems, which would have had a negative effect on the objectivity of the results. In the end, each informant had to evaluate about 38 sentences, a number we considered viable.

2.4 Language Pairs

We chose English and Catalan as source languages because we were interested in knowing whether the closeness between source and target languages, being the pair Catalan-Spanish the closer and the pair English-Spanish the further apart, influenced the MTness of translations. The reasons why we chose Spanish as the target language

were, on the one hand, the availability of rule-based and statistical MT systems for Catalan->Spanish/English->Spanish directions and, on the other hand, because we would not have felt confident about the judgements of older informants on translations in Catalan or English, especially those with basic learning levels. Apart from not having learned English, in their youth they had to study in Spanish and their reading was basically in Spanish, as teaching and publishing in Catalan was prohibited. This is the reason why Spanish was more likely to be known by the informants.

3 MTness Typology

After analysing the segments underlined by at least two informants we established a list of 14 types that we have classified in four groups: lexical, syntactic, semantic and formatting. We describe each group of types. The name of the type is followed by a code, which will be used in later references.

3.1 Lexical Types

3.1.1 Words not Pertaining to the Target Language (NO-L2)

These are words which are not recognised as pertaining to the target language. For example, in (1) 'Missatges' is a Catalan word.

(1) Protocolo de acceso IMAP (Internet Message Access Protocolo o Protocolo de Acceso de **Missatges** de Internet)

3.2 Syntactic Types

3.2.1 Wrong Syntactic Agreement (W-AGR)

This type covers the lack of agreement between subject and verb, adjective and noun, and so on. Here are two examples of errors in syntactic agreement.

(2) i) Las ayudas estatales no **debe** seguir adelante
ii) Los gobiernos son víctimas de **sus propios** laberinto

In (2i) the verb in the third person singular ('debe') does not agree with the subject in the plural ('Las ayudas estatales'). In (2ii) the determiner ('sus') and the adjective ('propios') are in the plural whereas the noun ('laberinto') is in the singular.

3.2.2 Inadequate Constituent Order (I-ORD)

This is the type for ungrammatical orderings of syntactic constituents. Here are some examples.

(3) i) Víctimas de la **española represión**

ii) He valorado mucho **del presidente Prodi declaraciones**

iii) El PNV recuerda a Zapatero que el Pacto del Tinell también permite una consulta si **el Congreso lo rechaza catalán.**

In (3i) the adjective 'española' is wrongly placed in a prenominal position. In (3ii) we see an example where the PP appears before the noun it complements. The proper order is *declaraciones del presidente Prodi.* Sometimes the incorrect ordering causes difficulties in understanding the sentence. For instance, in (3iii) the proper ordering should have been *si el Congreso catalán lo rechaza* but the position of the adjective after the verb makes the sentence difficult to understand.

3.2.3 Syntactic Gap (SYNT-GAP)

A syntactic gap is a missing constituent that should have appeared according to the argument structures of verbs and nouns and other syntactic restrictions. For instance, in (4) the direct object of the verb 'retornar' is missing.

(4) El Senado veta los presupuestos **y retorna al Congreso** por primera vez en la democracia.

Examples of this type are detected because there are combinations which are not grammatical, as in (5) where a gerund form cannot appear between a verb in a finite form ('está') and another verb in infinitive ('incorporar'). The informant expected a preposition to bridge them.

(5) Actualmente, el consejo está **hablando** incorporar esos mecanismos en el artículo 7.

3.2.4 Word Overgeneration (OVER-WRD)

This is the case when a word, or a sequence of words, does not perform any syntactic, semantic or cohesive role in the sentence. By deleting them, the sentence often makes more sense, as in (6) where the verb 'sigo' alone is enough to convey the meaning of the sentence.

(6) Quiero dar las gracias al comisario por su introducción, pero sigo **estando** con algunas preguntas.

Many of these cases are detected because of a non-grammatical part of speech (POS) combination, as in (7), where the combination *preposition+preposition+conjunction* is not allowed.

(7) Una serie de enmiendas que deben hacerse **en** para que la propuesta separación de poderes trabajo.

3.2.5 Word Repetitions (WRD-REP)

These are the cases where two identical word-forms with the same POS are in one syntactic constituent or very close to each other, as in (8i) and (8ii)

(8) i) Por consiguiente, negociamos en **un un** minima base y tenemos una

mínima carta, en particular respecto de los derechos sociales.
ii) El **paseo** resulta un duro **paseo** en la mesa de trabajo de Mac o dentro de Mi Ordenador.

3.2.6 Inadequate Part of Speech (I-POS)

In this case, the POS of a word is inadequate according to the context in which it appears. For instance, in (9i), an adjective cannot appear ('concreto') after the auxiliary verb 'haber', and in (9ii), the verbal form 'cenar' instead of the noun 'cena' is expected.

(9) i) He **concreto** mencionado algunos de los factores que nos permita para determinar inmediatamente si la reforma es suficiente o no.
ii) Se traen por ir a la playa, pero también por salir a **cena**.

3.2.7 Inadequate Verbal Form (I-VERBF)

This type covers non-finite verbs that should have appeared in finite forms and vice versa. For example, in (10) after the pronominal 'se', a verb in participle cannot appear because 'se' always precedes finite verbs.

(10) Queremos, a petición de que se **diferida** por tercera vez por razones políticas.

Inconsistencies of the verbal mood (indicative and subjunctive) are also covered in this type.

3.3 Semantic Types

3.3.1 Semantic Gaps (SEM-GAP)

Semantic gaps are missing constituents that are necessary to understand the sentence. For instance, in (11) the noun that the adjective is expected to modify is missing.

(11) Por último, necesitamos una definición más precisa de **la relevantes** del mercado, porque cada vez más, el mercado no es el mercado nacional.

SEM-GAP is different from SYNT-GAP because the latter is not linked to interpretation of the sentence. Indeed, a correct interpretation of the sentence leads to the reader's detecting the missing syntactical constituent, as we have seen in (4).

3.3.2 Semantic Incoherence (SEM-INCOH)

This type covers syntactic constituents which do not fit the semantic restrictions of the noun or the verb. For instance (12).

(12) **Los Bocados detectan** al Vallès una reavivada de asaltos nocturnos a viviendas.

"Bocados" (bites) is the mistranslation of the name for the Catalan police force (Mossos). This makes the interpretation of the sentence absurd because the subject

does not fit the selectional semantic restrictions of the verb ‘detectar’ (to detect).

3.3.3 Noisy Segments (NOI-SEG)

Noisy segments are those that make the sentence absolutely incomprehensible. They cannot be classified in a specific type because they are the result of the confluence of different syntactic and semantic resolutions which are not replicated in other translations. In (13) we see an example.

(13) Robert era un golfista ávido, cántara de softball, y **bombín quienes ganaban** un Campeonato de Estado en 1951

3.3.4 Contextual Incoherence (CON-INCOH)

Words and constituents that are contextually incoherent are those that do not fit the context of the discourse where they appear. Unlike SEM-GAP and SEM-INCOH, they do not violate selectional semantic restrictions imposed on the structure of the sentence. For instance, ‘bloody’ can be translated as ‘sangrante’ or ‘sangrienta’ but ‘dia sangrante’ is not correct for the context in (14i). In (14ii), the translation of the Catalan verb *volem (*we want/we fly) into Spanish as *volamos* (we fly) makes the translation incongruent.

(14) i) Una cincuentena de muertes en otro **día sangrante** en Irak.
ii) Es el Estatuto que, a día de hoy, Catalunya necesita y los catalanes **volamos**

Other CON-INCOH are linguistic phenomena such as apocopation (primero/primer) where context does not affect the meaning of words but their form.

(15) **El primero** ministro de Ucrania impugna las elecciones y reitera que no dimitirá.

3.4 Formatting Types

These types are TYPO-E, which covers the inadequate use of upper case and lower case, missing or inadequate punctuation marks, etc, and STR-CHAR, which are strange characters that appear because of an incorrect codification of the original text.

4 MTness Types in Human and Machine Translations

As seen in Table 1, more than a half of the MTness types were found in HT (these types appear in shaded cells), but the MTness instances only amounted to 50 out of about 1300 instances in total. Besides, the agreement between the informants that evaluated the HT was sparse. The agreement in identifying MTness types with more than 1 example was not over 50%. Likewise, we were unable to work out why 44 examples of MTness in HT and 223 examples in MT were underlined; this is why they were not counted in any type.

Table 1. Frequency and average agreement of MTness types in both MT and HT

TYPES	IN HT		IN MT	
	#	% Agreement	#	% Agreement
W-AGR			105	66.94
I-POS			14	58.75
I-VERBF			52	32.32
I-ORD			60	83.19
STR-CHAR			46	93.33
SINT-GAP	3	0	137	39.18
E-TYPO	5	40	42	16.36
NO-L2	21	28.57	183	70.72
OVER-WRD	1	0	41	63.63
WRD-REP	1	0	8	64.58
NOI-SEG	1	0	88	36.96
SEM-GAP	1	100	15	72.91
SEM-INCOH	6	50	60	52.06
CON-INCOH	11	45.45	154	48.16

This indicates the degree of subjectivity in appreciating human likeness and machine likeness. This subjectivity depends on factors such as domain and world knowledge (16i)[1]where the informant (ii16)as in ,stylistic and grammatical exigency , or the personal distaste about finding ,expected a definite article before the noun where the informant (iii16) as in ,words that remind one of the source language which also ,considered 'convidamos' a verb too similar to the Catalan verb 'convidar' .means 'invite'

(16) i) **Anonymity Proxy** para Windows es de dominio público (NO-L2)
ii) Anuncio sobre **pérdida** de un título académico (SYNT-GAP)
iii) Os **convidamos** a descubrir la luz y el color (CON-INCOH)

Nonetheless, it is unusual to find a case of NOI-SEG in a human translation. We attribute the only case found to the inability of the informant to grasp the meaning from the context.

The common MTness types are more frequent and more widely appreciated in MT, being SEM-GAP and NO-L2 the ones that are agreed upon most. Whereas, E-TYPO is very subjective. Agreement about NOI-SEG, SYNT-GAP and CON-INCOH are below 50%, which is noteworthy at first glance. The percentage of NOI-SEG can be

[1] Similar cases to (16i) include the translation of 'Windows OS' as 'Ventanas OS' or 'Los Tiempos' for the newspaper 'The Times'.

explained in terms of the way the experiment was developed. Some of the informants focused on underlining some segments while others, when facing a completely incomprehensible sentence, underlined larger segments that included them, or they even underlined the whole sentence. In SYNT-GAP cases, the disagreement is often due to the unconscious mental process of 'filling the gaps' when reading a sentence with missing articles, prepositions and other discrete syntactic constituents. This is not the case with constituents that are more prominent when grasping the meaning of the sentence, such as negative particles. There are other reasons as well. If the missing syntactic constituent is never present in the speaker's native language, the probability of disagreement arises. For instance, some informants did not appreciate the missing preposition 'a' before an animate direct object because in Catalan no direct object is preceded by a preposition. Stylistic tastes about using or not using a constituent also explains disagreements in SYNT-GAP, as we explained in (16ii).

The disagreement due to domain and world knowledge, which explains NO-L2 instances in human translations, also applies in cases of CON-INCOH and even SEM-INCOH. For instance, we saw *fichero adjunto* ('attached file') underlined. SEM-INCOH examples are mainly found in domain terminology, as the lack of semantic compositionality in some terms strikes people who are not familiar with the domain.

As for strict MT types, we see that STR-CHAR and I-ORD are the MTness types with the most agreement. Actually the agreement of strict MT types is considerable, except for I-VERBF, mainly because verbal tense and mood inconsistencies seem to be appreciated depending on the taste of the informants or their priority in underlining other examples they think are more crucial (finite/non-finite verbal forms, as in (10))

Except W-AGR which is the fourth most frequent type, the strict MT types with the most agreement are not the most frequent types. The types at the top of the frequency ranking also appear in HT (NO-L2, CON-INCOH, SINT-GAP). As some examples of these types were appreciated by subjective factors (stylistic taste, domain and world knowledge, or knowledge of the source language), we decided to analyse all the MTness examples with agreement and find what distinguishes them from the ones with no agreement at all. We realised that the examples with the most agreement, on the one hand, violate the model of the target language at a syntactic and semantic level. Those are examples with POS combinations which are not possible in the target language, as in a SINT-GAP case like (5), or semantic combinations which are not to be found in the semantic model of this language, as in (12). Likewise, even if a constituent is semantically feasible, when its word combination reminds people of a more probable one in the language model, the constituent is likely to be considered an example of MTness. This is the case of NPs with no apocopated determiners, as in (15).

5 and Future Work

In this paper we have presented a typology of MTness according to a comprehensive empirical study where human and machine translations underwent the Turing test. In the evaluation process, the detection of examples of this typology should be

implemented. We have also seen how MT and HT share some MTness types, according to some informants, and hence have proved how subjective the appreciation of human and machine likeness can be. As our goal is to implement this typology in automatic MT evaluations, in order to measure machine likeness, we could have suggested detecting types that were agreed on most often as attributable to MT only. However, we would have neglected the fact that word forms, syntactic structures and semantic patterns that violate the target language model are generally detected, no matter whether they are MTness instances whose types were also found in HT.

So we are planning to implement the detection of examples pertaining to all the types of the typology. However, we intend to establish weights for the examples detected in order to ponderate the MTness of the translation and present a score. Examples that violate the syntactic or semantic pattern of the target language model would have more weight than those that do not. In cases where there is no violation of the language model, a strict MT type would add more weight than a MT/HT type.

References

1. Papineni, K., Roukos, S., Ward, T. and Zhu, W-J.: BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of ACL, Philadelphia, PA. (2001)
2. Kulesza, A. and Shieber, S. M.: A Learning Approach to Improving Sentence-Level MT Evaluation. Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation, Baltimore. (2004)
3. Amigó, E., Gimenez, J., Gonzalo, J. and Márquez, L.: MT-Evaluation: Human-like vs. human acceptable. Proceedings of ACL. (2006)
4. Jelinek, R.: Modern MT Systems and the Myth of Human Translation: Real World Status Quo. Proceedings of the International Conference on Translating and the Computer 26, 18-19 November 2004, London. London: Aslib. (2004)
5. Reeder, F.: In One Hundred Words or Less. MT Evaluation Workshop MT Summit VIII. Santiago de Compostela. (2001).
6. Climent, S., and Moré, J.: A Cheap Evaluation Method Based on the Notion of Machine-Translationness. Proceedings of the Metis-II Workshop “New Approaches to Machine Translation.” 83- 90. (2007)