Machine Translationness: Machine-likeness in Machine Translation Evaluation

Joaquim Moré, Salvador Climent
Universitat Oberta de Catalunya
Avda Tibidabo, 39
jmore@uoc.edu, scliment@uoc.edu

Abstract

Machine translationness (MTness) is the linguistic phenomena that make machine translations distinguishable from human translations. This paper intends to present MTness as a research object and suggests an MT evaluation method based on determining whether the translation is machine-like instead of determining its human-likeness as in evaluation current approaches. The method rates the MTness of a translation with a metric, the MTS (Machine Translationness Score). The MTS calculation is in accordance with the results of an experimental study on machine translation perception by common people. MTS proved to correlate well with human ratings on translation quality. Besides, our approach allows the performance of cheap evaluations since expensive resources (e.g. reference translations, training corpora) are not needed. The paper points out the challenge of dealing with MTness as an everyday phenomenon caused by the massive use of MT.

Keywords: machine translationness, machine translation evaluation, MTS

1. Introduction

Machine translationness (MTness) is the linguistic phenomena that make machine translations distinguishable from human translations. It is the flavour of machine translation. The objectives of this paper are to characterize MTness from a linguistic point of view and to suggest a method of determining how machine-like a translation is. The machine-likeness degree is indicated by the MTS metric which proved suitable for machine translation evaluations (MTE). In section 2 we overview the human and machine-likeness treatment in MTE. In section 3 we explain how MTness is scored. The evaluation of the method is explained in section 4, followed by the conclusions and future work.

2. Human and machine-likeness in MTE

A machine translation system is an example of a machine that emulates intelligent human behaviour. Therefore, the interest in evaluating machine translation (MT) output has lain in determining whether this output is indistinguishable from a human translator output. In human evaluations, the items rated- fluency and adequacy- are qualities of human translations, and in automatic evaluations, metrics such as BLEU or NIST rate translation quality by taking human translations as references. This implies the high cost of creating human references that capture the legitimate translation variations of the same source.

(Amigó et al., 2006) claim that improving metrics according to human likeness is more reliable than improving metrics based on the correlation with human judgements. Current metrics do not describe human likeness at the sentence level individually, so they suggest to combine multiple metrics that capture partial aspects of human likeness into a single measure of quality.

An important trend in MTE, which takes machine-likeness into account, is based on this generalization: human-like MT translations are good whereas machine-like MT translations are bad. This generalization sustains the classification approach in MTE. The evaluation becomes a classification task: if the translation is classified as machine-like the translation is bad, if it is classified as human-like then it is good (Kulesza and Shieber, 2004) and (Corston-Oliver et al., 2001). The classification task is performed via machine learning but the machine learning approach does not characterize machine translationness because the training corpora are human and machine translations of the same source. Training corpora are also costly to create, except for an institution or a company with an overwhelming production of human and machine translations. Even in this case, the classifier learns from domain-specific corpora. If the sources were from other domains the results would probably be different.

Evaluating machine translations according to qualities of machines is an approach that has not been developed so far. Our proposal suggests assessing machine translations according to what they are (translations produced by a machine) and not to what they resemble (human translations). This motivates the linguistic study of the MTness features, that is, the linguistic phenomena that characterize machine-made translations. This study proposes a MTness typology, which is quite different from traditional MT errors typologies. The MTness typology is the characterization of linguistic features that can be applied for evaluating texts, whereas MT error typologies are focused on post-editing and improving MT systems (Farrús et al., 2011).

3. Scoring machine translationness

Scoring the machine translationness of a text requires the previous automatic detection of MTness instances. MTness detection implies to know what to detect and how to detect it by using NLP tools and resources. The *what* is a linguistic phenomenon that belongs to an MTness typology. The *how* consists in finding the items that force the NLP tool to build a linguistic representation of the text which is not recognized by the intuitive knowledge of the language. Finally, the scoring takes into account two conclusions we took from the studies on machine translationness perception by common people (Moré and Climent, 2008). First, the fact that one single word can spoil the translation and second the accumulative effect of MTness instances across different linguistic levels.

3.1. MTness typology

The MTness typology in Moré and Climent (2008) is summarized in table 1. Apart from typifying the MTness linguistic features, the typology is also the source of verification for MTness detection.

3.2. MTness detection

MTness detection can be stated as: detect the use of words and dependencies between words that do not match the native speaker′s intuitive knowledge of the language. The challenge is to detect as many mismatches as possible by using state-of-the-art natural language processing resources.

3.2.1. Word use and dependency representations

The knowledge about the use of words is represented in dictionaries, and the dependencies between words are represented by parsers. Parsers model the processing of a sentence by native speakers according to their intuitive knowledge. Although state-of-the-art parsers do not always behave perfectly, we assume that parsing representations are as consistent with intuitive knowledge as possible, bearing in mind their limitations. So an MTness instance is regarded as the linguistic item that forces the parser to build a representation that is not recognised by the intuitive knowledge of the language.

According to (Melĉuk, 1988) there are three types of dependencies: syntactic, morphological and semantic. Dependency parsers represent these dependencies in a labelled tree, called *typed dependency tree*. Figure 1 and figure 2 show an example of typed dependency tree by the Txala parser, which is the parser we used for the experimental assessment of our evaluation method (see section 4)[1].

Each line describes the syntactic, morphological and lexico-semantic information of a word in a tree node. The syntactic function is introduced first, and the form, lemma, grammatical category values and the Wordnet synset(s) of the word are between parentheses. Category values correspond to features such as part of speech, number, person, gender, etc, that are displayed morphologically. The line indentation corresponds to the arc from the dependent to its governor, and the indentation length of a line that describes a dependent node is two spaces longer than the indentation of the governor. So, in figure 2, the dependent nodes of the root are described in lines 2, 5 and 8. Notice that the dependent nodes of the governor are grouped between brackets and they share the same indentation length. These nodes are respectively the governors in smaller typed dependency trees whose dependent nodes are grouped between brackets as well. We will call these smaller typed dependency trees *dependent subtrees*.

The grouped nodes are ordered according to the X-bar theory (Chomsky, 1970). The node for the head, which is generally the governor, appears first. Then it follows the node for the word which functions as the specifier (e.g. subject of the verb, noun determiners), if any, and then the complements and finally the adjuncts.

Lexical	
MTness Type	**MTness Type Description**
NO-L2	Words which are not recognised as pertaining to the target language and are not loan words
Syntactic	
MTness Type	**MTness Type Description**
I-AGR	Morphological values that do not comply to the grammatical agreement restrictions between syntactic constituents (verb-subject agreement, noun-adjective, etc.)
I-POS	The part of speech (PoS) of a word is inadequate according to the context in which it appears.
I-VERBF	Non-finite verbs that should have appeared in finite forms and vice versa. Inconsistencies in the verbal mood (indicative and subjunctive)
I-ORD	Wrong order of syntactic constituents (e.g determiner and noun, verb and clitics, prepositional phrases displaced, etc.)
OVER-WRD	A word, or a sequence of words, not performing any syntactic role in the sentence. By deleting them, the sentence is syntactically well formed
WRD-REP	Two identical word-forms in the same syntactic phrase or in two phrases which are close to each other
SYNT-GAP	A missing constituent in a subcategorization structure.
Semantic	
MTness Type	**MTness Instance Description**
SEM-GAP	Missing constituents in an argument structure that are necessary to understand the sentence.
SEM-INCOH	Absurd interpretations because arguments do not fit the semantic restrictions of the predicate.
CON-INCOH	Arguments that do not violate semantic restrictions of the predicate but do not fit the context where they appear.
Formatting	
MTness Type	**MTness Instance Description**
STR-CHAR	Strange character
I-TYPO	Inadequate use of upper case and lower case, missing or inadequate punctuation marks, etc

Table 1: Summary of the MTness typology

[1] The Txala parser is a tool of the open source Freeling library

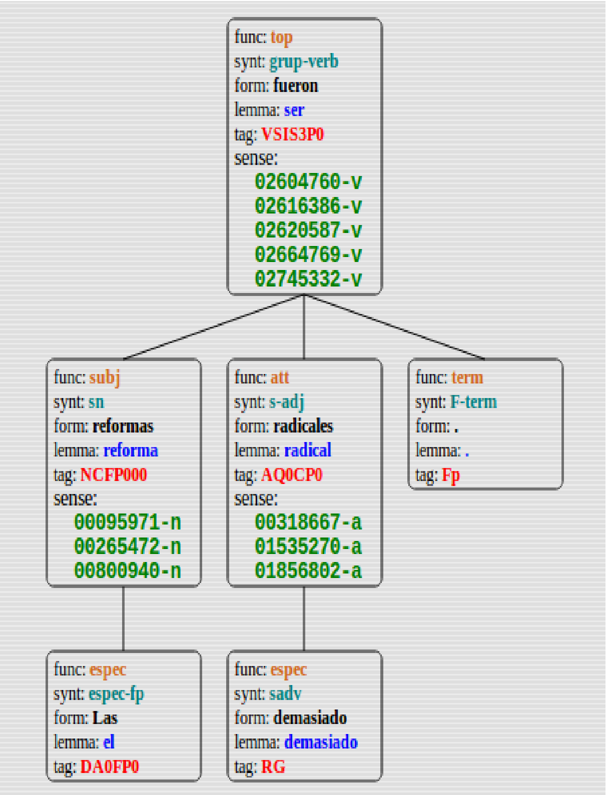
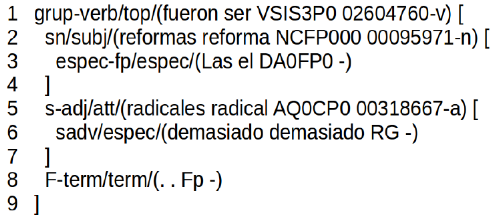


Figure 1: Example of a tree representation of dependencies



```
grup-verb/top/(fueron ser VSIS3P0 02604760-v) [
  sn/subj/(reformas reforma NCFP000 00095971-n) [
    espec-fp/espec/(Las el DA0FP0 -)
  ]
  s-adj/att/(radicales radical AQ0CP0 00318667-a) [
    sadv/espec/(demasiado demasiado RG -)
  ]
  F-term/term/(. . Fp -)
]
```

Figure 2: Example of output of the dependency parser

3.2.2. Detection of MTness instances in the use of words

MTness instances in the use of words are NO-L2 and STRCHAR instances; that is, lexical items that are not recognised as language words by the intuitive knowledge of native speakers. The typed dependency tree is traversed and each lemma is verified whether it matches a lemma in a monolingual dictionary of the target language. If not, the lemma is looked up in a source-target bilingual dictionary. The bilingual dictionary consists of a list of pairs where the first element of the pair is the lemma of a source word and the second element is a list of translation equivalents. So a pair is searched for in the bilingual dictionary where the lemma of the word matches both the source and an equivalent. If no pair is found, then the word form is an MTness instance. The use of the bilingual dictionary prevents loan words such as *golf e-Book* or *hip-hop* from being considered NO-L2 instances, whereas untranslated words are detected. (see table 2).

Word	MTness?	Explanation
mónadas	NO	The word form matches a word form in the monolingual dictionary
Nassau	NO	The word form matches a word form in the gazetteer
ostinato	NO	Bilingual pair: ostinato - ostinato
allegiance	YES	Bilingual pair: allegiance - fidelidad, lealtad
Caribbean	YES	Bilingual pair: Caribbean - Caribe

Table 2: Examples of MTness instances in the use of words

3.2.3. Detection of MTness instances in the relations between words

We focused on relations between words when the governor is a noun or a verb because the three types of dependencies are more clearly appreciated. The method consists in checking whether there are linguistic items that force the parser to build a dependency tree representation (DTR) which is not consistent with linguistic intuition. The DTR can represent the root tree or a dependent subtree.

We distinguish three types of DTRs, each corresponding to a type of dependency.

1. Syntactic DTR
2. Morphological DTR
3. Semantic DTR

A DTR from a machine translation is called *hypothesis DTR* because its consistency with linguistic knowledge must be assessed. The consistency is assessed by one of the following actions:

1. Matching hypothesis DTRs with reference DTRs
2. Testing the real use of co-occurrent words

a) Hypothesis DTRs matching with reference DTRs The linguistic consistency of a hypothesis DTR is assessed when the hypothesis DTR matches a *reference DTR*; that is, a DTR that represents a dependency structure which is recognized by the intuitive language knowledge. So a hypothesis syntactic DTR has no MTness instances if the DTR matches a reference syntactic DTR. A hypothesis morphological DTR is right if it matches a reference morphological DTR and a hypothesis semantic DTR is not odd if it matches a semantic DTR. On the contrary, MTness instances are to be found in non-matched hypothesis DTRs.

a.1) Creation of reference DTRs In order to obtain reference DTRs, the dependency parser parses corpora with texts that are representative of the native speakers′ use of the target language. The resulting typed dependency trees display all the lexical, syntactic, morphological and semantic information of the sentences in the representative texts. For each root tree and dependent subtree, the DTR creation puts linguistic annotations from the typed dependency tree onto a tuple with at most three elements. The central element is for the linguistic annotation of the head. The other two are for the annotations of the specifier or complement, if any.

According to (Lloberes et al., 2010), the evaluation of the parser we used showed that around 80% of the dependency trees had a correct head and a correct head and dependency relations. This means that bad reference DTRs may be present. Yet, after counting the number of times each DTR describes the representative sentences, a frequency threshold of reference DTRs was established for the detector to consider a reference DTR reliable.

Although the percentage of possible ill reference DTRs is not significant enough to dismiss the use of a dependency parser, we noticed that detecting MTness instances by matching DTRs was more reliable for detecting syntactic and morphological MTness instances than for finding semantic MTness instances. Automatic semantic labelling depends on procedures such as word-sense-disambiguation whose results are still far to be reliable in any language. So we leave open for the future the possibility of detecting MTness instances by matching semantic DTRs.

a.1.1) Creation of reference syntactic DTRs There are two types of syntactic DTRs. The first one is the phrase DTR (DTR_p), which is the DTR that describes the dependency relations in terms of syntactic phrases and syntactic functions. The first position holds the phrase and function of the specifier, the central position holds the type and function of the phrase that dominates the head, and the third position holds the type and syntactic functions of the complements. The second type of syntactic DTR is the subcategorization DTR (DTR_s). The subcategorization DTR describes subcategorization relations when the governor has a specific lemma.

a.1.2) Creation of reference morphological DTRs Reference morphological DTRs are tuples with two positions. When the head has a specifier, the first position of the tuple is for the category values of the specifier, and the second position is for the category values of the head. Then, for each complement, a DTR is created whose first position is for the category values of the head, and the second position is for the category values of the complement. The reason is the assumption that the head restrictions are not applied at the phrase level but from head to dependent individually.

a.2) Creation of hypothesis DTRs Hypothesis DTRs are created from the typed dependency tree of a machine translation. Syntactic and morphological hypothesis DTRs are created the same way as reference DTRs. Figure 3 shows the hypothesis DTR of the sentence *Aquel restaurante sirve platos excelentes* (That restaurant serves excellent dishes).

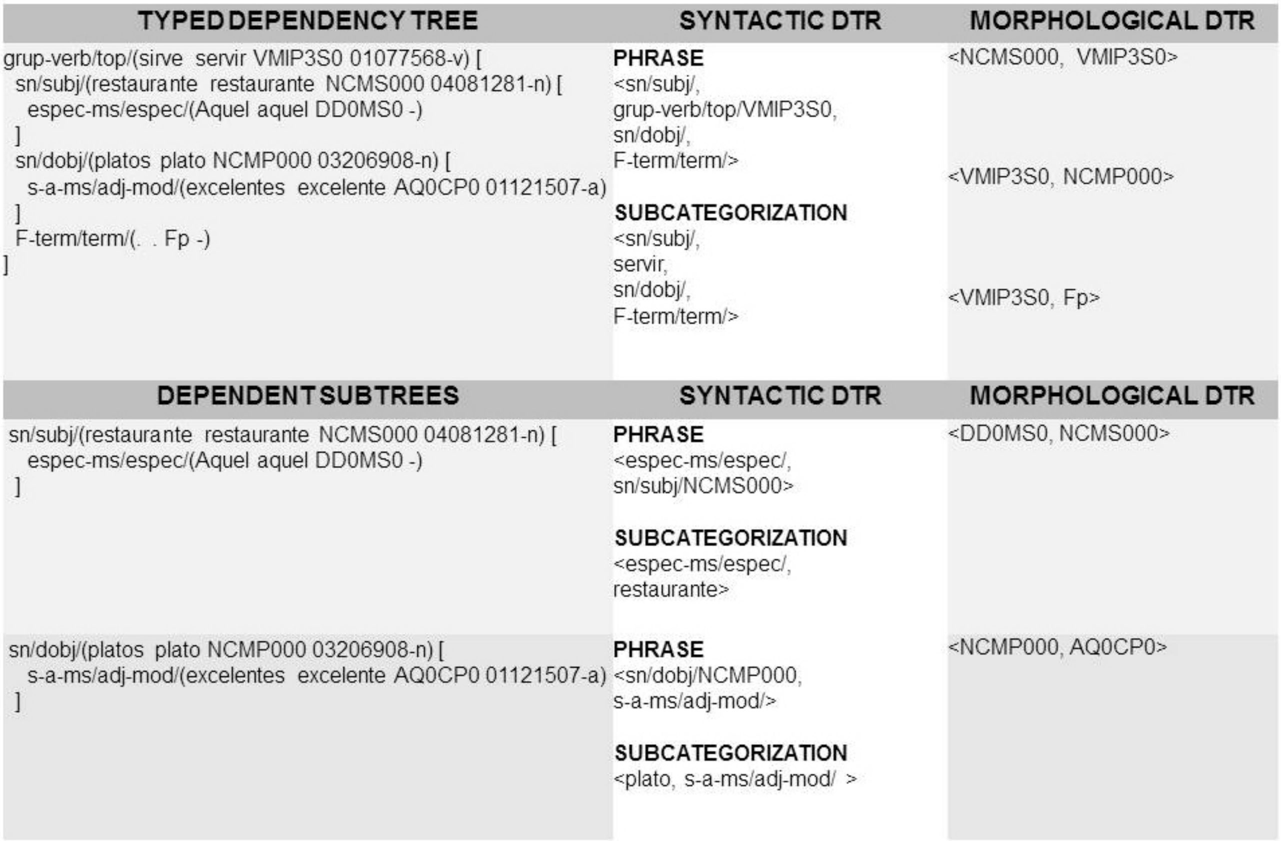
a.3) Detection of MTness instances in syntactic DTRs From the experiment on MTness perception by human readers, we noticed that there were segments (*noisy segments*) that caused noise in understanding the translation. These segments are generally represented by phrase DTRs that do not match any reference phrase DTR. For example, I-POS and SYNT-GAP instances where a noun phrase dominates an adjective and the noun is missing. This is the reason why the category values of the head appear in phrase DTRs.

We also noticed from the experiment that MTness instances are salient if they call up the right translation to the mind of the reader. These kind of instances can be detected in subcategorization DTRs, and the evocation of the right translation is modelled by obtaining the *expected syntactic DTR*. The expected syntactic DTR is the reference subcategorization DTR whose edit distance to the hypothesis is the shortest. The MTness instance is detected by analyzing the necessary operations on the hypothesis DTR to get the expected syntactic DTR. If a symbol must be inserted in order to get the expected syntactic DTR, and this symbol indicates a syntactic role (subject, direct object, indirect object or prepositional complement) then a complement with this syntactic role is expected. This is the case of SYNT-GAP and SEM-GAP instances. Finally, if a symbol must be deleted and this symbol indicates a syntactic role then a constituent with this role is not expected. This is the case of OVER-WRD, with an over generation of specifiers or complements.

a.4) Detection of MTness instances in morphological DTRs We assume that these MTness instances are salient if they call up the right grammatical category values to the reader′s mind. The evocation of the right values is modeled by obtaining the expected morphological DTR, the same way as the expected syntactic DTR. If the expected DTR is the result of replacing the value of part of speech, number, gender, or verb form/mood with a different value, then the DTR represents an MTness instance. This method detects instances of I-POS (inconsistent part of speech), I-AGR (inconsistent agreement), or I-VERBF (inadequate verbal form).

b) Testing the real use of co-occurrent words MTness instances can also be detected with the following assumption: in a dependency tree, if the governor does not occur with the specifier or one of the complements in the largest representative corpus available- the Web- then the dependency tree represents an MTness instance. A search engine is requested to find contexts where the specifier or a complement co-occurs with the governor.

The requests are performed with queries interpretable by the search engine, and the key words are the word forms of the head and the word form of the specifier or the complement, because the search engine matches word forms. The order of the key words indicates the order in which these words should appear in the matching contexts. The order specifier-head-complement is suitable for retrieving contexts in languages such as English, Spanish or Catalan. For other languages with a different argumental order, the queries should be adapted.



TYPED DEPENDENCY TREE	SYNTACTIC DTR	MORPHOLOGICAL DTR
grup-verb/top/(sirve servir VMIP3S0 01077568-v) [sn/subj/(restaurante restaurante NCMS000 04081281-n) [espec-ms/espec/(Aquel aquel DD0MS0 -)] sn/dobj/(platos plato NCMP000 03206908-n) [s-a-ms/adj-mod/(excelentes excelente AQ0CP0 01121507-a)] F-term/term/(. . Fp -)]	**PHRASE** <sn/subj/, grup-verb/top/VMIP3S0, sn/dobj/, F-term/term/> **SUBCATEGORIZATION** <sn/subj/, servir, sn/dobj/, F-term/term/>	<NCMS000, VMIP3S0> <VMIP3S0, NCMP000> <VMIP3S0, Fp>
DEPENDENT SUBTREES	**SYNTACTIC DTR**	**MORPHOLOGICAL DTR**
sn/subj/(restaurante restaurante NCMS000 04081281-n) [espec-ms/espec/(Aquel aquel DD0MS0 -)]	**PHRASE** <espec-ms/espec/, sn/subj/NCMS000> **SUBCATEGORIZATION** <espec-ms/espec/, restaurante>	<DD0MS0, NCMS000>
sn/dobj/(platos plato NCMP000 03206908-n) [s-a-ms/adj-mod/(excelentes excelente AQ0CP0 01121507-a)]	**PHRASE** <sn/dobj/NCMP000, s-a-ms/adj-mod/> **SUBCATEGORIZATION** <plato, s-a-ms/adj-mod/ >	<NCMP000, AQ0CP0>

Figure 3: Hypothesis DTRs of the translation *Aquel restaurante sirve platos excelentes*

The MTness instance is detected if the search engine retrieves no results or the number of results is below 3. We establish a threshold of 3 because it is possible that the co-occurring words match contexts of non revised machine translated documents. We assume that the number of these misleading contexts is generally below 3. If the number of results is over 3, then the number of matching snippets are counted. Matching snippets are those where the words co-occur with no punctuation marks that initiate another clause in between (e.g. period, semicolon, parenthesis, etc.). If the number of matching snippets is below 3 then the dependency tree represents an instance of MTness.

3.3. The MTS metric

MTS (MTness Score) is a metric that rates the machine translationness of a piece of text (translation unit). MTS values range from 0 to 1. 0 means that no traces of MTness were detected and 1 signifies that the piece of text was unquestionably produced by a machine. Values between 0 and 1 indicate how close the translation unit is to a piece of text where all the words are affected by machine translationness. The score must be consistent with the fact that one single word can spoil the translation and, on the other hand, the score must capture the effect caused by the overlapping of MTness instances at different linguistic levels. An MTness instance may impact across different linguistic levels and the more levels impacted the stronger is the MTness appreciation. For example, a noun phrase which is wrongly assigned the subject role becomes an I-AGR instance when its morphological agreement values and those of the verb do not match.

In the experiments about MTness perception, lexemes not used in the target language caused the most impact, and the agreement between informers was high (over 90% for STR-CHAR and 70% for NO-L2). Therefore translations with NO-L2 and STR-CHAR instances are rated the highest. The overlapping of MTness instances at different linguistic levels are modelled by combining the values of the following three metrics:

- Syntactic mts: Metric that rates MTness when only ill syntactic dependencies are detected
- Morphological mts: Metric that rates MTness when only morphologically inconsistent dependencies are found.
- Co-occurrent mts: Metric that rates MTness when only inconsistent co-occurrent words are found.

3.3.1. The mts calculation

The reasoning behind the mts calculation is to rate how close the translation is to the worst of the situations, that is a translation with the highest MTness at a linguistic level. The highest syntactic mts indicates that all the nodes of the typed dependency tree appear in syntactic DTRs with MTness instances. The highest morphological mts indicates that all the nodes of the typed dependency tree appear in morphological DTRs with MTness instances. Finally, the highest co-occurrent mts indicates that all the dependent words of the typed dependency tree cannot co-occur with their governors.

a) reference mts string The *reference mts string* models the worst of the situations. The reference mts string is created by first indexing all the lines of the parser′s output, which represent the nodes of the dependency tree. The index is the line number and the ordered indexes are concatenated in a string. This is the *node string* (NS). The reference is obtained by replacing the indexes of the NS by the symbol 'M' ('M' stands for 'machine translationness').

b) hypothesis mts string The hypothesis mts string is generated with the *Dependency Index Tuples* (DiT). For the root tree and for each dependent subtree a DiT is created with the indexes of the governor, the indexes of the dependents and the index of the end of the dependency tree. When the root tree or dependent subtree represents an MTness instance, the DiT indexes of the dependents are replaced by the symbol 'M' in the node string[2].

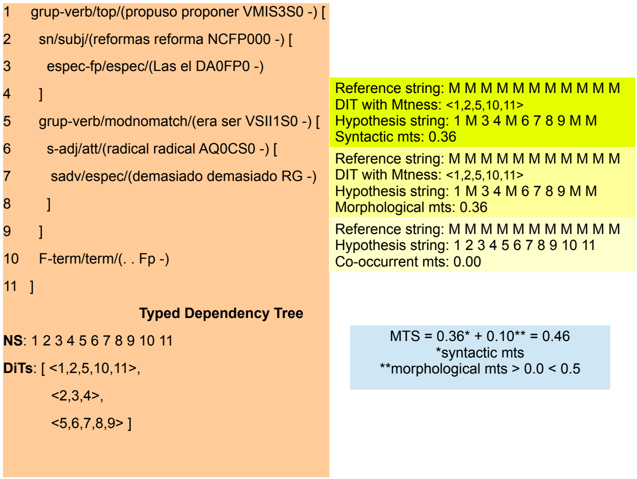
c) The syntactic, morphological and co-occurrent mts calculation The distance of the hypothesis to the worst situation at a linguistic level is rated by using a metric that takes the precision and recall of a hypothesis string with respect to a reference. We chose ROUGE-L (Chin-Yew and Och, 2004) because this metric takes into account the consecutive positions of MTness symbols in the hypothesis. We assume that the more consecutive the positions are, the more impact in the perception of MTness.
To calculate the syntactic mts the hypothesis string is created by retrieving the DiTs of the root tree and dependent subtrees whose syntactic DTRs describe MTness instances. Then the indexes of the dependents affected are replaced by the symbol 'M' in the node string. Once the reference and the hypothesis are created, the ROUGE-L is calculated. The morphological and co-occurrent mts are calculated the same way. For the morphological mts, the indexes are those of the dependents in morphological DTRs that describe MTness instances, and for co-occurrent mts, the indexes are those of the dependents which cannot co-occur with their governor.

3.3.2. The MTS calculation

When the MTness instance is NO-L2 or STR-CHAR the MTS value is 1. Otherwise, the value is calculated according to the cumulative effect of MTness instances across different linguistic levels, which is modeled this way:[3]

- Score the partial MTS with the highest mts from the syntactic, morphological and co-occurrent mts
- For each remaining mts over 0.5, increase the partial MTS by two tenths
- For each remaining mts over 0 and below 0.5, increase the partial MTS by one tenth
- Equal the definitive MTS to the current partial MTS value

As an example, let us see the MTS calculation of the translation *las reformas propuso era demasiado radical* (the reforms he proposed was too much) (Figure 4).
The expected syntactic DTR of the root tree (table 3) indicates that the direct object of the verb *propuso* (proposed) is missing. So the hypothesis DTR describes a SYNT-GAP instance. In the DiT of the root tree, the dependents have the indexes 2, 5, 10 and the limit index of the tree is 11. Therefore, these indexes in the node string are replaced by the symbol 'M' to generate the hypothesis for calculating the syntactic mts.



```
1   grup-verb/top/(propuso proponer VMIS3S0 -) [
2     sn/subj/(reformas reforma NCFP000 -) [
3       espec-fp/espec/(Las el DA0FP0 -)
4     ]
5     grup-verb/modnomatch/(era ser VSII1S0 -) [
6       s-adj/att/(radical radical AQ0CS0 -) [
7        sadv/espec/(demasiado demasiado RG -)
8       ]
9     ]
10    F-term/term/(. . Fp -)
11  ]
```

Typed Dependency Tree

NS: 1 2 3 4 5 6 7 8 9 10 11

DiTs: [<1,2,5,10,11>,
<2,3,4>,
<5,6,7,8,9>]

Reference string: M M M M M M M M M M M
DIT with Mtness: <1,2,5,10,11>
Hypothesis string: 1 M 3 4 M 6 7 8 9 M M
Syntactic mts: 0.36

Reference string: M M M M M M M M M M M
DIT with Mtness: <1,2,5,10,11>
Hypothesis string: 1 M 3 4 M 6 7 8 9 M M
Morphological mts: 0.36

Reference string: M M M M M M M M M M M
Hypothesis string: 1 2 3 4 5 6 7 8 9 10 11
Co-occurrent mts: 0.00

MTS = 0.36* + 0.10** = 0.46
*syntactic mts
**morphological mts > 0.0 < 0.5

Figure 4: MTS calculation of *las reformas propuso era demasiado radical*

Hypothesis DTR_s	Expected DTR_s
<sn/subj/,proponer/grup-verb/modnomatch/ >	<sn/subj/,proponer/grup-verb-inf/**dobj** >

Table 3: SYNT-GAP instance

The fact that *las reformas* is wrongly typed as the subject of the main verb *propuso* has consequences at the morphological level. The expected morphological DTR of the root tree (table 4) indicates that the expected number feature of the subject should agree with the verb. So the hypothesis DTR describes an I-AGR instance.

Hypothesis DTR	Expected DTR
<NCF**P**000, VMIS3S0 >	<NCM**S**000, VMIS3S0 >

Table 4: I-AGR instance

So the indexes 2, 5, 10 and the limit index of the tree in the node string are replaced with the symbol 'M' to generate the hypothesis for calculating the morphological mts.
The highest mts value corresponds to the value of the syntactic mts and the morphological mts. If we take the syntactic mts value as the partial MTS (0.36), the definite MTS

[2] The DiT index of the governor is not replaced by the symbol 'M' because the governor is not an MTness instance by itself but the dependents that are wrongly related to the governor

[3] Sometimes the mts in one type is so high that when added to other mts, the sum is over 1. In that case, the value is normalized to 1.

value is obtained by adding one tenth because the morphological value is over 0 and below 0.5. So the MTS of the example is 0.46

When the translation has more than one sentence, the MTS score is the result of merging the MTS of each sentence. The calculation is similar to the calculation in single translations but the factors are MTS scores instead of mts. First we score the partial MTS, which is the highest MTS, and then for each remaining MTS over 0.5, the MTS is raised by two tenths. On the other hand, for each remaining MTS over 0 and below 0.5 the MTS is raised by one tenth.

4. Evaluation of the method

We evaluated the method by analysing the Pearson correlation between MTS scores and human quality perception. The MTS variable should decrease as the human rating increases (the higher quality the lower MTS). Therefore, the coefficient should be a negative fraction, away from 0 and tending to -1. Three people were asked to assess 196 Wordnet glosses machine translated from English into Spanish.

We were interested in the linguistic intuition of monolingual ordinary readers in detecting disfluent and inaccurate translations. So we did not need bilingual evaluators to judge disfluent translations and judge the inaccuracy of odd and absurd translations. We realized that a standard ARPA scale was suitable for our experiment. This scale has five points: 1-Incomprehensible, 2- Disfluent, 3- Non-native, 4- Good, 5- Flawless Spanish (the language of our experiment). Although the scale is for fluency we considered that translations with MTness instances that affected accuracy could be rated as incomprehensible.

The correlation to human rating was -0.71 (p-value <¡ 2.2e-16). The correlation value for MTS was higher than the correlation obtained with ngram-based metrics (BLEU, NIST, METEOR, ROUGE-L and GTM) and reference-based metrics that also process a dependency tree representation (HWCM). The MTS correlation was also higher than the correlation of classification-based methods (Gamon et al., 2005) and (Mutton et al., 2007).

The MTS correlation result was good even by using freely available resources. The parser is open source, licenced under GPL, the dictionaries were generated from Wiktionary resources, the API to the search engine was free at the time of the method assessment[4], and the reference DTRs were obtained by parsing freely available documents.

Our method is cheaper because translation references and training corpora are not needed and the evaluation is performed with freely available resources. It is true that our method depends on the availability of parsers and corpora in the target language, and not all the languages have this kind of resources yet. So the challenge is a large-scale application of the method for as many languages as possible.

One might wonder how can we guarantee that the reference corpus, from which reference DTRs are created, covers all the hypothesis space. Actually, the same question is relevant for n-gram metrics, because we also may wonder if translation references cover all the legitimate translation variations. On the other hand, how can we prevent the risk that something that is not in the reference corpus, but is correct in the target language, will be considered as an error if the reference corpus is inadequate? The same question is relevant for translation references. If the references do not cover all the hypothesis space, a good translation can be regarded as bad. All in all, the methodological question of the hypothesis space coverage must be taken into account.

5. Conclusions and future work

In this paper we presented an evaluation method that assesses machine translations according to what they are (translations produced by a machine) and not to what they resemble (human translations). The method is centered on the machine qualities of machine translations rather than human translation qualities, as in state-of-the-art methods. This method proved to correlate better with translation quality judgements, and the costs are significantly lower.

As future work, we intend to detect I-ORD (inadequate order) and WRD-REP (the same two words are very close together). The parser output is hierarchically- not linearly-ordered. Therefore instances of I-ORD and WRD-REP are not detected by retrieving an expected syntactic DTR. So we leave open the detection of these instances with a parser that displays word position information. Anyway, I-ORD and WRD-REP are often consequences of dependency structures that have already been assessed as having an MTness instance.

Machine translationness ratings can be applied for other uses beyond machine translation evaluation. The MTS metric can be an important indicator to prevent the consequences of the massive use of MT (plagiarism and other forms of cheating, detection of unsupervised MT documents published on the Web, etc.). The aplicability of the MTness typology across different languages is also an interesting field of research.

6. Acknowledgements

This article has been presented thanks to the project Scenario Knowledge Acquisition by Textual Reading (SKATeR) TIN2012-38584-C06] (2013-2016) from the Spanish Ministry of Economy and Competitivity.

7. References

Amigó, E., Giménez, J., Gonzalo, J., and Márquez, L. (2006). MT evaluation: Human-like vs. human acceptable. In *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 17–24.

Chin-Yew, L. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004).*

Chomsky, N., (1970). *Remarks on nominalization.* Waltham: Ginn.

Corston-Oliver, S., Gamon, M., and Brockett, C. (2001). A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of the 39th Annual*

[4] API to the Bing search engine

Meeting of the Association for Computational Linguistics (ACL), pages 140–147.

Farrús, M., Costa-Jussà, M. R., Marino, J. B., Poch, M., Hernández, A., Henríquez, C., and Fonollosa, J. A. (2011). Overcoming statistical machine translation limitations: error analysis and proposed solutions for the catalan-spanish language pair. *Language resources and evaluation*, 45:181–208.

Gamon, M., Aue, A., and Smets, M. (2005). Sentence-level MT evaluation without reference translations: beyond language modeling. In *Proceedings of EAMT*, pages 103–111.

Kulesza, A. and Shieber, M. (2004). A learning approach to improving sentence-level mt evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 75–84.

Lloberes, M., Castellón, I., and Padró, L. (2010). Spanish Freeling dependency grammar. In *Proceedings of the LREC-2010*, pages 693–699.

Melĉuk, I. (1988). *Dependency Syntax: Theory and Practice*. The SUNY Press, Albany, N.Y.

Moré, J. and Climent, S. (2008). A machine translationness typology for MT evaluations. In *Proceedings of the European Association for Machine Translation*, pages 130–139.

Mutton, A., Dras, M., Wan, S., and Dale, R. (2007). Gleu: Automatic evaluation of sentence-level fluency. In *Proceedings of ACL 2007*.