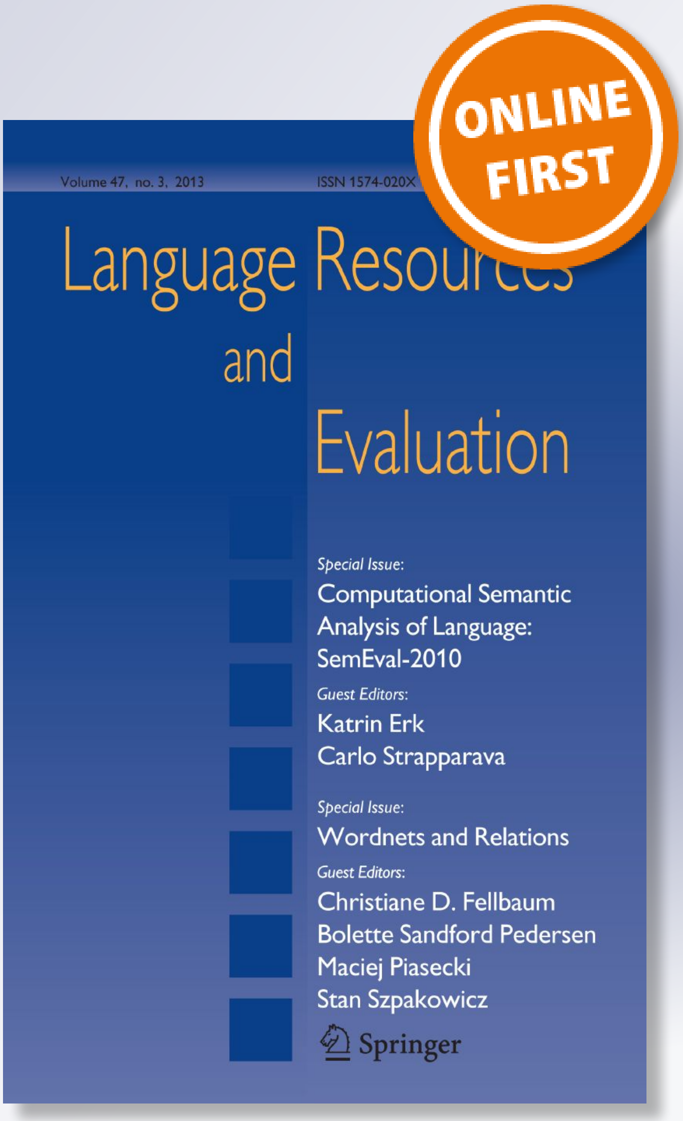
VERTa: a linguistic approach to automatic machine translation evaluation

Elisabet Comelles & Jordi Atserias

Language Resources and Evaluation

ISSN 1574-020X

Lang Resources & Evaluation
DOI 10.1007/s10579-018-9430-2



ONLINE FIRST

Volume 47, no. 3, 2013 ISSN 1574-020X

Language Resources and Evaluation

Special Issue:
Computational Semantic Analysis of Language: SemEval-2010

Guest Editors:
Katrin Erk
Carlo Strapparava

Special Issue:
Wordnets and Relations

Guest Editors:
Christiane D. Fellbaum
Bolette Sandford Pedersen
Maciej Piasecki
Stan Szpakowicz

Springer

Lang Resources & Evaluation
https://doi.org/10.1007/s10579-018-9430-2

ORIGINAL PAPER

VERTa: a linguistic approach to automatic machine translation evaluation

Elisabet Comelles[1] · **Jordi Atserias**[2]

Abstract Machine translation (MT) is directly linked to its evaluation in order to both compare different MT system outputs and analyse system errors so that they can be addressed and corrected. As a consequence, MT evaluation has become increasingly important and popular in the last decade, leading to the development of MT evaluation metrics aiming at automatically assessing MT output. Most of these metrics use reference translations in order to compare system output, and the most well-known and widely spread work at lexical level. In this study we describe and present a linguistically-motivated metric, VERTa, which aims at using and combining a wide variety of linguistic features at lexical, morphological, syntactic and semantic level. Before designing and developing VERTa a qualitative linguistic analysis of data was performed so as to identify the linguistic phenomena that an MT metric must consider (Comelles et al. 2017). In the present study we introduce VERTa's design and architecture and we report the experiments performed in order to develop the metric and to check the suitability and interaction of the linguistic information used. The experiments carried out go beyond traditional correlation scores and step towards a more qualitative approach based on linguistic analysis. Finally, in order to check the validity of the metric, an evaluation has been conducted comparing the metric's performance to that of other well-known state-of-the-art MT metrics.

Keywords Machine translation · Machine translation evaluation · MT metric · Linguistic features · Qualitative approach

✉ Elisabet Comelles
elicomelles@ub.edu

[1] Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain

[2] University of the Basque Country, Paseo Manuel de Lardizábal, 1, 20018 Donosti, Spain

Published online: 15 October 2018

1 Introduction

Machine translation (MT) is one of the most complete tasks within the field of natural language processing (NLP). MT is the automatic translation of one text from a source language into a target language; to put it simply, MT implies using a computer to translate text or speech from one language to another. This is one of the most challenging tasks inside the field of NLP because it implies most types of knowledge that humans possess (i.e. grammar, semantics, knowledge of the world, etc.). The complexity behind MT lies in automatically translating a text as a person does, with all the knowledge that it implies, creating a new text in the target language, with all the knowledge of the target language and target culture that goes with it.

Directly linked to MT there is a subtask, machine translation evaluation, which is intended to check (or evaluate) the quality of the automatic translation produced. As pointed out by Hutchins and Somers (1992), there are several types of evaluation which can be performed at different stages: (a) evaluation performed during the development of a system; (b) evaluation once the system has been developed before offering it to a potential user; (c) evaluation of the system by its potential buyers and users; and (d) evaluation of the system by the final recipients of translations.

In most of these stages there is one common point, the linguistic quality of the MT output. In order to evaluate this MT output one can focus on the assessment of the MT quality or on error analysis. Whereas the former deals with aspects such as assessing the accuracy or fidelity in translating the meaning of the source sentence or assessing if the target sentence can be understood, the latter focuses on identifying and classifying errors made by the MT system.

Both types of evaluations were initially performed by human evaluators. This has the advantage that MT developers are provided with a wide range of assessments regarding partial aspects of MT quality (ALPAC report 1966; White et al. 1994; Snover et al. 2006; Lo and Wu 2011; Macketanz et al. 2017). In addition, human evaluators possess all that knowledge that MT systems try to emulate. On the other hand, performing this type of evaluation is very expensive, time-consuming and subjective—sometimes the inter- and/or intra-annotator agreement is rather low (Turian et al. 2003; Ye et al. 2007; Callison-Burch et al. 2012). As a reaction to these drawbacks and since MT developers required fast and reliable MT evaluations, the MT community started developing and using automatic MT evaluation measures, the framework of this study.

Automatic MT evaluation metrics are supposed to be faster, cheaper and more objective than human evaluation. Actually, the use of this type of evaluation has been widely extended among MT developers because they can carry out fast evaluations of their MT systems and immediately use the results obtained to improve them. This is the main reason why in the last decade a wide range of MT metrics has been developed. Most of them work as similarity measures and use reference translations to compare them to the MT output or hypothesis. Among these there are BLEU (Papineni et al. 2001), NIST (Doddington 2002), METEOR (Banerjee and Lavie 2005), SMT and HWCM (Liu and Gildea 2005), TER (Snover et al. 2006), SR (Giménez 2008), MEANT (Lo and Wu 2012) or DiscoTK (Joty et al. 2014), just to name some of them. Although automatic MT metrics are widely used, they have also received criticism due to the fact that reference translations are required (Lommel 2016). As a response to this criticism other metrics are aimed at estimating MT Quality, in other words, predicting the quality of MT output when reference translations are not available, such as those metrics proposed by Specia et al. (2009, 2010, 2011).

From those metrics using similarity measures, some do not use linguistic information at all, such as BLEU and NIST among others; some use character n-grams, for example BEER (Stanojević and Sima'an 2015), the ChrF family of metrics (Popović 2015, 2017) or CharacTer (Wang et al. 2016); some use information at lexical level (e.g. synonyms, stemming, paraphrasing) such as METEOR, M-TER and M-BLEU (Agarwal and Lavie 2008), TERp (Snover et al. 2009), SPEDE (Wang and Manning 2012) or MPEDA (Zhang et al. 2016); some use morphological information (e.g. information about suffixes, roots, prefixes) such as AMBER (Chen et al. 2012) and INFER (Popović 2012); some use information regarding morphology and syntax (e.g. part-of-speech (PoS) tags, constituents, dependency relations) such as SMT and HWCM (Liu and Gildea 2005), Owczarzak et al. (2007a, b), SP, CP and DP metrics (Giménez 2008), DepRef (Wu et al. 2013) or UOWREVAL (Gupta et al. 2015); some make use of information related to semantics, such as SR and DR metrics (Giménez 2008), SAGAN-STS (Castillo and Estrella 2012), UMEANT (Lo and Wu 2013) and MEANT 2.0 (Lo 2017). Most of the above mentioned metrics evaluate partial aspects of MT output (e.g. vocabulary, syntax, semantics); however, in the last years MT metrics have been more oriented towards evaluating MT quality in general and MT researchers have struggled to find the best way to combine different types of MT metrics either by using machine learning techniques (Albrecht and Hwa 2007a, b; Yang et al. 2011; Gautam and Bhattacharyya 2014; Joty et al. 2014; Yu et al. 2015, Ma et al. 2017) or trying more simple approaches such as MAXSIM (Chang and Ng 2008), ULC (Giménez and Márquez 2010b), IPA and STOUT (González et al. 2014).

The above mentioned metrics range from very simple metrics, usually aiming at partial aspects of quality, to highly sophisticated ones, using a large amount of information and machine learning techniques. It must also be highlighted that the performance of these metrics depends on how well they correlate with human judgements and they are developed and improved taking into account these correlations.

Giménez and Márquez (2010b) reported that linguistic information and especially their combination of linguistic features correlated well with human judgements in several evaluation campaigns. However, little qualitative analysis on the use and influence of linguistic features, regardless of how well or badly they correlate with human judgements, has been performed. We consider that this qualitative analysis is also appropriate since, although correlation with human judgements is the standard method to evaluate the performance of a metric, it is highly dependent on the degree of intra-/inter-annotator agreement (Turian et al. 2003; Callison-Burch et al. 2012). Furthermore, when using more sophisticated metrics that combine linguistic information at different levels, such as those reported above, it is hard to interpret their score since this type of metrics uses such highly heterogeneous types of linguistic features that it is difficult to know to what extent and how each linguistic feature is contributing to the evaluation of MT output.

The present study introduces VERTa,[1] a linguistically-motivated MT metric. The development of VERTa is based on the analysis of linguistic features relevant to the evaluation of MT output. Thus, the experiments performed to develop VERTa pursue the aim of shedding some light on the suitability, influence and combination of linguistic information to evaluate adequacy and fluency, especially by highlighting the effectiveness and benefits of a more qualitative approach based on linguistic information. Finally, a comparison between VERTa's performance and that of other well-known metrics is also provided in order to check the validity of VERTa as an evaluation metric.

2 Metric architecture and description

VERTa is an MT metric that compares each hypothesis segment with the corresponding reference segment(s) according to different types of linguistic information.

When approaching the design and development of VERTa, a thorough linguistic analysis was conducted in order to identify the linguistic information to be considered when comparing hypothesis and reference segments. So as to conduct this analysis, part of the newswire datasets provided in the MetricsMatr evaluation task[2] was used. This data consisted of 100 segments (Arabic to English) of the NIST Open-MT06[3] data, the MT output from eight different MT systems and 4 reference translations. All segments were taken into account regardless of the system providing them, in order to have a more precise correlation and avoid being system-biased. The rest of data was kept unseen in order to evaluate the metric.

From the linguistic phenomena identified, the most relevant ones were selected and classified (Comelles et al. 2017). Even though most of them are interrelated and interact, they were classified into the following levels for the sake of analysis: lexical information (e.g. lexical semantics), morphological information (e.g. PoS, morphosyntactic features, etc.), syntactic information (e.g. word order and alternations) and semantic information (e.g. sentence semantics), following Farrús et al. (2010). This classification covering the different levels of language was more appropriate to our needs, mainly because it is a wide and language-independent classification which allows us to deal not only with errors but also with positive characteristics that must be considered. The analysis conducted was of great help to confirm those linguistic features that had already been used by state-of-the-art MT metrics, but also to highlight other kinds of linguistic information relevant to the evaluation of MT, such as the use of hyponyms/hypernyms as regards lexical

[1] Sources available at https://github.com/jatserias/VERTa.

[2] http://www.statmt.org/wmt10/evaluation-task.html.

[3] http://www.itl.nist.gov/iad/mig/tests/mt/2006/.

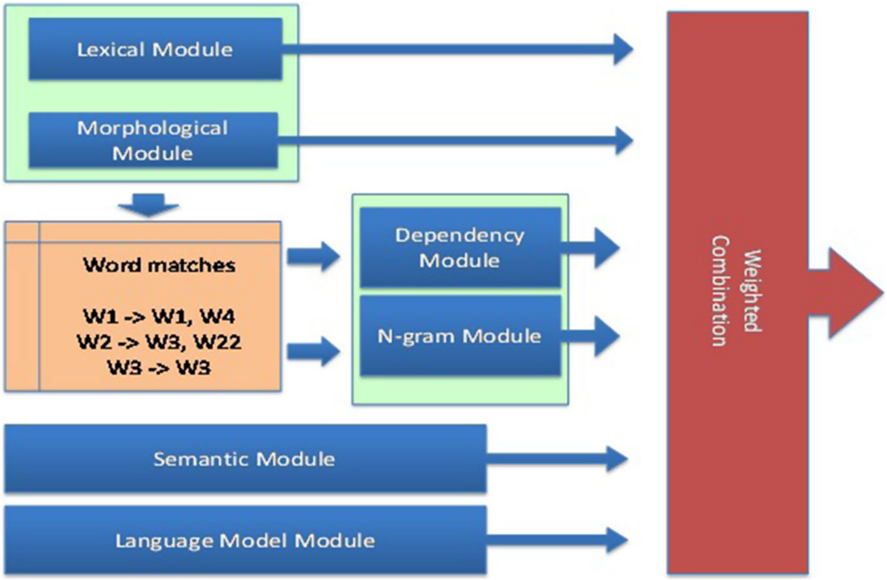


Fig. 1 VERTa's architecture

semantics and valency alternations as regards syntax, to mention a couple of examples. Likewise, it supports the use and combination of linguistic information to ensure a wide and holistic approach to MT evaluation; and finally, it shows that the importance of the linguistic traits used varies depending on the language assessed and the type of evaluation.

According to the classification established in the initial linguistic analysis, VERTa consists of several modules working at different levels: Lexical Module, Morphological Module, Dependency and Semantic Module. Moreover, an N-gram Module accounting for similarity between chunks, as well as a Language Model (LM) Module are included. The fact of organising the linguistic features in different modules or levels allows different types of evaluations (i.e. adequacy, fluency and ranking), thus checking the suitability of linguistic features for each type.

Each module in VERTa works first individually and the final score is the Fmean of the weighted combination of the Precision and Recall of each module in order to get the results which best correlate with human judgements (see Fig. 1). This way, the different modules can be weighed depending on their importance regarding the type of evaluation and language evaluated. In addition, the modular design of this metric makes it suitable for all languages. Even those languages that do not have a wide range of NLP tools available could be evaluated, since each module can be used in isolation or in combination.

The first module applied in VERTa is the Lexical Module, which serves as the base of the alignment. VERTa allows two possible alignments: the first one only takes into account the matches set in the Lexical Module, whereas the second one uses a combination of both Lexical and Morphological modules (i.e. PoS) which implies a more restrictive alignment. The type of alignment used will depend on the type of evaluation.

The matching procedure follows a greedy approach. The best non-conflicting token to token alignment is chosen iteratively (based on the matching score between tokens). Higher level alignments (e.g. n-grams) are based on the token/word level alignment. The metric implementation is configurable so that different strategies can be implemented in order to build better (more global) alignments.

All modules (except for the Language Model) use a weighted precision and recall over the number of matches of the particular element of each level (words, dependency triplets, n-grams, etc.) as shown below.

$$P = \frac{\sum_{\partial \in D} w_{\partial} * nmatch_{\partial}(\nabla(h))}{(\nabla(h))} \quad R = \frac{\sum_{\partial \in D} w_{\partial} * nmatch_{\partial}(\nabla(r))}{(\nabla(r))}$$

where r is the reference, h is the hypothesis and ∇ is a function that given a segment will return the elements of each level (e.g. words at lexical level and triplets at dependency level). D is the set of different types of matching. $nmatch_{\partial}()$ is a function that returns the number of matches of type ∂ (e.g. the number of lexical matches at the lexical level or the number of dependency triplets that perfectly match at the dependency level). Finally, W is the set of weights [0 1] associated to each of the different types of matching in order to combine the different kinds of matches considered in that level.

VERTa uses the Fmean to combine Precision and Recall measures. If there's more than one reference, the maximum Fmean among all references is returned as the score. When the scores per module are calculated the final score is a weighted average of the different scores (Fmean) of the modules.

All modules forming VERTa and the linguistic features used are described in detail in the following subsections.

2.1 Lexical module

The Lexical Module compares lexical items from the hypothesis segment with those in the reference segment. The approach followed in this module was inspired by METEOR (Banerjee and Lavie 2005) in the sense that the module relies on lexical items and lexical semantic relations. However, while the most recent versions of METEOR (Denkowski and Lavie 2011, 2014) deal with semantics by means of synonymy and paraphrase tables, our metric does not only use synonymy but it also makes good use of the lexical semantic relations[4] shown in Table 1, such as hypernymy and hyponymy avoiding the use of paraphrase tables which have to be built up for each language and domain. Moreover, VERTa also employs the information provided by lemmas and partial lemmas (i.e. first characters in a lemma), whereas METEOR relies only on stemming. In addition, a system of weights is also applied on the different matches established depending on their importance as regards semantics.

[4] The lexical semantic relations used are obtained from WordNet 3.0 (Fellbaum 1998).

Table 1 Lexical matches and examples

	Match	Examples	
		Hypothesis	Reference
1	Word-form	*east*	*east*
2	Lemma	*is_BE*	*are_BE*
3	Synonym	*believed*	*considered*
4	Hypernym	*barrel*	*keg*
5	Hyponym	*keg*	*barrel*
6	Partial lemma	*danger*	*dangerous*

2.2 Morphological module

The Morphological Module allows for combining lexical and morphological information or using morphological information by itself. When used in the combinatory fashion, the module is based on the matches established in the Lexical Module in combination with PoS tags from the annotated data.[5]

On the other hand, when only morphological information is used, it is only based on PoS matches between the hypothesis and reference segments. The aim of this module is to compensate for the broader coverage of the Lexical Module, preventing matches such as *invites* and *invite*, which although similar in terms of meaning, differ on their morphosyntactic information. Therefore, this module seems to be more appropriate to assess the fluency of a segment rather than its adequacy. In addition, although this module may not play a key role when assessing English output, it might be particularly useful when evaluating languages with a richer inflectional morphology (e.g. Romance languages).

2.3 Dependency module

The use of the Dependency Module proves to be effective in order to establish similarities between equivalent sentences which show a different constituent order. In Example 1, the adjunct of time *today Wednesday* occupies different positions in the hypothesis and reference strings. By means of the dependency analysis, we can state that although located differently inside the sentence, both subject and adjunct depend on the verb (see Table 2).

Example 1
HYP: *Ramallah (West Bank) 2–15 (AFP)—The executive committee of the PLO said today Wednesday that…*
REF: *Ramallah (West Bank) 2/15 (AFP)—Today, Wednesday, the Executive Committee of the Palestine Liberation Organization expressed the opinion…*

[5] The data has been annotated by the Stanford Log-Linear Part of Speech Tagger (Toutanova et al. 2003), included in the Stanford CoreNLP suite.

Table 2 Comparison between hypothesis and reference triplets

Hypothesis	Reference
nsubj(committee, said)	nsubj(committee, expressed)
tmod(today, said)	tmod(today, expressed)

Table 3 Dependency matches

	Match type	Match description
1	Exact	Label1 = Label2 Head1 = Head2 Mod1 = Mod2
2	No_label	Label1 ≠ Label2 Head1 = Head2 Mod1 = Mod2
3	No_mod	Label1 = Label2 Head1 = Head2 Mod1 ≠ Mod2
4	No_head	Label1 = Label2 Head1 ≠ Head2 Mod1 = Mod2

This module works at sentence level and follows the approach used by Owczarzack et al. (2007a, b) and He et al. (2010) with some changes and linguistic additions in order to adapt it to our metric combination. One of the differences between the above mentioned proposals and VERTa's is that they used an LFG parser and MALT parser respectively, whereas the parser used in VERTa is the Stanford parser (De Marneffe et al. 2006). The reason why the Stanford parser is used is because after conducting an evaluation (Comelles et al. 2010) where the performance of several dependency parsers was assessed that proved to be the best in terms of linguistic quality.

Similar to the Morphological Module, the Dependency Module also relies first on those matches established at lexical level—word-form, synonymy, hypernymy, hyponymy, lemma and partial lemma—in order to capture lexical variation across dependencies and avoid relying only on surface word-form.

Then, by means of flat triplets with the form Label(Head, Mod) four different types of dependency matches are considered (see Table 3) and weights can also be assigned to each type of match.

These matches are applied in the order established in Table 3. First VERTa looks for Exact matches (i.e. triplets in the hypothesis and reference segments are identical). Then, the metric moves to the No_label match, thus comparing triplets that show identical head and modifier but different label, as shown below.

Example 2
HYP:…***all*** *Palestinian political* ***parties***…
REF:…***all*** *the Palestinian political* ***parties***…

$$\mathrm{predet}(\mathbf{parties}, \mathbf{all}) = \mathrm{det}(\mathbf{parties}, \mathbf{all})$$

With the triplets left, VERTa tries to establish matches between those triplets that show the same label and head but different modifier, as illustrated next.

Example 3
HYP: *…the situation more **difficult** and complicated and **serious**…*
REF: *…the situation is more **difficult**, complicated and **dangerous**…*

$$\textbf{conj\_and}(\textbf{difficult}, \text{dangerous}) = \textbf{conj\_and}(\textbf{difficult}, \text{serious})$$

Finally, the metric looks for triplets that share the same label and modifier but different head:

Example 4
HYP: *…He **said** "I **believe** that the situation…"*
REF: *…He **added** "I **think** the situation…"*

$$\textbf{ccomp}(\text{said}, \textbf{believe}) = \textbf{ccomp}(\text{added}, \textbf{think})$$

The No_head match, was also proposed by Owczarzak et al. (2007a, b); however, He et al. (2010) disregarded this type of match in their proposal. Although no arguments were given for such a decision, we might think that it did not correlate well with human judgements. In our metric, we decided to use it because we were interested in checking its suitability, not only as regards correlation with human judgements but also regarding linguistic analysis.

Following He et al. (2010)'s approach, dependency labels are given different weights depending on their suitability and importance depending on the type of evaluation.

As regards the way the final score for this module is calculated, a couple of parameters are considered: the type-of-match weight and the dependency-relation weight. Therefore, each triplet match combines the weight given to the type of match and the weight assigned to the dependency label. Then matches are added up and precision and recall are calculated.

Finally, a set of language-dependent rules has been added with two goals: (1) capturing similarities between different syntactic structures conveying the same meaning, in case the dependency matches overlook them (e.g. active–passive alternation, post-modifier *of*-PP and possessive '*s*); and (2) restricting certain dependency relations (e.g. subject word order when translating from Arabic to English). Thanks to the linguistic analysis performed beforehand, the development of this set of rules (a total of 10 working at phrase and clause level) was quite easy and straightforward.

2.4 N-gram module

The N-gram Module matches chunks in the hypothesis and reference segments, similar to BLEU. However, it can rely either on the matches set by the Lexical Module or the matches set by the Morphological Module; in other words, combining lexical matches and PoS information or using PoS information isolated. Chunks length may go from bigrams to sentence length, depending on the type of evaluation.

The use of this module allows the combination of both linguistic and statistical approaches and enables us to deal with word order inside the sentence by means of a more simple approach than the parsing of constituents.

2.5 Semantic module

Semantics plays an important role in the evaluation of adequacy. This has also been claimed by Lo and Wu (2010) who report that their metric based on semantic roles (SR) outperforms other well-known metrics when adequacy is assessed. The Semantic Module in VERTa does not use information on SRs since dependency relations are thought to be halfway between syntax and semantics, thus one of our hypotheses is that the Dependency Module could also provide information in this sense. The Semantic Module uses other semantic information at both lexical and sentence level: NEs, Time Expressions and Sentiment analysis.

Regarding NEs, named entity recognition (NER) and named entity linking (NEL) are used. Following previous NE-based metrics (Reeder et al. 2001; Giménez 2008) the NER component captures similarities between NEs in the hypothesis and reference segments. In order to identify NEs the Supersense Tagger (Ciaramita and Altun 2006) is used. On the other hand the NEL component focuses only on those NEs that appear on Wikipedia, which allows for linking NEs in the hypothesis and reference segments regardless of their external form. Thus, *EU* and *European Union* will be captured as the same NE, since both of them are considered as the same organisation in Wikipedia. The NEL component uses a graph-based NEL tool inspired by Hachey et al. (2011) which links NEs in a text with those in Wikipedia pages.

As regards the time expressions (TIMEX) component, it matches temporal expressions in the hypothesis and reference segments regardless of their form. The tool used is the Stanford Temporal Tagger (Chang and Manning 2012) which recognizes not only points in time but also duration. By means of the TIMEX component, different syntactic structures conveying the same time expression can be matched, such as *on February 3rd* and *on the third of February*.

Finally, Sentiment analysis has been added using the dictionary strategy described in Atserias et al. (2012). Sentiment analysis provides information regarding the contextual polarity of the sentence, whether it has a positive or negative connotation.

2.6 Language model module

The Language Model (LM) Module works differently from the rest of modules, in the sense that it neither tries to find similarity matches between the hypothesis and reference segments, nor tries to compare them. This module is only applied to the hypothesis segment and uses the News LM[6] to calculate the degree (log probability) to which the hypothesis segment is expected compared to what occurs in the corpus used to build the language model. A language model assigns a probability to a sequence of words (N-grams), thus it is possible to obtain the most frequent N-grams for a specific domain. By using a language model we aim at accounting for those segments that, even being syntactically different from their corresponding

[6] This LM was used as a baseline feature in the WMT13 Quality Estimation Task (http://www.statmt.org/wmt13/quality-estimation-task.html).

reference translations, are still fluent; in other words, we will be able to check the correct construction and plausibility of the hypothesis, even if it is very different or not included in any of the reference segments. The use of LMs is also widely extended in Quality Estimation.

3 Experiments and results

This section describes the experiments conducted and results obtained with the aim of studying and testing the suitability of the linguistic features used in VERTa, the influence of each module and the best way to combine them in order to evaluate adequacy or fluency. The experiments conducted take correlation coefficients as a point of departure and focus on providing linguistic evidence, supported with examples, of the suitability of those linguistic features used and the influence of each module and their combination. Thus, so as to perform such a fine-grained evaluation of the linguistic information, experiments are carried out at segment level. Each module is first tested separately and later in combination. Modules' weights were first assigned manually, following linguistic criteria; although later in order to calculate an upper-bound for the weight tuning, all possible weight combinations were tuned automatically using a 0.01 step.

The experiments of both adequacy and fluency were based on scores instead of ranking, since we consider scores to be more informative for our research. The Pearson Correlation Coefficient (1914/1924/1930) was used to compare the scores provided by the metric with those provided by human judges. Traditionally, researchers use correlation as a way to measure the performance of their metrics and to check the suitability of the features used. In our case, the information obtained from correlating VERTa's scores with human judgements was used as a guide to know whether we were making progress and we were advancing in the correct way. However, since we were especially interested in checking the suitability of linguistic information in order to evaluate MT, besides using information provided by correlations as a guide, we also performed a qualitative and detailed analysis of the metric's output every time linguistic features were added and/or combined. This analysis was possible due to the fact that VERTa does not only provide a score per segment but it also provides an XML file where linguistic features used in each module and their corresponding matches can be traced (Comelles and Atserias 2016). Therefore, every time a new linguistic feature was added and/or combined, first the correlation with human judgements was checked as a hint to see whether they improved or worsened. In both cases a set of segments either improving or worsening their scores were selected and analysed in depth in order to study how linguistic features influenced the metric for better or for worse, and a final decision on the use of such features could be made.

3.1 Experiments on adequacy

So as to perform these experiments part of the development data provided in the MetricsMaTr 2010 shared-task was used. From the data provided by the organization we used a total of 800 segments (Arabic to English) of the NIST Open-MT06 data, the MT output from 8 different MT systems (100 segments/system) and 4 reference translations. The human judgments used were based on adequacy (7-point scale, straight average). In order to calculate correlations at segment level Pearson correlation was applied between our metric and the adequacy judgments.

Our hypothesis was that those linguistic features that should have a stronger influence when evaluating adequacy were lexical semantics, syntactic information and sentence semantics. In VERTa's modules these features were included in the Lexical Module, the Dependency Module, the N-gram Module and the Semantic Module. Therefore, experiments were performed first, module per module, and later in a combinatory fashion.

According to Pearson correlation, the most effective intra-module settings are as follows:

- The Lexical Module proves most effective (0.743) when word-forms and synonyms receive the maximum weight (1), whereas lemmas and partial lemmas are assigned lower weights, 0.8 and 0.6 respectively.
- The Dependency Module performs best when the following matches and weights are used: Exact and No_label match (maximum weight, 1), No_mod match (0.9) and No_head match (0.7). In addition, dependency categories are also assigned different weights depending on how informative they are, thus most of the categories receive the maximum weight (1) except for *det*, *num* and _[7] that receive (0.5). Finally a set of language-dependent rules have been added to restrict the position of the subject, which improves the correlation with human judgements up to 0.752.
- The N-gram Module achieves its best results (0.701) when based on lexical items and with a shorter n-gram distance (i.e. bigrams).
- According to the experiments performed, the Semantic Module shows a low correlation with human judgements on adequacy. This is due to the fact that only partial aspects of translation are considered in this module, whereas human judgements cover the adequacy of the entire hypothesis segment. From the features contained in this module, the one that correlates best individually is NEs recognition (0.338), whereas Sentiment analysis correlates the worst (0.132). However, all components have been finally used since the correlation of the whole module improves (0.390) when all of them are combined.

As shown in Table 4, not all modules are suitable for the evaluation of adequacy. As expected the combination of the Lexical and Dependency Modules proves to be the most effective to assess adequacy, although neither the N-gram Module nor the Semantic Module should be disregarded. The Lexical Module has the strongest influence (0.47), followed by the Dependency Module (0.43).

A closer analysis of these results and of those examples that most benefit from this combination shows the following:

[7] *det* stands for determiner; *num* stands for numeral and _ refers to those intermediate categories that help moving from standard dependencies to collapsed dependencies.

Table 4 Weighted combination of modules and Pearson correlation results

Modules combination	Pearson correlation
All modules—same weight	0.617
Lexical M. (0.47), Dependency M. (0.43), N-gram M. (0.05) and Semantic M. (0.05)	**0.781**

Scores in bold indicate the best correlations obtained

- Firstly, the Dependency Module infers relations that might be disregarded if only the Lexical Module is taken into account, as illustrated in Example 5.[8]

Example 5
HYP: ***He*** *said* ***“that all these positions unfair to*** *the right* **people,** *US, and* **we** *now* **possess** *an* **Islamic** *or the* **Palestinians and Arabs options”.**
REF: ***He*** *added, “We emphasized* ***that all these positions*** *are* ***unfair to*** *our* ***people*** *but that* ***we have alternative Palestinian, Arab****,* ***and Islamic*** *resources.”*

In the hypothesis segment the copula verb *are* is missing, however, the meaning is not affected, and a potential reader could still infer that *all these positions are unfair*. If only the Lexical Module was taken into account, this no-match would penalise the hypothesis segment; however, if the Dependency Module is used, the relation between the subject *positions* and the Subject Complement (Cs) *unfair* is still preserved as shown below (see Table 5), where the analysis of the hypothesis segment accounts for a dependency relation between *position* and *unfair*, even though the type of relation cannot be established due to the missing copula verb. In addition, by means of the Dependency Module the clauses introduced by *and* and *but* in the hypothesis and reference segments, respectively, can also be connected to the previous clause. Although the meaning of both connectors is clearly different, it does not seem to affect the meaning of the whole sentence. Finally, the Dependency Module also accounts for the last part of the sentence which shows a different word order as well as a clearly disfluent clause *we now possess an Islamic or the Palestinians and Arabs option*s, but whose meaning can still be understood.

- Secondly, the Dependency Module accounts for matches between different syntactic structures that express the same meaning, as the X-Complement (XCompl) and the Oblique (Obl) dependents illustrate in Example 6.

Example 6
HYP: *This series of events in the Beba province [Subj] started [Verb]* ***burning five churches*** *[XCompl] in the 3rd February [Adj]*
REF: *The series of incidents [Subj] began [Verb]* ***with the burning of five churches in Bibb County on February 3rd*** *[Obl].*

[8] Lexical module matches in bold and N-gram module matches underlined.

- Finally, some components in the Semantic Module prove considerably effective, such as the TIMEX components, which links equivalent temporal constructions regardless of their external form, as shown by the phrases in bold in the example below.

Example 7
HYP: …*HAMAS who won the legislative elections* ***in late January***…
REF: …*the movement, which won the legislative elections* ***at the end of January***…

3.2 Experiments on fluency

In order to carry out these experiments, data containing human judgements on fluency was used. This data was granted by the National Institute of Standards and Technology (NIST)[9] and the Language Data Consortium (LDC),[10] from their NIST 2005 Open Machine Translation (OpenMT) Evaluation campaign.[11] This data includes MT output from Arabic into English from 6 different systems, 4 reference translations and 5-scale human judgements on fluency. From this data, 600 segments (100 segments/system) were used as development data in order to conduct experiments on fluency, the rest of the data was kept unseen to conduct an evaluation of the metric. Similar to the experiments on adequacy, Pearson correlation was applied between the metric and the fluency judgments to calculate correlations at segment level.

Our hypothesis suggests that those linguistic features that should be more suitable to evaluate the fluency of a segment are those related to syntactic and morphosyntactic information, which are covered by the Morphological Module, Dependency Module, N-gram Module and LM Module. Therefore, those four modules were tested first individually and later in combination.

According to Pearson correlation, each module shows its best performance when the following intra-module settings are used:

- The Morphological Module works best (0.217) when lexical matches are combined with PoS and all matches weigh the same.
- In the Dependency Module, the most effective type of match to evaluate the fluency of a segment is the Exact match (0.310). Even though the correlation with human judgements indicates that the combination of Exact match + No_Mod Match achieves the best results, the linguistic analysis has proved that the only positive effect of this combination is to widen the coverage of matches related to lexical semantics (i.e. semantically related words not captured by the Lexical Module). On the other hand, such combination overlooks the omission/mistranslation of determiners, prepositions and conjunctions which cannot be disregarded when fluency is assessed. As for the dependency labels, they should be organized into three categories (i.e. top nodes—dependency relations

[9] https://www.nist.gov/.

[10] https://www.ldc.upenn.edu/.

[11] https://catalog.ldc.upenn.edu/LDC2010T14.

Table 5 Dependency matches corresponding to Example 5

Hypothesis	Reference	Match
dep(unfair, positions)	nsubj(unfair, positions)	No_label match
conj_and(unfair, possess)	conj_but(unfair, have)	No_label match
dobj(possess, Islamic)	dobj(have, Palestinian)	No_head match
conj_or(Islamic, Palestinians)	NO MATCH	No match
conj_and(Palestinians, Arabs)	conj_and(Palestinian, Arab)	Exact match
dep(Palestinians, options)	amod(Palestinian, alternative)	No_label match

affecting the arguments of the verb, auxiliary verbs and copular verbs; middle nodes—dependency relations affecting adjuncts and phrase level modifiers and complements; and ultimate nodes—dependency relations related to punctuation marks, and unlabeled constituents), which receive different weights: 1, 0.5 and 0, respectively. Finally, from the language-dependent rules added, those that allow for comparing different syntactic structures conveying the same meaning slightly improve the correlation with human judgements (0.383), since they broaden the restrictive coverage of the Exact match.

- The N-gram Module shows its best performance with large n-grams (bigrams to sentence-grams) calculated over PoS (0.345).
- The Language Model Module shows its best correlation with human judgements (0.257) when the News LM is used.

When combined (see Table 6), the Dependency Module is clearly the module that most contributes to the performance of the metric, next is the LM Module followed closely by the N-gram Module. Finally, the Morphological Module contributes slightly to the performance of the metric. The N-gram Module and LM Module complement each other, since the first accounts for PoS n-grams while the second focuses on n-grams over lexical items that might not occur in the reference translations. The small contribution of the Morphological Module can also be explained because a) the N-gram Module is already taking into account PoS information, covering issues such as agreement; and b) English does not show a rich inflectional morphology, thus individual PoS matching is not that important.

Some of the grammaticality issues that could be detected with the use of the modules combination reported above are the following:

- Sentences without subject. In English all sentences must contain a subject in order to be grammatical, however this is still a problem for some machine translation engines which are either unable to translate the subject or provide an incorrect translation, mainly using 3rd person singular pronoun *he* in its place. Missing subjects affect not only adequacy but also fluency, as shown in Example 8. *Bouzoubaa*, the subject of the main clause in the hypothesis sentence, is missing, thus affecting the grammaticality of the segment. The use of the

Table 6 Weighted combination of modules and Pearson correlation results

Modules combination	Pearson correlation
All modules—maximum weight	0.403
Morphological M.(0.04), Dependency M.(0.37), N-gram M.(0.29) and LM M.(0.30)	0.434

Dependency Module with the Exact match and a higher weight to top-level dependency relations help to detect this type of issues.

Example 8
HYP: *In an interview with the newspaper le "**<u>Ø</u>**[12] **confirmed** that the persons involved in terrorist cases in the Netherlands…".*
REF: *In an interview with the "Aujourd'hui le Maroc" newspaper, **<u>Bouzoubaa</u> stressed** that the people involved in the terror cases in Holland…*

- Lexicogrammatical patterns. The type of complements that verbs take plays an important role in the grammaticality of a sentence. Examples 9 and 10 illustrate their importance.

Example 9
HYP: *He **said** Ardogan station "TV" television that "the European Union cannot address…"*

The default pattern that verb *say* enters is SVOObl (say something to somebody), however, this verb can also subcategorize for a clause complement (ClCompl) realised by a that-clause. In this case, the pattern would be SVClCompl (*say that*…). Thus, the dependency parser analyses the chunk *Ardogan station TV television that "the European Union cannot address*… as the direct object of the main verb, where *address* and *television* are linked by the dependency tag *dep* which indicates that this is an unnatural grammatical structure. In this case, the verb used should have been *tell* which accepts *tell somebody something*. Furthermore, it must also be noticed that *Ardogan* should occupy the subject position instead of *He*.

In Example 10, attention should be paid to the chunk in bold *see each warned of Morroccan terrorist acts committed in the Netherlands*.

Example 10
HYP: *The minister added, "which is why I said to **see each warned of Morroccan terrorist acts committed in the Netherlands**."*

The verb *see* subcategorizes for a direct object, however, due to a bad translation there is no noun that could work as the head of the direct object. As a consequence, the analysis provided by the dependency parser links *see* and *warned* by means of the tag *dep*, indicating, again, that there is an unnatural grammatical structure.

[12] Missing subject.

- Word order of immediate constituents. Sometimes a constituent itself might show a correct internal grammatical structure, but it might occupy an ungrammatical position at clause level resulting in an ungrammatical sentence. Example 11 illustrates this fact.

Example 11
HYP: *Baghdad 24–12 (AFP)—**accused** [Shiite leader of the hardline young issued] [today, Friday,] [Israel and the United States and Britain] [of being behind the bloody attacks against the cities, Najaf and Kerbala last Sunday, which claimed the lives of 66 people dead and some 200 injured].*
REF: *Baghdad 12–24 (AFP)—[The young radical Shiite leader Muqtada Al-Sadr] **accused** [today, Friday], [Israel, the United States and Britain] [of being behind the bloody attacks that targeted the two cities of Najaf and Karbala last Sunday and in which 66 people were killed and about 200 injured].*

In Example 11 hypothesis, the NP realising the subject has not been translated properly and, in addition reordering is needed, as it occupies the position of the object. Consequently, the sentence is clearly disfluent and although some of the immediate constituents present a correct internal grammatical structure, the grammaticality of the whole sentence is clearly affected. The grammaticality of the constituents internal structure is mainly captured by the N-gram Module, which provides better results (see Table 7) than the Dependency Module which is clearly affected by the ungrammatical position of the immediate constituents.

- Word order inside the phrase. The English default word order Pre-modifier + Noun is not always kept in machine translation. This does not affect the meaning of the sentence but its fluency, as illustrated by the phrases *detainees Moroccans* and *Moroccan detainees* in Example 12. In this case, the role played by the N-gram Module and the LM Module is crucial, since the dependency parser can sometimes handle word order differences and analyse correctly those chunks even if the pre-modifier follows the noun, instead of preceding it.

Example 12
HYP: *He said that in Spain "suspected of some **detainees Moroccans** clearly they participated directly or indirectly in preparation…"*
REF: *Bouzoubaa said that in Spain "some **Moroccan detainees** are clearly suspected of having directly or indirectly participated in the preparations…"*

Last but not least, it is worth mentioning the use of the LM Module. The LM model works as a complement to the reference translations, since those grammatical chunks not covered by the reference segments can be covered by the LM. This is the case of Example 13 where the use of the LM moves the score of the metric from 1.4 (using dependency and N-gram Modules) up to 2.5, coinciding with the human judgement for this segment.

Example 13
HYP: *He said the official, who asked* ***to remain anonymous**, "we support if the meeting is aimed at* ***helping the Palestinian Authority*** *at the level of economic and encourage them to undertake reforms".*
REF: *The official, who wished* ***to remain anonymous**, said "we support this meeting if the aim is to* ***help the Palestinian Authority*** *economically and to encourage it to make reforms".*

The Dependency Module accounts for the chunks in bold whereas the N-gram Module matches the chunks underlined. In addition, by employing an LM we can account for the grammaticality of other chunks, such as *if the meeting is aimed at*, which were not covered by any of the previous modules because it does not occur in the reference sentence. Thus, using an LM in combination with other modules aimed at checking the grammaticality of a segment turns into a positive contribution.

4 Evaluation

This section describes the evaluation carried out in order to check the validity of VERTa to evaluate adequacy and fluency separately. To this aim an evaluation at segment level has been performed and the metric has been compared to other well-known MT metrics included in the Asiya framework[13] (Giménez and Márquez 2010a; González and Giménez 2014).

4.1 Evaluating adequacy

In order to carry a meta-evaluation on adequacy, the unseen part of the newswire dataset (Arabic-English) described in Sect. 3 was used. This corpus contains 1192 segments translated by 8 different systems (149 segments/system), 4 reference translations and adjusted human judgements for adequacy. In order to check VERTa's performance, the same dataset has also been evaluated by several other metrics contained in the Asiya framework:

- BLEU: accumulated BLEU score up to 4-grams.
- METEOR-ex, METEOR-st, METEOR-sy and METEOR-pa: METEOR using only exact matching (METEOR-ex), adding stem matching (METEOR-st), plus synonymy matching (METEOR-sy), plus paraphrase matching (METEOR-pa).
- SP-Op(*) and SP-Oc(*): metrics using shallow parsing. SP-Op(*) calculates the average lexical overlap over PoS tags. SP-Oc(*) calculates the average lexical overlap over all chunk types.
- DPm-Ol(*), DPm-Oc(*) and DPm-Or(*). These measures capture similarities between dependency trees in the hypothesis and reference segments and use the MALT v1.7 parser to analyse the segments. DPm-Ol(*) calculates overlapping

[13] http://asiya.lsi.upc.edu/.

Table 7 Score per module corresponding to Example 11

Modules	Score obtained
N-gram module	0.2702
Dependency module	0.1690

between words hanging at all levels, DPm-Oc(*) calculates overlapping between grammatical categories, and finally, DPm-Or(*) calculates overlapping between grammatical relations.

- CP-Op(*) and CP-Oc(*).[14] These measures compare similarities between constituent parse trees in the hypothesis and reference segments. The Charniak and Johnson (2005)'s Max-Ent reranking parser is used to obtain the constituent trees. CP-Op(*) calculates lexical overlap over PoS and CP-Oc(*) calculates lexical overlap according to the phrase constituent.
- SR-Or, SR-Or(*) and SR-Mr(*). These metrics compare Semantic Roles similarities between the hypothesis and reference segments. SR-Or deals with Semantic Roles overlap regardless of their lexical realization. SR-Or(*) computes the average lexical overlap over all Semantic Roles types. SR-Mr(*) calculates the average lexical matching over all Semantic Roles types.
- NE-Me(*) and NE-Oe(*). This set of metrics compares the hypothesis and reference segments according to their NEs. The NE-Me(*) calculates the average lexical matching over all NEs whereas the NE-Oe(*) calculates the average lexical overlap over NEs.
- Combination of metrics 1: The ULC (Unified Linear Combination) combination of metrics that are representative of each linguistic level in Asiya (Giménez and Márquez 2008). This set of metrics includes: BLEU, NIST, -TER, -TERp-A, ROUGE-W, METEOR-ex, METEOR-pa, METEOR-st, METEOR-sy, DP-HWCM_c-4, DP-HWCM_r-4, DP-Or(*), CP-STM-4, SR-Or(*), SR-Mr(*), SR-Or, DR-Or(*), DR-Orp(*). They are combined by means of the normalized arithmetic mean of all metrics' scores.
- Combination of metrics 2: The ULC combination of metrics that according to Giménez and Márquez (2010b) show the best performance in several data sets to evaluate quality. This combination of metrics is: ROUGE-W, METEOR-sy, DP-HWCM_c-4, DP-HWCM_r-4,[15] DP-Or(*), CP-STM-4, SR-Or(*), SR-Mr(*), SR-Or, DR-Or(*), DR-Orp(*).

[14] Although both SP and CP metrics use the Penn Treebank PoS tagset, SP metrics use a different tool to automatically annotate sentences [SVM tool (Giménez and Márquez 2004) and BIOS (Surdeanu and Turmo 2005)], thus its different performance.

[15] In the original combination of metrics, there were two metrics that are not available in the Asiya framework nowadays, DP-HWCM_c and DP-HWCM_r, and which have been substituted by the variants DP-HWCM_c-4 and DP-HWCM_r-4.

The modules in VERTa were set and weights were assigned according to the experiments on adequacy, as follows: Lexical Module (0.47), Dependency Module (0.43), N-gram Module (0.05) and Semantic Module (0.05).

Correlations with human judgements obtained by these metrics have been compared to the correlation obtained by VERTa and both a quantitative and a qualitative analysis of the results has been conducted.

Table 8 shows the Pearson correlation obtained by the metrics described above and by VERTa.

4.1.1 *Analysis of the results*

According to the results obtained, VERTa stands out from the rest of the metrics obtaining a correlation of 0.728, whereas the closest metrics get 0.650 (Combination 1), 0.629 (SP-Oc/CP-Oc metrics) and 0.616 (CP-Op). The key to VERTa's excellent performance is the combination of linguistic information at different levels that enriches the metric and allows for a more flexible use.

A closer analysis of the results shows that those metrics working at lexical level (BLEU and METEOR family) obtain similar results. It is interesting to notice that in the METEOR family, the more linguistic information used, the worse the correlation obtained. The only type of information which improves its correlation is the use of stemming; however, the use of synonymy has the opposite effect. This is quite surprising since adequacy is being evaluated, thus the use of synonymy relations seemed to be appropriate. Actually, the use of synonyms in VERTa has proved effective to increase the metric's correlation with human judgements.

Regarding those metrics using syntactic information, their performance seems to contradict the common belief that this type of metrics is the most effective one to evaluate the fluency of a segment (not recommending their use for the evaluation of adequacy), since some of them (SP-Oc, CP-Oc and CP-Op) obtain a good correlation with human judgements on adequacy. It is noticeable that those that work at chunk and phrase constituency level achieve the best results. Hence, this seems to indicate that word order is also important when evaluating adequacy, confirming the modules combination in VERTa, where the N-gram Module also proved suitable. On the other hand, those metrics working with dependency trees do not obtain good results. Within the DPm familiy, both DPm-Ol (0.543) and DPm-Or (0.574) show a better performance than DPm-Oc (0.268) because they compare lexical items, the former, and dependency relations, the latter. However their low performance in comparison with VERTa's Dependency Module might be due to the fact that both metrics are much more rigid than VERTa. The key factors for VERTa's better performance are that (a) in VERTa's Dependency Module, information regarding lexical semantics has also been taken into account; (b) VERTa's Dependency Module considers different types of matches and rules which lead to a more flexible coverage of dependency relations and allows for similarity between different syntactic structures conveying the same meaning, even if they are not totally grammatical; (c) in VERTa, the least informative dependency relations are assigned very low weights. Another factor that might also be worth considering when comparing VERTa's Dependency Module and the DPm family is the selection of the dependency parser used to perform the analysis (see Comelles et al. 2010's paper on evaluating constituency and dependency parsers); VERTa uses the Stanford parser, whereas the DPm family makes use of the MALT parser.

Finally, as regards semantically-related metrics—SR-based metrics and NEs-based metrics—they did not obtain a good correlation. Actually, a better performance was expected, especially from those using Semantic Roles information. According to Lo and Wu (2010), this type of information is especially useful when evaluating adequacy; however, results obtained by the SR-metrics contradict their statement. It must be noticed that those metrics that compare Semantic Roles taking into account lexical items—SR-Or(*) (0.392) and SR-Mr(*) (0.307)—work better than that which disregards their lexical realization (SR-Or), which gets 0.182. VERTa does not use information on Semantic Roles but the semantic relations within a sentence can be captured by the Dependency Module, since dependency relations are considered to be an interface between syntax and semantics.

Last but not least, NEs-based metrics obtained a low correlation (0.304 for NE-Me(*) and 0.332 for NE-Oe(*)), similar to those obtained by the NE-based components in VERTa. These results were expected since, as explained in Sect. 3.1, NEs are just a partial aspect of the segment and human judgements used for correlation assess the entire hypothesis segment. Nevertheless, it must be noticed that even though these metrics do not correlate well in isolation, they slightly contribute when combined with other modules.

In addition, since VERTa combines linguistic features at different levels, two combinations of some of the metrics available in Asiya have also been used. Results for Combination 1[16] confirm our hypothesis that the combination of several metrics working at different levels correlate better with human judgements than single metrics working at a specific level. On the other hand, according to the correlations obtained, VERTa outperforms significantly Combination 1, which gets 0.650. This is mainly due to the fact that VERTa's individual modules are more flexible and use more linguistic information than those in that combination. In addition, it must also be highlighted that metrics in Combination 1 are combined using the normalized arithmetic mean of all metrics scores, whereas VERTa selects and weighs each module depending on the type of evaluation. Finally, Combination 1 uses a wide range of metrics so it is difficult to check the influence of each metric and whether any of them represents a drawback to this type of evaluation. Thus, it seems that the combination of such a large amount of metrics is not that effective and a selection of metrics covering the key linguistic features related to the meaning of a sentence (e.g. those in VERTa) has proved more useful to evaluate adequacy.

4.2 Evaluating fluency

Once VERTa has been evaluated on adequacy, it is the turn for fluency. To this aim, the unseen part of the news-related corpus (Arabic-English) described in Sect. 3.2 has been used. This corpus contains 1192 segments translated by 6 different systems

[16] The one obtaining the best results between the two.

Table 8 Pearson correlation for adequacy. Comparing VERTa metric and a selection of well-known metrics

Metric	Pearson correlation
VERTa	**0.728**
Metric combination 1	**0.650**
SP-Oc(*)	**0.629**
CP-Oc(*)	**0.629**
CP-Op(*)	**0.616**
Metric combination 2	0.578
BLEU	0.577
DPm-Or(*)	0.574
METEOR-st	0.571
SP-Op(*)	0.570
METEOR-sy	0.569
METEOR-ex	0.568
METEOR-pa	0.552
DPm-Ol(*)	0.543
SR-Or(*)	0.392
NE-Oe(*)	0.332
SR-Mr(*)	0.307
NE-Me(*)	0.304
DPm-Oc(*)	0.268
SR-Or	0.182

Scores in bold indicate the best correlations obtained

(149 segments/system), 4 reference translations and human judgements on fluency per segment. VERTa's performance was compared to well-known metrics such as BLEU, the METEOR family and some of the linguistically-based metrics available in the Asiya framework. Most of these metrics have been already described in Sect. 4.1, whereas others have been added because they are more fluency-oriented. These are:

- DP-HWCM_c-4 and DP-HWCM_r-4 metrics, variants of Liu and Gildea (2005)'s HWCM metric, which consider different head-word chain types. DP-HWCM_c-4 considers syntactic categories whereas DP-HWCM_r-4 considers syntactic relations and both of them calculate the average accumulated proportion of category/relation chains up to length 4;
- Confidence Estimation (CE) measures (Specia et al. 2010), also available in the Asiya framework, which are suitable to check the fluency of a segment. CE measures do not need reference translations, they can be target-based (just focusing on target segments) or source/target-based (using both source and target sentences). From those CE measures available in Asiya we selected three target-based measures since their hypothesis is that the likelier the sentence (according to a language model), the more fluent it is. Hence they are suitable in order to check the fluency of a segment. These three measures are: CE-ippl, CE-ippl-c and CE-ippl-p. CE-ippl calculates the inverse perplexity of the target

segment according to a pre-defined language model. CE-ippl-c metric combines the use of a language model with phrase chunks tags. Finally, CE-ippl-p metric uses a language model calculated over sequences of PoS tags. For further details please refer to Asiya technical manual (González and Giménez 2014).

Table 9 reports the results obtained by VERTa and the selected set of metrics, when comparing their scores to human judgements on fluency by means of Pearson correlation coefficient.

4.2.1 Analysis of the results

VERTa's correlation with human judgements on fluency is worse than the correlation obtained on adequacy. This was not unexpected since similar results were obtained when the experiments were performed (Sect. 3.2). Nonetheless, it must be noticed that VERTa clearly outperforms the metrics it is compared against.

Although BLEU is one of the widest used metrics to evaluate MT quality and has also been claimed to correlate well with human judgements on fluency, it has not proved effective to evaluate fluency with our data. This was somehow anticipated given the strict word order considered by BLEU and its matches. As for metrics in the METEOR family, they got similar results, although METEOR-sy, which covers exact matches, stemming and synonymy relations, obtains the best correlation (0.327). VERTa outperforms both BLEU and METEOR due to the combination of n-grams calculated over PoS instead of lexical items and the use of the LM Module, which seems a more suitable strategy to check the grammaticality of a sentence.

From those metrics using shallow parsing—SP-Oc(*) and SP-Op(*)—the former, which accounts for all successfully translated phrases, achieves good results (0.311). This is due to the fact that the metric checks that all words inside a specific phrase have been translated correctly and, indirectly, it accounts for correct word order inside the phrase. As for metrics using dependency trees information, DPm-Or(*) shows a good correlation (0.357) in line with the Dependency Module in VERTa, and especially with the assignment of specific weights to dependency relations occupying different positions in the parsing tree. On the other hand, although Liu and Gildea (2005) claimed that their HWCM metric achieved good results as regards fluency, that is not the case with the two variants tested DP-HWCM_c-4 (0.250) and DP-HWCM_r-4 (0.248), which VERTa clearly outperforms.

As for metrics working at constituent level, the CP-Oc(*) metric obtains the best correlation from all metrics used (0.368), except for VERTa. Without doubt, the use of syntactic information on constituents, namely lexical overlap according to the phrase constituent, proves effective to evaluate the fluency of a segment; thus, highlighting again the importance of word order, not only in phrase chunks but most importantly inside phrase constituents to check the grammaticality of a sentence.

On the other hand, as expected, semantically-based metrics do not achieve good correlations with human judgements on fluency. This is especially remarkable in NE metrics which show a very poor performance (0.080 for Ne-Me(*) and 0.072 for NE-Oe(*)), in line with VERTa's NE components.

Table 9 Pearson correlation for fluency. Comparing VERTa and a selection of well-known metrics

Metric	Pearson correlation
VERTa	**0.455**
CP-Oc(*)	**0.368**
DPm-Or(*)	0.357
METEOR-sy	0.327
METEOR-pa	0.318
CP-Op(*)	0.315
SP-Oc(*)	0.311
METEOR-ex	0.308
METEOR-st	0.307
SR-Or(*)	0.304
BLEU	0.293
SP-Op(*)	0.284
DP-HWCM_c-4	0.250
DP-HWCM_r-4	0.248
DPm-Oc(*)	0.247
DPm-Ol(*)	0.237
SR-Mr(*)	0.237
CE-ippl-p	0.207
CE-ippl	0.193
CE-ippl-c	0.146
NE-Me(*)	0.080
NE-Oe(*)	0.072

Scores in bold indicate the best correlations obtained

A new set of metrics has been used to evaluate fluency, CE metrics, namely CE-ippl, CE-ippl-c and CE-ippl-p. Unfortunately, none of them obtain a good correlation, being CE-ippl-p the best one (0.207), in sharp contrast to VERTa's LM Module. In our metric, the use of a LM proved highly effective to check the grammaticality of a sentence, thus a better performance was expected from CE metrics. Their low correlation migh be due to the LM used, based on the Europarl corpus,[17] a different genre from the newswire corpus used to conduct this evaluation and to the fact that they were used isolated. On the other hand, from this set of metrics, CE-ippl-p obtained the best results. This metric uses an LM calculated over sequences of PoS tags, which strengthens the idea that PoS tags and word order are appropriate to evaluate fluency, and that LM-based measures contribute to the evaluation when combined with other information, as used in VERTa and reported in Sect. 3.2.

To conclude, according to the results obtained, a collaborative approach such as that proposed in VERTa, which combines information on dependency relations, PoS tags and word order, is the most appropriate to evaluate the grammaticality of a sentence. The combination of different linguistic features, once again, outperforms single metrics.

[17] http://www.quest.dcs.shef.ac.uk/quest_files/lm.europarl-nc.en.

5 Conclusions and future work

In the present study we have presented VERTa, an MT evaluation metric based on linguistic information, which has been useful to check the suitability of the linguistic features selected and how they should interact to better measure adequacy and fluency in English.

Several experiments were conducted on a per-module basis and also in a combinatory fashion until we found out which linguistic features should be employed and how they should be used. The resulting features and their combinations go beyond a quantitative analysis and head towards a more qualitative approach, thus moving away from combining a wide range of metrics, which makes it difficult to check their contribution to the analysis, and from using machine learning techniques that require a large amount of data. Our analysis to identify and select the linguistic information and how it should be combined has linked traditional correlations with human judgements with a linguistic analysis of the data every time a new linguistic feature was added. The use of correlations has been useful as a point of departure for our analysis, to refine weights and guide our understanding of the modules in VERTa and their interaction.

The linguistic combination to evaluate adequacy that we propose involves mainly information at lexical level (i.e. word-form, synonyms, hypernyms, hyponyms, lemma and partial lemma) and at syntactic level (i.e. dependency relations). Besides, in a lower degree it also requires word-order features (i.e. n-grams) and other semantically related features (NER, NEL, Time Expressions and Sentiment analysis). The translation of these features into VERTa's modules is the use of the Lexical, Dependency, N-gram and Semantic Modules.

As regards fluency, the linguistic combination involves using mainly information regarding dependency relations, word order and PoS features. Actually, it must be highlighted that at this point linguistic features have interacted with an LM, thus combining a reference-based approach with a target-based approach. As regards VERTa's modules, this information corresponds to an important contribution of the Dependency, LM and N-gram Modules and a minor use of the Morphological Module.

During the experiments, state-of-the-art linguistic features were revisited and we found out that dependency relations, traditionally more fluency-oriented, can also be used to evaluate adequacy, achieving very good results, indeed. In addition, we have also tested the use of unfrequently used features related to textual entailment: NE linking, Time Expressions identification and Sentiment analysis. Although their individual use does not help in the evaluation of MT output, we have proved that the interaction of NER, NEL, Time Expressions and Sentiment analysis is effective to evaluate adequacy in combination with other adequacy-oriented linguistic features. On the other hand, we have tested linguistic features that had not been used before. From these features, our experiments indicate that the use of hypernymy and hyponymy relations should not be entirely disregarded in MT evaluation.

In the evaluation performed the results obtained were highly satisfactory since VERTa outperformed the other metrics in both adequacy and evaluation. This validates the linguistic analysis performed and the features selected to develop VERTa. VERTa has also proved suitable for the ranking of sentences (Comelles and Atserias 2014, 2015). However, its performance in that task was not as outstanding as in the evaluation described on this paper. This is mainly due to the following: (a) the metric was not designed for the ranking of sentences but for the evaluation of adequacy and fluency separately; (b) the combination of modules in VERTa was just roughly adapted to the new task; and (c) no linguistic analysis was performed. From a linguistic point of view, identifying, selecting and combining the linguistic information suitable to rank sentences requires a thorough analysis not only of the data but also of the criteria used by the evaluators when facing the task of ranking.

The work presented in this study is just a small step towards the qualitative analysis of linguistic features in MT evaluation. Actually, there is still a long way to go in order to improve MT metrics from a qualitative perspective. Here we offer a summary of those lines we would like to study further.

Firstly, since our work has been partly inspired by Giménez and Márquez (2010a, b), who used correlations with human judgements and different datasets to find the best combination of linguistic features to evaluate MT quality, we would like to widen the scope of our qualitative analysis by using a larger amount of data and including different datasets. The use of a larger amount of data would allow us to reach more conclusive weights, especially as regards intra-module settings.

Secondly, some of the features used in the Semantic Module, mainly NEs and Time Expressions, are aimed at matching expressions that contain the same meaning but differ in their form. We think that the NEL and Time Expressions metrics could be used in a pre-process stage to identify these expressions conveying the same meaning but differing in their form and substitute them for a normalized form. This normalization will probably help the NLP tools used for parsing both hypothesis and reference segments, thus probably resulting in a better performance of the metric.

Finally, during our experiments we found that the mistakes made by NLP tools inevitably affect the performance of our metric. Some of the errors that we have already detected are usually caused by the PoS tagger, especially when the hypothesis segment is analysed. The most common errors are the misanalyses of verbs as nouns and vice versa, and not distinguishing proper nouns from uppercase common nouns. The former is very common when part of a segment has not been translated, whereas the latter tends to occur in headlines. Since PoS tagging is the first step in the parsing process, these mistakes are propagated through the parsing chain affecting the metric's performance. Thus, we are particularly interested in detecting parser errors and exploring the impact that they have on VERTa.

On a different note, we would also like to explore the application of VERTa to another NLP task: recognizing textual entailment (RTE). Actually, textual entailment (TE) has been used in some MT metrics (Padó et al. 2009; Castillo and Estrella 2012) since somehow, RTE and evaluation of MT using references (at least when evaluating adequacy) are not that far, as both of them compare a hypothesis and reference segment and try to find out if they are semantically similar. We would like to check if VERTa can also be useful in this NLP task.

Acknowledgements We would like to thank LDC and NIST for kindly providing the data used in this study. This work has been funded by the Spanish Government (Project TUNER, TIN2015-65308- C5-1-R, MINECO/FEDER, UE).

References

Agarwal, A., & Lavie, A. (2008). METEOR, M-BLEU and M-TER: Flexible Matching and Parameter Tuning for High-Correlation with Human Judgments of Machine Translation Quality. In *Proceedings of the ACL2008 Workshop on Statistical Machine Translation*. Columbus, Ohio, USA.

Albrecht, J. S., & Hwa, R. (2007a). A re-examination of machine learning approaches for sentence-level MT evaluation. In *Proceedings of the 45th annual meeting of the association for computational linguistics (ACL), Prague, Czech Republic* (pp. 880–887).

Albrecht, J. S., & Hwa, R. (2007b) Regression for sentence-level MT evaluation with pseudo references. In *Proceedings of the 45th annual meeting of the association for computational linguistics, Prague, Czech Republic* (pp. 296–303).

Atserias, J., Blanco, R., Chenlo, J. M., & Rodriguez, C. (2012). *FBM-Yahoo at RepLab 2012*. CLEF (Online Working Notes/Labs/Workshop).

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of ACL workshop on intrinsic and extrinsic evaluation measures for MT and/or summarization, Michigan, USA*.

Callison-Burch, Ch., Koehn, P., Monz, Ch., Post, M., Soricut, R., & Specia, L. (2012). Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the seventh workshop on statistical machine translation, Montréal, Canada* (pp. 10–51).

Castillo, J., & Estrella, P. (2012). Semantic textual similarity for MT evaluation. In *Proceedings of the seventh workshop on statistical machine translation, Montréal, Canada* (pp. 52–58).

Chang, A. X., & Manning, Ch. D. (2012). SUTIME: A library for recognizing and normalizing time expressions. In *Proceedings of the 8th international conference on language resources and evaluation, Istanbul, Turkey*.

Chang, Y. S., & Ng, H. T. (2008). MAXSIM: A maximum similarity metric for machine translation evaluation. In *Proceedings of the ACL-08: HLT, Columbus, Ohio, USA* (pp. 55–62).

Charniak, E., & Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL), Michigan, USA*.

Chen, B., Kuhn, R., & Foster, G. (2012). Improving AMBER, an MT evaluation metric. In *Proceedings of the 7th workshop on statistical machine translation, Montréal, Canada* (pp. 59–63).

Ciaramita, M., & Altun, Y. (2006). Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*.

Comelles, E., Arranz, V., & Castellon, I. (2010). Constituency and dependency parsers evaluation. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural, 45,* 59–66.

Comelles, E., Arranz, V., & Castellon, I. (2017). Guiding automatic MT evaluation by means of linguistic features. *Digital Scholarship in the Humanities, 32*(4), 761–778.

Comelles, E., & Atserias, J. (2014). VERTa participation in the WMT14 metrics task. In *Proceedings of the ninth workshop on statistical machine translation, Baltimore, USA*.

Comelles, E., & Atserias, J. (2015). VERTa: A linguistically-motivated metric at the WMT15 metrics task. In *Proceedings of the tenth workshop on statistical machine translation, Lisbon, Portugal*.

Comelles, E., & Atserias, J. (2016). Through the eyes of VERTa. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural, 57,* 181–184.

De Marneffe, M.C., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th edition of the international conference on language resources and evaluation (LREC-2006), Genoa, Italy*.

Denkowski, M., & Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the 6th workshop on statistical machine translation, Edinburgh, Scotland, UK* (pp. 85–91).

Denkowski, M., & Lavie, A. (2014). Meteor universal. Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation, Baltimore, Maryland, USA* (pp. 376–380).

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd international conference on human language technology, San Diego, California* (pp. 138–145).

Farrús, M., Costa-Jussà, M. R., Mariño, J. B., & Fonollosa, J. A. R. (2010). Linguistic-based evaluation criteria to identify statistical machine translation errors. In *Proceedings of the 14th annual conference of the European association for machine translation, Saint Raphael, France*.

Fellbaum, Ch. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Gautam, S., & Bhattacharyya, P. (2014). Layered: Metric for machine translation evaluation. In *Proceedings of the ninth workshop on statistical machine translation, Baltimore, Maryland, USA* (pp. 387–393).

Giménez, J. (2008). *Empirical machine translation and its evaluation*. PhD thesis, Universitat Politècnica de Catalunya, Spain.

Giménez, J., & Márquez, L. (2004). SVM tool: A general POS tagger generator based on support vector machines. In *Proceedings of the 4th international conference on language resources and evaluation (LREC'04), Lisbon, Portugal* (pp. 43–46).

Giménez, J., & Márquez, Ll. (2008). Discriminative phrase selection for statistical machine translation. In C. Goutte, N. Cancedda, M. Dymetman, & G. Foster (Eds.), *Learning machine translation. NIPS workshop series*. Cambridge: MIT Press.

Giménez, J., & Márquez, Ll. (2010a). Asiya: An open toolkit for automatic machine translation (meta-) evaluation. *The Prague Bulletin of Mathematical Linguistics, 94,* 77–86.

Giménez, J., & Márquez, Ll. (2010b). Linguistic measures for automatic machine translation evaluation. *Machine Translation, 24*(3–4), 77–86.

González, M., Barrón-Cedeño, A., & Márquez, L. L. (2014). IPA and STOUT: Leveraging linguistic and source-based features for machine translation evaluation. In *Proceedings of the ninth workshop on statistical machine translation, Baltimore, Maryland, USA* (pp. 394–401).

González, M., & Giménez, J. (2014). *Asiya: An open toolkit for automatic machine translation (meta-) evaluation. Technical manual 3.0*. Barcelona: Universitat Politècnica de Catalunya.

Gupta, R., Orăsan, C., & van Genabith, J. (2015). ReVal: A simple and effective machine translation evaluation metric based on recurrent neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing, Lisbon, Portugal* (pp. 1066–1072).

Hachey, B., Radford, W., & Curran, J. R. (2011). Graph-based named entity linking with Wikipedia. In *Proceedings of the 12th international conference on web information system engineering* (pp. 213–226).

He, Y., Du, J., Way, A., & van Genabith, J. (2010). The DCU dependency-based metric in WMT-MetricsMATR 2010. In *Proceedings of the 5th workshop on statistical machine translation, Uppsala, Sweeden* (pp. 349–353).

Hutchins, J. W., & Somers, H. L. (1992). *An introduction to machine translation*. London: Academic Press.

Joty, S., Guzmán, F., Márquez, L., & Nakov, P. (2014). DiscoTK: Using discourse structure for machine translation evaluation. In *Proceedings of the ninth workshop on statistical machine translation, Baltimore, Maryland, USA* (pp. 402–408).

Liu, D., & Gildea, D. (2005). Syntactic features for evaluation of machine translation. *Proceedings of ACL workshop on intrinsic and extrinsic evaluation measures for MT and/or summarization* (pp. 25–32).

Lo, Ch. (2017). Meant 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the second conference on machine translation, volume 2: Shared tasks papers, Copenhagen, Denmark* (pp. 589–597).

Lommel, A. (2016). Blues for BLEU: Reconsidering the validity of reference-based MT evaluation. In *Proceedings of the LREC 2016 workshop "translation evaluation – from fragmented tools and data sets to an integrated ecosystem", Portoroz, Slovenia* (pp 63–70).

Lo, Ch., & Wu, D. (2010). Semantic vs. syntactic vs. N-gram structure for machine translation evaluation. In *Proceedings of the fourth workshop on syntax and structure in statistical translation, Beijing* (pp. 52–60).

Lo, Ch., & Wu, D. (2011). Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, Portland, Oregon, USA* (pp. 220–229).

Lo, Ch., & Wu, D. (2012). Unsupervised vs. supervised weight estimation for semantic MT evaluation metrics. In *The sixth workshop on syntax, semantics and structure in statistical translation (SSST-6), Jeju Island, South Korea.*

Lo, Ch., & Wu, D. (2013). MEANT at WMT2013: A tunable, accurate yet inexpensive semantic frame based MT evaluation metric. In *Proceedings of the eighth workshop on statistical machine translation, Sofia, Bulgaria* (pp. 422–428).

Ma, Q., Graham, Y., Wang, S., & Liu, Q. (2017). Blend: A novel combined MT metric based on direct assessment CASICT-DCU submission to WMT17 metrics task. In *Proceedings of the second conference on machine translation, volume 2: Shared tasks papers, Copenhagen, Denmark* (pp. 598–603).

Macketanz, V., Avramidis, E., Burchardt, A., Helcl, J., & Srivastava, A. (2017). Machine translation: Phrase-based, rule-based and neural approaches with linguistic evaluation. *Cybernetics and Information Technologies, 17*(2), 28–43.

Owczarzak, K., van Genabith, J., & Way, A. (2007a). Dependency-based automatic evaluation for machine translation. In *Proceedings of SSST, NAACL-HLT/AMTA workshop on syntax and structure in statistical translation* (pp. 80–87).

Owczarzak, K., van Genabith, J., & Way, A. (2007b). Labelled dependencies in machine translation evaluation. In *Proceedings of the ACL workshop on statistical machine translation, Czech Republic* (pp. 104–111).

Padó, S., Galley, M., Jurafsky, D., & Manning, Ch D. (2009). Measuring machine translation quality as semantic equivalence: A metric based on entailment features. *Machine Translation, 23*(2–3), 181–193.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). *Bleu: A method for automatic evaluation of machine translation*. RC22176 (technical report). IBM T.J. Watson Research Center.

Pearson, K. (1914, 1924, 1930). *The life, letters and labours of Francis Galton* (3 volumes).

Popović, M. (2012). Class error rates for evaluation of machine translation output. In *Proceedings of the 7th Workshop on Statistical Machine Translation* (pp. 71-75). Montréal, Canda.

Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation, Lisbon, Portugal* (pp. 392–395).

Popović, M. (2017). chrF ++: Words helping character n-grams. In *Proceedings of the second conference on machine translation, Volume 2: Shared tasks papers, Copenhagen, Denmark* (pp. 612–618).

Reeder, F., Miller, K., Doyon, J., & White, J. (2001). The naming of things and the confusion of tongues: An MT metric. In *Proceedings of the workshop on MT evaluation "who did what to whom?" at machine translation summit VIII* (pp. 55–59).

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th conference of the association for machine translation in the Americas (AMTA)* (pp. 223–231).

Snover, M., Madnani, N., Dorr, B., & Schwartz, R. (2009). Fluency, adequacy or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the 4th workshop on statistical machine translation at the 12th meeting of the european chapter of the association for computational linguistics (EACL-2009), Athens, Greece*.

Specia, L., Cancedda, N., Dymetman, M., Turchi, M., & Cristianini, N. (2009). Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th annual conference of the EAMT, Barcelona, Spain* (pp. 28–35).

Specia, L., Hajlaoui, N., Hallet, C., & Aziz, W. (2011). Prediting machine translation adequacy. In *Proceedings of the 13th translation summit, Xiamen, China* (pp. 513–520).

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

Stanojević, M., & Sima'an, K. (2015). BEER 1.1: ILLC UvA submission to metrics and tuning task. In *Proceedings of the tenth workshop on statistical machine translation, Lisbon, Portugal* (pp. 396–401).

Surdeanu, M., & Turmo, J. (2005). Semantic role labeling using complete syntactic analysis. In *Proceedings of CoNLL shared task*.

Toutanova, K., Klein, D., Manning, Dh., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL* (pp. 252–259).

Turian, J. P., Shen, L., & Melamed, I. D. (2003). Evaluation of machine translation and its evaluation. In *Proceedings of MT SUMMIT IX*.

Wang, M., & Manning, Ch. (2012). SPEDE: Probabilistic edit distance metrics for MT evaluation. In *Proceedings of the 7th workshop on statistical machine translation, Montréal, Canada*.

Wang, W., Peter, J., Rosendahl, H., & Ney, H. (2016). CharacTer: Translation edit rate on character level. In *Proceedings of the first conference on machine translation, Berlin, Germany* (pp. 505–510).

White, J. S., O'Connell, T., & O'Mara, F. (1994). The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proceedings of the 1st conference of the association for machine translation in the Americas (AMTA)* (pp. 193–205).

Wu, X., Yu, H., & Liu, Q. (2013). DCU participation in WMT13 metrics task. In *Proceedings of the eighth workshop on statistical machine translation, Sofia, Bulgaria* (pp. 435–439).

Yang, M. Y., Sun, S. Q., Zhu, J. G., Li, S., Zhao, T. J., & Zhu, X. N. (2011). Improvement of machine translation evaluation by simple linguistically motivated features. *Journal of Computer Science and Technology, 26*(1), 57–67.

Ye, Y., Zhou, M., & Lin, Ch. (2007). Sentence level machine translation evaluation as a ranking problem: One step aside from BLEU. In *Proceedings of the second workshop on statistical machine translation, Prague, Czech Republic* (pp. 240–247).

Yu, H., Ma, Q., Wu, X., & Liu, Q. (2015). CASICT-DCU participation in WMT15 metrics task. In *Proceedings of the tenth workshop on statistical machine translation, Lisbon, Portugal* (pp. 417–421).

Zhang, L., Weng, Z., Xiao, W., Wuan, J., Chen, Z., Tan, Y., Liand, Ma., & Wang, M. (2016). Extract domain-specific paraphrase from monoligual corpus for automatic evaluation of machine translation. In *Proceedings of the first conference on machine translation. Volume 2 shared task papers, Berlin, Germany* (pp. 511–517).

Links

https://github.com/jatserias/VERTa. Accessed September 9, 2018.
http://www.statmt.org/wmt10/evaluation-task.html. Accessed September 9, 2018.
http://www.itl.nist.gov/iad/mig/tests/mt/2006/. Accessed September 9, 2018.
http://www.statmt.org/wmt13/quality-estimation-task.html. Accessed September 9, 2018.
https://www.nist.gov/. Accessed September 9, 2018.
https://www.ldc.upenn.edu/. Accessed September 9, 2018.
https://catalog.ldc.upenn.edu/LDC2010T14. Accessed September 9, 2018.
http://asiya.lsi.upc.edu/. Accessed September 9, 2018.
http://www.quest.dcs.shef.ac.uk/quest_files/lm.europarl-nc.en. Accessed September 9, 2018.