Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 181-184 recibido 29-03-2016 revisado 11-04-2016 aceptado 22-04-2016

Through the Eyes of VERTa

A Través de los Ojos de VERTa

Elisabet Comelles
Universitat de Barcelona (UB)
Gran Via de les Corts Catalanes, 585, 08007,
Barcelona (Spain)
elicomelles@ub.edu

Jordi Atserias
IXA Group
University of the Basque Country
(UPV/EHU)
Spain
jordi_atserias001@ehu.eus

Abstract: This paper describes a practical demo of VERTa for Spanish. VERTa is an MT evaluation metric that combines linguistic features at different levels. VERTa has been developed for English and Spanish but can be easily adapted to other languages. VERTa can be used to evaluate adequacy, fluency and ranking of sentences. In this paper, VERTa's modules are described briefly, as well as its graphical interface which provides information on VERTa's performance and possible MT errors.
Keywords: Machine translation evaluation, Automatic metric, Demo, Spanish, Linguistic knowledge, Error analysis

Resumen: Este artículo describe la demostración práctica de VERTa para el castellano. VERTa es una métrica de evaluación de traducción automática que combina información lingüística a diferentes niveles. VERTa ha sido desarrollada para el inglés y el castellano pero se puede adaptar fácilmente a otras lenguas. La métrica puede evaluar la adecuación, la fluidez y ranking de frases. En este artículo se describen brevemente los módulos de VERTa y su interficie gráfica, la cual proporciona información sobre el rendimiento de la métrica y posibles errores de traducción.
Palabras clave: Evaluación de la traducción automática, Métrica automática, Demo, Castellano, Conocimiento lingüístico, Errores de traducción

1 Introduction

Automatic Machine Translation (MT) Evaluation has become a key field in Natural Language Processing due to the amount of texts that are translated over the world and the need for a quick, reliable and inexpensive way to evaluate the quality of the output text. Therefore, a large number of metrics have been developed, which range from very simple metrics, such as BLEU to more complex ones, which involve combining a wide variety of linguistic features using machine-learning techniques (Gautam and Bhattacharyya, 2014; Joty et al., 2014; Yu et al., 2015) or in a more simple and straightforward way (Giménez and Márquez, 2010; González et al., 2014). Nevertheless, little research has been carried out in order to explore the suitability of the linguistic features used and how they should be combined, from a linguistic point of view. In order to address this issue, VERTa, a linguistically-motivated metric (Comelles and Atserias, 2015), has been developed. This metric uses a wide variety of linguistic features at different levels and aims at moving away from a biased evaluation, providing a more holistic approach to MT evaluation.

This paper reports a demo of VERTa and aims at exploring the results provided by the metric and its potential use as an error analysis tool. Therefore, we provide a brief description of the different modules in the Spanish version of VERTa and how the results and the information in these modules can be visualized to better understand the metric's performance and to help developers carry out error analysis. VERTa is available at http://grial.ub.edu:8080/VERTaDemo/.

ISSN 1135-5948 © 2016 Sociedad Española para el Procesamiento del Lenguaje Natural

2 VERTa

VERTa claims to be a linguistically-motivated metric because before its development a thorough analysis was carried out in order to identify which linguistic phenomena an MT evaluation metric should take into account when evaluating MT output by means of reference translations. With the results of this analysis (Comelles, 2015) we decided on the linguistic features that would be more appropriate and on how they should be combined depending on whether Adequacy or Fluency was evaluated. Therefore, VERTa consists of six modules which can work independently or in combination: *Lexical Similarity Module (L)*, *Morphological Similarity Module (M)*, *N-gram Similarity Module (N)*, *Dependency Similarity Module (D)*, *Semantic Similarity Module (S)* and *Language Model (LM) Module*[1].

All metrics use a weighted precision and recall over the number of matches of the particular element of each level (words, dependency triples, n-grams, etc.).

Next, all modules forming VERTa are described.

2.1 Lexical Similarity Module

The Lexical Similarity Module captures similarities between lexical items in the hypothesis and reference sentences. This module does not only use superficial information such as the wordform, but it also takes into account lemmatization, lexical semantics (i.e. synonymy, hypernymy and hyponymy), and partial lemma. In addition, different weights are assigned depending on their importance as regard semantics and/or fluency.

2.2 Morphological Similarity Module

This module uses the information provided by the Lexical Module in combination with Part-of-Speech (PoS) tags[2].

Similar to the Lexical Similarity Module, this module matches items in the hypothesis and reference segments and a set of weights is assigned to each type of match.

This module aims at making up for the broader coverage of the Lexical Module, thus preventing matches such as *invites* and *invite*, which although similar in meaning differ in their morphosyntactic features.

2.3 Dependency Similarity Module

The Dependency Module captures similarities beyond the external structure of a sentence and uses dependency structures to link syntax and semantics. Thus, this module allows for identifying sentences with the same meaning but different syntactic constructions (e.g. active – passive alternations), as well as changes in word order.

This module works at sentence level and follows the approach used by Owczarzack et al. (2007) and He et al. (2010) with some linguistic additions in order to adapt it to our metric combination. Similar to the Morphological Module, the Dependency Similarity metric also relies first on those matches established at lexical level − word-form, synonymy, hypernymy, hyponymy and lemma − in order to capture lexical variation across dependencies and avoid relying only on surface word-form. Then, by means of flat triples with the form Label (Head, Mod) obtained from the parser[3], four different types of dependency matches are designed (i.e. complete, partial-no-label, partial-no-head, partial-no-mod) and weights are assigned to each type of match.

In addition, VERTa also enables the user to assign different weights to the dependency categories according to the type of evaluation performed.

Finally, a set of language-dependent rules has been implemented in order to a) widen the range of syntactically-different but semantically-equivalent expressions, and b) restrict certain dependency relations (e.g. subject, object).

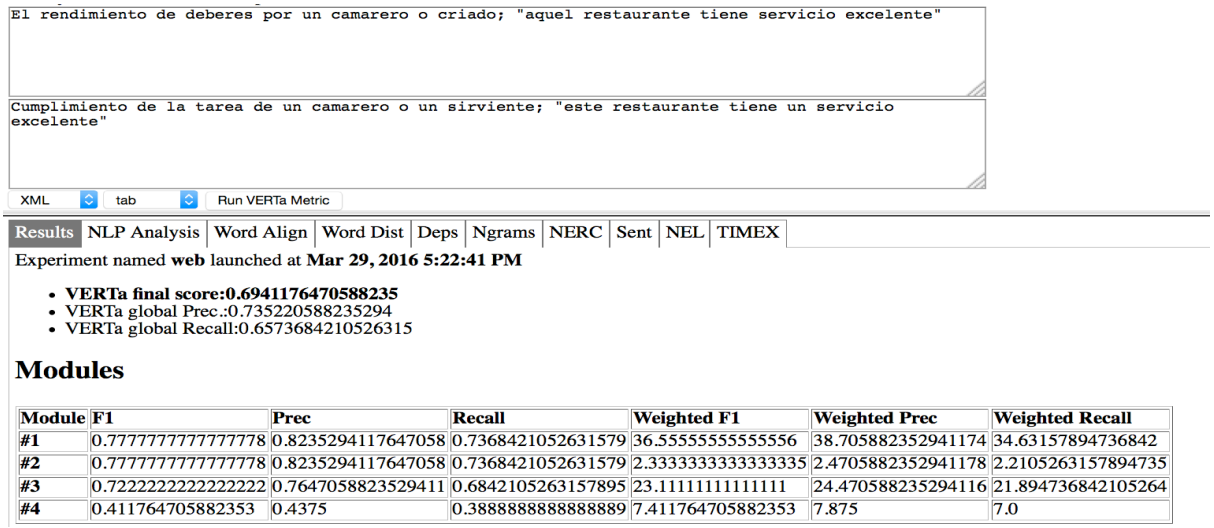
2.4 N-gram Similarity Module

This module matches chunks in the hypothesis and reference segments. N-grams can be calculated over lexical items (considering the information provided by the Lexical Module),

[1] Neither the Semantic Similarity Module nor the Language Model Module are available in the Spanish version of the metric.

[2] The text is tagged using Freeling (Padró and Stanilovsky, 2012).

[3] Both hypothesis and reference strings are annotated with dependency relations by means of Freeling dependency parsing (Lloberes et al., 2010).



Module	F1	Prec	Recall	Weighted F1	Weighted Prec	Weighted Recall
#1	0.7777777777777778	0.8235294117647058	0.7368421052631579	36.55555555555556	38.705882352941174	34.63157894736842
#2	0.7777777777777778	0.8235294117647058	0.7368421052631579	2.3333333333333335	2.4705882352941178	2.2105263157894735
#3	0.7222222222222222	0.7647058823529411	0.6842105263157895	23.11111111111111	24.470588235294116	21.894736842105264
#4	0.411764705882353	0.4375	0.3888888888888889	7.411764705882353	7.875	7.0

Figure 1: VERTa's home page, global score and scores per module

over PoS and over the combination of lexical items and PoS. The n-gram length can go from bigrams to sentence-length grams. This module is particularly useful when evaluating Fluency because it deals with word order.

3 VERTa GUI: a Graphical Interface

VERTa GUI allows users to visualize the similarity between two segments module by module. The reference and hypothesis segments are entered in different text boxes and VERTa GUI does not only return the global score but also the score per module (see Figure 1).

In addition, this visual interface also allows users to navigate the different modules in VERTa, by means of a set of tabs. By clicking on each tab (see Figure 2), users are taken to the corresponding module: Word align and word distance (Lexical and Morphological Modules), Deps (Dependency Module), Ngrams (N-gram Module), and NERC, Sent, NEL and TIMEX, corresponding to the Semantic Module.

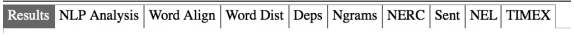


Figure 2: Tabs taking to each module in VERTa

The information contained in these tabs is not only useful in order to check the metric's performance, but also in order to identify possible MT errors.

3.1 A Practical Case

Although VERTa was initially developed to evaluate English a new version has been developed for Spanish. This version uses all modules in VERTa except for the Semantic Similarity Modules and the Language Module Modules. In the demo reported in this paper, VERTa is used to evaluate adequacy. To this aim, the combination of modules is the following:

- Lexical Module: 0.46
- PoS Module: 0.03
- Dependency Module: 0.32
- Ngram Module: 0.19

The metric can account for the semantic similarity between two sentences such as those in example 1 where the hypothesis segment conveys the meaning of the source segment, despite not being very natural.

SOURCE: *the performance of duties by a waiter or servant; "that restaurant has excellent service "*

HYP: *El rendimiento de deberes por un camarero o criado; "aquel restaurante tiene servicio excelente".*

REF: *Cumplimiento de la tarea de un camarero o un sirviente; "este restaurante tiene un servicio excelente"*

As shown in Figure 3, synonymy helps in matching *deberes* ("duties") and *tareas* ("tasks"), as well as *criado* ("manservant") and *sirviente* ("servant") in the Lexical Module. In addition, possible errors can also be identified by the elements not matched (coloured in red), such as the lack of determiners preceding *deberes* and *servicio* in the hypothesis segment.

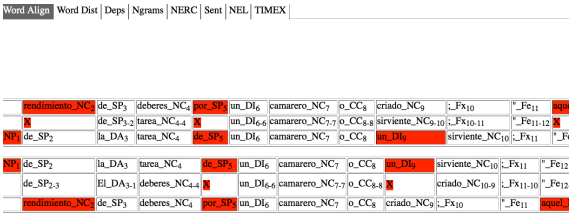


Figure 3: Example of matches in the lexical module

In addition, the n-gram module matches chunks in the hypothesis segments to those in the reference segments. Those chunks that can be matched are highlighted in green.

The demo can be accessed at http://grial.ub.edu:8080/VERTaDemo/.

4 Conclusions and Future Work

This paper has described the Spanish version of VERTa, a linguistically-motivated MT metric, and its graphical interface. The architecture of the metric and the modules in the Spanish version have been described. In addition, its graphical interface, VERTa GUI has been presented in order to show the metric's performance and the usefulness of the information provided by VERTa's modules.

In the future we're planning to add the Semantic Similarity Module and the Language Model Module to the Spanish version. In addition, we are also considering a better way to extract and display information on error analysis.

Acknowledgements

This work has been funded by the Spanish Government (project TUNER, TIN2015-65308-C5-1-R)

References

Comelles, E. and J. Atserias. 2015. VERTa: A Linguistically-Motivated Metric at the WMT15 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. EMNLP, pp. 366-372, Lisbon.

Comelles, E. 2015. *Automatic Machine Translation Evaluation: A Qualitative Approach*, University of Barcelona, Barcelona.

Gautam, S. and P. Bhattacharyya. 2014. LAYERED: Metric for Machine Translation Evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation,* ACL-2014, pp. 387-393, Baltimore.

Giménez, J. and Ll. Márquez. 2010. Linguistic Measures for Automatic Machine Translation Evaluation. *Machine Translation*, 24(3–4):77–86.

González, M., A. Barrón-Cedeño, and Ll. Márquez. 2014. IPA and STOUT: Leveraging Linguistic and Source-based Features for Machine Translation Evaluation. En *Proceedings of the Ninth Workshop on Statistical Machine Translation,* ACL-2014, pp. 394-401, Baltimore.

He, Y., J. Du, A. Way and J. van Genabith. 2010. The DCU Dependency-based Metric in WMT-Metrics MATR 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics* MATR, WMT 2010, pp. 349-353, Uppsala.

Joty, S., F. Guzmán, Ll. Márquez and P. Nakov. 2014. DiscoTK: Using Discourse Structure for Machine Translation Evaluation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation,* ACL-2014, pp. 402-408, Baltimore.

Lloberes, M., I. Castellón and Ll. Padró. 2010. Spanish FreeLing Dependency Grammar. In *Proceedings of the Seventh conference on International Language Resources and Evaluation,* LREC'10, pp. 693-699, Malta.

Owczarzak, K., J. van Genabith and A. Way. 2007. Labelled Dependencies in Machine Translation Evaluation. In *Proceedings of the ACL Workshop on Statistical Machine Translation,* ACL, pp. 104-111. Prague.

Padró, Ll. and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference,* LREC 2012, pp.2473-2479. Istanbul.

Yu, H., Q. Ma, X. Wu and W. Liu. 2015. CASICT-DCU Participation in WMT2015 Metrics Task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, EMNLP, pp. 417-421, Lisbon.