METEOR for Multiple Target Languages using DBnary.

Zied Elloumi zied.elloumi@imag.fr
Hervé Blanchon herve.blanchon@imag.fr
Gilles Serasset gilles.serasset@imag.fr
Laurent Besacier laurent.besacier@imag.fr
LIG - GETALP - Univ. Grenoble Alpes

Abstract

This paper proposes an extension of METEOR, a well-known MT evaluation metric, for multiple target languages using an in-house lexical resource called DBnary (an extraction from Wiktionary provided to the community as a Multilingual Lexical Linked Open Data). Today, the use of the synonymy module of METEOR is only exploited when English is the target language (use of WordNet). A synonymy module using DBnary would allow its use for the 21 languages (covered up to now) as target languages. The code of this new instance of METEOR, adapted to several target languages, is provided to the community. We also show that our *DBnary augmented* METEOR increases the correlation with human judgements on the WMT 2013 and 2014 *metrics* dataset for English-to-(French, Russian, German, Spanish) language pairs.

1. Introduction

Machine translation (MT) is the process of automatically translating a text in a source language into a corresponding text in a target language. In order to evaluate and compare the quality of several MT systems, we need to rate the translation hypothesis produced by each MT system either with the help of human experts (subjective evaluation) or compare it to pre-existing human translations (using automatic evaluation metrics, objective evaluation). In practice, subjective evaluation considers various aspects to grade the translation quality, such as adequacy, fluency, and intelligibility. However, subjective evaluation conducted *a posteriori* often costs too much (in term of human resources) and, thus, objective evaluation metrics (fast and cheap as long as references are available) are often preferred nowadays.

One drawback with automatic evaluation metrics is that they compare the MT hypothesis with few (and sometimes only one) reference translations. This is definitely not enough to capture lexical variation in translation. For this reason, metrics which exploit synonymy or stem similarities, such as METEOR (Banerjee and Lavie, 2005), exhibit better correlation with human judgement. METEOR maps words with the same stem or the same synset using lexico-semantic resources. However, so far, the full potential of METEOR is only exploited when English is the target language (use of WordNet).

Contribution This paper proposes an extension of METEOR for multiple target languages using a lexical resource called DBnary (Sérasset, 2015). DBnary is an extraction in RDF of the lexical data of multiple editions of Wiktionary. It has several millions of triples describing lexical entries of the extracted languages, and more than 4.6 million translations from 21 languages to more than 1500 target languages. The modified code allowing to call METEOR for new target languages (French, Russian, German, Spanish) is made available to the research community. More target languages (today 21 in total) could be plugged very quickly by interested users using the same lexical resource (DBnary notably includes Bulgarian, Dutch, English, Finnish, French, German, (Modern) Greek, Indonesian, Italian, Japanese, Latin, Lithuanian, Malagasy, Norwegian, Polish, Portuguese, Russian, Serbo-Croat, Spanish, Swedish, Turkish). We also present initial experiments on the WMT 2013 and 2014 *metrics* dataset and show that our new METEOR slightly increases correlation with human judgments of translation quality, for language pairs with a target language different than English.

2. State of the art

Since METEOR was first introduced in 2005, it has been improved and extended to include more features and accommodate more languages for a subset of its features.

2.1. METEOR: the basics

Banerjee and Lavie (2005) introduced METEOR to overcome several weakness of BLEU (Papineni, 2002) and NIST (Doddington, 2002) they identified as: the lack of recall, an indirect only measure of level of grammatical wellformedness, the lack of explicit word-matching between translation and reference, and the use of geometric averaging of n-grams.

The goal of METEOR was to aim for better correlation with human judgments of translation quality using not only word-to-word alignment between the translation hypothesis and the reference translation(s). The alignment is incrementally produced by a three-leveled mapping approach between the hypothesis and the reference, using additional resources if needed: exact match of the surface forms of the words, exact match of the computed stems of the words, synonymy overlap through shared WordNet "synset" of the words. The second mapping level uses a stemmer and the third mapping level uses Enlgish WordNet.

While no free WordNets are available for languages such as French, Spanish or German, current implementation of METEOR for such languages do not support the third mapping level.

2.2. METEOR: the recent extensions

METEOR-NEXT (Denkowski and Lavie, 2010a), was introduced to better correlate with human-targeted Translation Edit Rate (HTER) (Snover et al., 2006), a semi-automatic post-editing based metric which measures the distance between a MT hypothesis and its post-edited version. The goal was to go beyond a strictly world-level metric with a new aligner supporting phrases (multi-word) matches. Thus, a fourth mapping level was added to implement this new feature using a paraphrase database. For English, the database was developed by Snover (2009a). Later, Denkowski and Lavie (2010b), released paraphrase databases for Czech, German, Spanish and French.

In 2014, *METEOR Universal* was released (Denkowski and Lavie 2014) that enabled the construction of the paraphrase database using only the parallel corpora used to develop the MT system (which was not the case in 2010).

In order to prevent synonyms/paraphrases corresponding to different senses to be treated as semantically equivalent, Apidianaki and Marie (2015) proposed *METEOR-WSD*. The English references are further disambiguated and annotated using *Babelfly* (Moro et al., 2014) for several language pairs (French, Hindi, German, Czech and Russian to English). For their experiment, Apidianaki and Marie (2015) got a better segment-level Kendall's τ correlation than METEOR for 4 language pairs when the paraphrase module was activated.

2.3. Lexical resources

2.3.1. WordNet

WordNet is a large lexical database for English, developed by linguists of Princeton University (Fellbaum, 1998). Nowadays, it has become an important and a very useful resource for NLP applications, such as machine translation, word sense disambiguation, cross-lingual information retrieval etc. WordNet links nouns, verbs, adjectives and adverbs to sets of cognitive synonyms (called synsets), where each synset represent a specific concept. Synsets are interconnected through conceptual semantic and lexical relations, including synonymy, antonymy, hyponymy etc. Note that words with multiple meanings belong to several synsets, and their senses are arranged by order of frequency. There are different versions of WordNet in languages other than English, such as Arabic WordNet, French WordNet, etc. However, these lexical resources in other languages are not freely available. As already said, METEOR uses WordNet to increase the chance of the MT output words to match the reference words.

The latest version of WordNet 3.0, contains in total 117 659 synsets: 82 115 noun synsets, 13 767 verb synsets, 18 156 adjective synsets and 3621 adverb synsets.

To Lemmatize forms, METEOR uses the *Morphy-7WN*[1] function included in WordNet. This function uses a two-step process to find lemma of a particular word *W*. Firstly, *Morphy* checks for exceptions in a list (containing morphological transformations that are not regular). If *W* is not in exception list, Morphy uses the rules of detachment for NOUN, VERB and ADJ categories (no rules applied to ADV). After each transformation, WordNet is searched for the resulting string in the syntactic category specified.

2.3.2. DBnary

DBnary is a multilingual lexical resource in RDF (Klyne & Carroll, 2004) collected at LIG (Sérasset, 2015). The lexical data is represented using standard vocabularies. The lexicon structure is defined using the LEMON vocabulary (McCrae et al., 2011). Most parts of speech informations are mapped to the *Lexinfo* or *OliA* standard vocabularies (Cimiano et al. 2011, Hellman et al. 2015), making it highly reusable in many contexts. It is available either as a set of downloadable files or as Linked Open Data directly accessible to browsers or applications. It may also be queried online using a public SPARQL endpoint.

The available lexical data is automatically extracted from 21 different language editions[2] of Wiktionary, the dictionary counterpart of Wikipedia. Among available lexical data, one may find 2.9M *lexical entries* (with parts-of-speech, canonical form for all of them, along with pronunciations when available and inflected forms for some languages). *Lexical entries* are subdivided into 2.5M *lexical senses* (with their definitions and some usage example).

1 MORPHY(7WN) manual page : https://wordnet.princeton.edu/man/morphy.7WN.html

2 Bulgarian, Dutch, English, Finnish, French, German, (Modern) Greek, Indonesian, Italian, Japanese, Latin, Lithuanian, Malagasy, Norwegian, Polish, Portuguese, Russian, Serbo Croat, Spanish, Swedish, Turkish

DBnary also contains more than 4.6M translations going from the 21 extracted sources languages to more than 1500 different target languages. Additionaly, DBnary contains lexico-semantic relations (syno/antonymy, hypo/hypernymy and mero/holonymy and troponymy).

Table 1 shows the size of the data for languages involved in the experiments later reported in this paper. Additional figures are available on the DBnary public web site[3].

	English	**French**	**Russian**	**German**	**Spanish**
# of entries	620,369	322,018	185,910	104,505	86,388
# of senses	498,451	416,323	176,335	116,290	126,411
# of synonyms	35,437	36,019	31,345	33,282	21,024

Table 1. Number of entries, senses, and lexico-semantic relations in *DBnary* for the target languages considered in this study.

3. METEOR-DBnary for multiple target languages

The principal goal of this study is to propose an evaluation metric that uses synonyms in order to improve MT evaluation for target languages other than English.

3.1. Resources prepared

As mentioned in the section 1, *METEOR* package uses the synset dictionary of WordNet, which is a rich resource of 147 306 unique synsets belonging to four categories (nouns, verbs, adjectives and adverbs) for English. To gather new external resources for our *augmented METEOR*, we downloaded and installed Dbnary dataset[4] and set up a virtuoso-opensource[5] server in order to interrogate Dbnary locally. Then, we launched SPARQL queries on DBnary in order to extract every synonymy relations in the database for English, French, Russian, German and Spanish. The result of the extraction is in the format of: *lemma -> synonym*. Next, we performed a processing to match the same format as WordNet (which is already compatible with METEOR). This treatment is to assign an ID for each lemma and to build a list of synonyms for each lemma under format ID eg *lemma -> ID_syn1 ID_Syn2 ID_Syn3*.

METEOR computes its scores using WordNet as an external lexical resource. In order to measure the difference between the use of DBnary synonyms and the use of WordNet synonyms, we built two English synonym dictionaries extracted automatically from Dbnary: one that contains the same four categories available in WordNet (called *METEOR-DB-4-catg*) and another containing all the existing categories in DBnary for English (called *METEOR-DB-All-catg*).

3 data, docs and examples are available at http://kaiko.getalp.org/about-dbnary/

4 http://kaiko.getalp.org/about-dbnary/dataset/

5 https://github.com/openlink/virtuoso-opensource

	METEOR-Baseline	METEOR-DB-4-catg	METEOR-DB-ALL-catg
online A	36.97 %	36.91 %	**37.13 %**
rbmt-1	33.74 %	33.60 %	**33.89 %**

Table 2 . METEOR-Baseline vs METEOR-DBnary for 2 systems picked up randomy from WMT14 data (French-English MT)

The results of *Table 2* above show that *METEOR-DBnary-4-catg* and *METEOR-Baseline* (based on WordNet) both obtain very similar scores while the size of the WordNet dictionary is 2,5 times larger than that of DBnary (4-catg). Moreover, using all existing categories in DBnary, we notice an increase of +0.20% in the final score. In other words, slightly more synonym matches are obtained with the latter metric based on the full English DBnary.

3.2. Lemmatisation issues

For English, METEOR uses the *Morphy-7WN* function as well as an exception list, attempting to find the lemma of a given word. However, for other target languages, it is very difficult to identify rules in order to lemmatize an inflected form. Thus, for the moment, we use TreeTagger (Schmid, 1994) to lemmatize words for other languages.

In order to avoid relaunching *TreeTagger* for each new entry, we adopt a preprocessing step that is needed before launching our modified METEOR. During this step, TreeTagger is run on the full evaluation corpus to collect a list of unique words with their respective lemmas.

We compare the impact of the lemmatization tool used (*Morphy vs TreeTagger*) by METEOR on the same two systems of WMT 2014 (see Table 3 results).

	METEOR-Morphy	METEOR-TTG
online A	36.97 %	**37.00 %**
rbmt-1	33.74 %	**33.76 %**

Table 3. Impact of lemmatization; METEOR-Morphy vs METEOR-TTG for 2 systems picked up randomy from WMT14 data (French-English MT)

In Table 3, *METEOR-TTG* shows a slight increase in the score, compared to *METEOR-Morphy*, because TreeTagger lemmatizes all categories (including Adverbs), whereas Morphy lemmatizes only three categories (Noun, Verb and Adjective).

4. Correlation with human judgements

In order to evaluate the correlation of our proposed *METEOR-DBnary* with human judgements of machine translation outputs, we used the data from the WMT13 Metrics Shared Task (Machacek and Bojar, 2013) for English-to-Spanish MT, and from the WMT14 Metrics Shared Task (Machacek and Bojar, 2014) for French-English, English-French, English-German and English-Russian MT.

We present the results in a similar fashion as in the WMT *metrics* task methodology using the following metrics. More details and formulas can be found in (Machacek and Bojar, 2013) or (Machacek and Bojar, 2014).

- System-level using *Pearson* correlation coefficient between system ranking based on human judgments *versus* METEOR (we will use our augmented metric and compare it to the baseline METEOR).
- Segment-level using *Kendall's* τ correlation between system ranking, at the sentence level, based on human judgments *versus* METEOR (we will use our augmented metric and compare it to the baseline METEOR).

Our results were obtained with two different configurations of METEOR:

- *METEOR-Baseline*: currently available *METEOR Universal* tool with the synonym module activated for English only (using the WordNet resource) - see *table 4*.
- *METEOR-DBnary* : our augmented-METEOR with the synonym module activated for English, French, Spanish, German and Russian, using our lexical resource DBnary - see *table 4*.

It is worth mentioning that, in order to use our new synonym dictionaries and evaluate our approach, we activated the synonym module in METEOR for the following languages: French Spanish Russian and German, by assigning a weight of 0.8 to each languages (same weight as for the English module).

		WMT14			WMT13
	FR-EN	**EN-FR**	**EN-RU**	**EN-GE**	**EN-ES**
METEOR-Baseline	**.975**	.941	.923	.263	.886
METEOR-DBnary	.973	**.943**	**.928**	**.320**	**.895**

Table 4. System-level correlations (*Pearson* Correlation Coefficient) between METEOR-Baseline (or METEOR-DBnary) and the WMT13/WMT14 human rankings.

		WMT14 τ			WMT13 τ
	FR-EN	**EN-FR**	**EN-GE**	**EN-RU**	**EN-ES**
METEOR-Baseline	**.**406	.280	.238	.427	.184
METEOR-DBnary	**.406**	**.284**	**.240**	**.435**	**.187**

Table 5. Segment-level correlations (*Kendall's* τ) between METEOR-Baseline (or METEOR-DBnary) and the the WMT13/WMT14 human rankings.

Table 4 shows that the use of DBnary slightly improves the system-level correlations of the METEOR score to human judgments in all language pairs except for French-English. *Table 5* shows the same trend for segment-level correlations which confirms that DBnary can be a useful resource for MT evaluation. The use of DBnary seems very promising for Russian and German as target languages.

Finally, *Table 6* shows the absolute values of both METEOR (Baseline *vs* DBnary) for the same language pairs and for a system randomly chosen in the WMT datasets (system *rbmt-1* is a rule-based machine translation system). As expected, METEOR score increases when used with DBnary since in that case the metric maps more words with the same meaning, using DBnary as lexical resource for synonymy.

		WMT 14		**WMT13**
	EN-FR	**EN-RU**	**EN-DE**	**EN-ES**
METEOR-Baseline	50.94	36.21	38.06	49.88
METEOR-DBnary	**52.34**	**37.60**	**41.51**	**51.04**

Table 6 : Comparison of METEOR-Baseline vs METEOR-DBnary (for system *rbmt-1*)

We present below some examples of matches obtained for METEOR-Baseline and for METEOR-Dbnary.

- **Example 1 : EN-FR (**system *rbmt-1***)**
  - **Reference :** Si la personne la plus puissante d'Europe peut être visée, alors les *dirigeants* d'entreprise sont sûrement *aussi* des cibles potentielles.
  - **Hypothesis:** Si la personne la plus puissante de l'Europe peut être visée, alors sûrement les *chefs* de file des affaires sont *également* les cibles potentielles.
  - **Synonym match : word → lemma → synonym list**
    - dirigeants → dirigeant → [*chef*, maître, leader, directeur]
    - chefs → *chef* → [tête, maître, cuisinier, leader, maître_queux, patron]

  => the lemma "*chef"* exists in the synonym list of the word "*dirigeant"*, thus "*dirigeants*" and "*chefs*" are considered as synonyms.
    - aussi → *aussi* → [ainsi, *également*, itou]
    - également → *également* → [*aussi*, pareillement, de_même, par_ailleurs]

  => METEOR considers "*aussi*" and "*également*" as synonyms, because "*aussi*" belongs in the synonym list of "*également*" and "*également*" exists in the synonym list of "*aussi*" .
  - **Segment score :**

    METEOR-Baseline : 0.6762
    METEOR-DBnary : 0.7290
- **Example 2: EN-FR (**system *rbmt-1***)**
  - **Reference :** J'estime qu'il est concevable que ces données soient *utilisées* dans leur intérêt_mutuel.
  - **Hypothesis :** Je pense qu'il est concevable que ces données soient *employées* pour le bénéfice_mutuel.

- **Synonym match : word → lemma → synonym list**
  - employées → employer → [occuper, *utiliser*]
  - utilisées → *utiliser* → [user]

=> During word-to-word alignment, METEOR considers the words *"utilisées"* (in REF) and *"employées"* (in HYP) as synonyms, because in the step of synonym match, we find that the lemma *"utiliser"* exists in the synonym list of the lemma *"employer"*.

- **Segment score :**

METEOR-Baseline : 0.6609
METEOR-DBnary : 0.7133

➢ **Example 3: EN-FR (**system *rbmt-1***)**

- Reference : Il me parlait, m'encourageait constamment, il *habitait* mon corps.

- Hypothesis: Il me parlerait, m'encouragent constamment, il a *vécu* dans mon corps.

- **Synonym match: word → lemma → synonym list**
  - habitait → *habiter* → [occuper]
  - vécu → vivre → [*habiter*, nourriture]

=> the lemma *"habiter"* exists in the synonym list of the word *"vivre"*, thus *"habitait"* and "*vécu*" are considered as synonyms.

- **Segment score :**

METEOR-Basline : 0.6743
METEOR-DBnary : 0.7688

5. Conclusion

We proposed an extension of METEOR, a well-known MT evaluation metric, for multiple target languages using our in-house lexical resource called DBnary. Our *augmented* METEOR obtained a better correlation with human judgements than the *baseline* METEOR, on the WMT 2014 metrics dataset for English-to-(French, Russian, German, Spanish) language pairs.

The modified code allowing to call METEOR for new target languages (French, Russian, German, Spanish) is made available to the research community from the following link (*http://kaiko.getalp.org/about-dbnary/meteor-with-dbnary/*).

In a near future, more target languages (today 21 in total) could be plugged very quickly by us or by interested users (please contact us if you want to contribute) using the same lexical resource (DBnary). The same adaptation of synonym matches could be done to TER-Plus (Snover et al., 2009b). Finally, using WSD, such as done in (Apidianaki and Marie, 2015), is another interesting avenue for improving correlation between automatic evaluation metrics and human judgements.

References

Apidianaki M., Marie B. (2015) METEOR-WSD: Improved Sense Matching in MT Evaluation. *Proceedings of the 9th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST'9)*, 4 June, Denver, Colorado, pp. 49--51.

Banerjee, S. and Lavie, A. (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization,*

Cimiano, P., Buitelaar, P., McCrae, J., & Sintek, M. (2011). LexInfo : A Declarative Model for the Lexicon-Ontology Interface. *Web Semantics: Science, Services and Agents on the World Wide Web*, *9*(1), 29–51. doi:10.1016/j.websem.2010.11.001

Denkowski, M. and Lavie, A. (2014). METEOR Universal: Language Specific Translation Evaluation for Any Target Language, *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation.*

Denkowski M., Lavie A. (2010a). "METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages", *Proceedings of the ACL 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR.*

Denkowski M., Lavie A. (2010b). Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level. I*n Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* , pages 250–253, Los Angeles, California, USA.

Doddington G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. *Proceedings of HLT 2002. San Diego, California*. March 24-27, 2002. 138-145.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Hellmann, S., Moran, S., Munich, L. M. U., Brümmer, M., Chiarcos, C., & Sukhareva, M. (2015). OLiA – Ontologies of Linguistic Annotation. *Semantic Web Journal - Special Issue on Multilingual Linked Open Data*, *6*(4), 379–386.

Klyne, G., & Carroll, J. J. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. (G. Klyne & J. Carroll, Eds.) *Structure*. World Wide Web Consortium. Retrieved from http://www.w3.org/TR/rdf-concepts/

Sérasset, G. (2015). The DBnary eco-system, data and APIs. *1st Summer Datathon on Linguistic Linked Open Data (SD-LLOD-15), 15 June,* Madrid, Spain.

Machacek, M. and Bojar, O. (2013). Results of the WMT13 Metrics Shared Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria.

Machacek, M. and Bojar, O. (2014). Results of the WMT14 Metrics Shared Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA.

McCrae, J., Spohr, D., & Cimiano, P. (2011). Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In G. Antoniou, M. Grobelnik, E. P. B. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, & J. Z. Pan (Eds.), *ESWC (1)* (Vol. 6643, pp. 245–259). Springer.

Moro A., Raganato A., Navigli R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Papineni K., Roukos S., Ward T., & Zhu W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of 40th meeting of the Association for Computational Linguistics*. Philadelphia, USA. July 7-12, 2002. 311-318.

Schmid, N. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Snover M., Dorr B., Schwartz R., Micciulla L., Makhoul J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of AMTA 2006. Cambridge, MA, USA*. August 8-12, 2006. 223-231.

Snover M., Madnani N., Dorr B., Schwartz R. (2009a). Fluency, Adequacy, or HTER? Exploring Different Human Judgments with a Tunable MT Metric. *Proceedings of Fourth Workshop on Statistical Machine Translation at the 12th Meeting of the European Chapter of the Association for Computational Linguistics*. March 30-31, 2009. 259-268.

Snover M., Madnani N., Dorr B., Schwartz R. (2009b). TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation Journal.* September, 2009. 117-127.