Word2Vec vs DBnary: Augmenting METEOR using Vector Representations or Lexical Resources?

Christophe Servan[1,2], Alexandre Bérard[1], Zied Elloumi[1,3], Hervé Blanchon[1] & Laurent Besacier[1]

[1]LIG – Univ. Grenoble Alpes
Domaine Universitaire
38401 St Martin d'Hères, France
firstname.lastname@imag.fr

[2]SYSTRAN
5 Rue Feydeau
75002 Paris, France
firstname.lastname@systran.fr

[3]LNE
29 Avenue Roger Hennequin
78190 Trappes, France
firstname.lastname@lne.fr

Abstract

This paper presents an approach combining lexico-semantic resources and distributed representations of words applied to the evaluation in machine translation (MT). This study is made through the enrichment of a well-known MT evaluation metric: METEOR. This metric enables an approximate match (synonymy or morphological similarity) between an automatic and a reference translation. Our experiments are made in the framework of the *Metrics* task of WMT 2014. We show that distributed representations are a good alternative to lexico-semantic resources for MT evaluation and they can even bring interesting additional information. The augmented versions of METEOR, using vector representations, are made available on our *Github* page.

1 Introduction

Learning vector representations of words using neural networks has generated a strong enthusiasm in the NLP research community. In particular, many contributions were proposed after the work of (Mikolov et al., 2013a; Mikolov et al., 2013b; Mikolov et al., 2013c) on training word embeddings. The main reasons for this strong interest are: the proposal of a simple and efficient neural architecture to learn word vector representations, the availability of an open source tool *Word2Vec*[1] and the rapid structuring of a user community[2]. Later on, several contributions have extended the work of Mikolov on word vectors to phrases (sequences of words) (Mikolov et al., 2013b; Le and Mikolov, 2014a) and to bilingual representations (Luong et al., 2015). All these vector representations capture similarities between words, phrases or sentences at different levels (morphological, semantic).

However, although these representations can be semantically informative, they do not exactly replace fine-grained information available in lexical-semantic resources such as *WordNet* (Fellbaum, 1998), *BabelNet* (Navigli and Ponzetto, 2010), or *DBnary* (Sérasset, 2012). Such lexical resources are also more easily interpretable by humans as shown in (Panchenko, 2016), but their construction is costly while word embeddings can be trained *ad infinitum* on any monolingual or bilingual corpora.

In short, both approaches (lexical resources and word embeddings) have their pros and cons. However, few studies have attempted to compare and combine them. Pioneering work of Faruqui et al. (2014) proposed to refine representations learning using lexical resources. The idea is to force words connected in the lexical network, to have a close representation (for example through a synonymy link). The technique proposed is evaluated on several benchmarks (word similarity, sentiment analysis, finding of synonyms). More recently, Panchenko (2016) and Rothe and Schütze (2015) extended word embeddings to sense embeddings and tried to compare them to lexical synsets.

[1]http://word2vec.googlecode.com/svn/trunk/

[2]https://groups.google.com/d/forum/word2vec-toolkit

Contributions: this article attempts to review the contribution of vector representations to measure sentence similarity. We compare them with similarity measures based on lexical resources such as *WordNet* or *DBnary*. Machine Translation (MT) evaluation was identified as a particularly interesting application to investigate, since MT evaluation is still an open problem nowadays. More precisely, we propose to augment a well known MT evaluation metric (METEOR (Banerjee and Lavie, 2005)) which allows an approximate matching (through synonymy or morphological similarity) between MT hypothesis and reference. The augmented versions of METEOR proposed (using word embeddings, lexical resources or both) allow us to objectively compare the contribution of each approach to measure sentence similarity. For this, correlations between METEOR and human judgements (of MT outputs) are measured within the framework of WMT 2014 *Metrics* task. The code of the augmented versions of METEOR is also provided on our *Github* page[3].

Outline: in section 2 (Related Work), we quickly present METEOR, lexical resources and word embeddings. Section 3 presents our propositions to augment METEOR in order to conduct a fair comparison between lexical resources and vector representations respectively. Section 4 presents our experiments made within the framework of WMT 2014, as well as quantitative and qualitative analyses. Finally, section 5 concludes this work and gives some perpectives.

2 Related Work

2.1 An automatic metric for MT evaluation: METEOR

2.1.1 The origins

METEOR was proposed to compensate BLEU's and NIST's weaknesses (Papineni et al., 2002; Doddington, 2002). In short, METEOR was created to better correlate with human judgements by using more than word-to-word alignments between a hypothesis and some references. The alignment is made according to three *modules*: the first stage uses exact match between word surface forms (*Exact* module), the second one compares word stems (*Stems* module) and the third one uses synonyms (*Synonym* module) from a lexical resource such as WordNet (available for English only in METEOR).

One contribution of this paper is to propose an alternative to *Stems* and *Synonym* modules: our proposed add-on will be called *Vectors* module later on.

2.1.2 Recent extensions of METEOR

METEOR-NEXT (Denkowski and Lavie, 2010a) was proposed to better correlate with HTER (Human-targeted Translation Edit Rate – *HTER* (Snover et al., 2006)). HTER is a semi-automatic post-editing based metric, which measures the edit distance between a hypothesis and a reference. METEOR-NEXT proposes to go further than just word-to-word alignment by using phrase-to-phrase alignments. For this, phrase databases were created for several languages like English (Snover et al., 2009), German, French or Czech (Denkowski and Lavie, 2010b). More recently, another version called *METEOR Universal* used bitexts to extract paraphrases (Denkowski and Lavie, 2014).

METEOR was also extended by using Word Sense Disambiguation (WSD) techniques (Apidianaki and Marie, 2015). The authors used *Babelfly* (Moro et al., 2014) for several langage pairs (translation from French, Hindi, German, Czech and Russian to English). A better correlation with human judgement at segment level was observed using WSD in METEOR.

Finally, to extend the use of *Synonym* module to target languages others than English, Elloumi et al. (2015) proposed to replace WordNet by DBnary (Sérasset, 2012). The new target languages equipped with a *Synonym* module were French, German, Spanish, Russian and English.

2.2 Lexical resources

2.2.1 WordNet

WordNet is a well known lexical resource for English. Created at the University of Princeton (Fellbaum, 1998), it is used in several NLP tasks such as Machine Translation, Word Sense Disambiguation,

[3] https://github.com/cservan/METEOR-E

Cross-lingual Information Retrieval, etc. WordNet links nouns, verbs, adjectives and adverbs to a set of synonyms called "synsets". Each synset represents a specific concept.

Synsets are linked to each other according to semantic, conceptual and lexical relations. Words with multiple meanings correspond to multiple synsets and meanings are sorted according to their frequency. WordNet is available in several languages (Arabic, French, etc.) but these versions are not freely available. In METEOR, only English WordNet is used to match hypothesis and reference words according to their meanings. It contains more than 117,000 synsets.

To extract lemmatized forms, METEOR uses a function called *Morphy-7WN1* which firstly checks special cases in an exception list and secondly uses rules to lemmatize words according to their syntactic class.

2.2.2 DBnary

DBnary is a multilingual lexical resource in RDF format (Klyne and Carroll, 2004). This resource has been collected by Sérasset (2012). Lexical data are represented using the LEMON vocabulary (McCrae et al., 2011). Most Part-of-Speech tags are linked with *Olia standards* or *Lexinfo* vocabularies (Chiarcos and Sukhareva, 2015; Cimiano et al., 2011) which makes them reusable in many contexts.

DBnary is downloadable or available online through a SPARQL access point. Lexical data are automatically extracted from Wiktionary, Wikipedia's dictionary for 21 languages[4].

	English	French	Russian	German
Number of entries	620 K	322 K	185 K	104 K
Number of meanings	498 K	416 K	176 K	116 K
Number of synsets	35 K	36 K	31 K	33 K

Table 1: Detail of the data used from DBnary for the languages targeted in this paper.

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).

DBnary also contains more than 4.6M translations going from the 21 extracted sources languages to more than 1500 different target languages. Additionally, DBnary contains lexicosemantic relations (syno/anto-nyms, hypo/hypero-nyms, etc.). 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[5].

Lemmatized forms for DBnary are based on the *TreeTagger* module (Schmid, 1995), which enables us to find the corresponding synsets.

2.3 Monolingual and bilingual embeddings

2.3.1 Overview

Learning word embeddings is an active research area (Bengio et al., 2003; Turian et al., 2010; Collobert et al., 2011; Huang et al., 2012). The main idea is to learn a word representation according to its context: the surrounding words (Baroni and Zamparelli, 2010). The words are projected on a continuous space and those with similar context should be close in this multi-dimensional space. When word vectors are available, a similarity between two words can be measured by a metric such as a cosine similarity.

Using word-embeddings for machine translation evaluation is appealing since they can be used to compute similarity between words or phrases in the same language (monolingual embeddings capture intrinsically synonymy or morphological closeness) or in two different languages (bilingual embeddings allow to directly compute a distance between two sentences in different languages). We use the *MultiVec* (Bérard et al., 2016) toolkit for computing and managing the continuous representations of texts. It includes word2vec (Mikolov et al., 2013a), paragraph vector (Le and Mikolov, 2014b) and bilingual distributed representations (Luong et al., 2015) features.

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

[5]`http://kaiko.getalp.org/about-dbnary/`

2.3.2 Use of vector representations in NLP evaluation

Zou et al. (2013) proposed to use bilingual word embeddings to detect similarities for word alignment. This information is used as an additional parameter in a phrase-based machine translation system. (Banchs et al., 2015) proposed to explore a metric funded on latent semantic analysis (Salton et al., 1975) to extract semantic embeddings and measure the similarity between two sentences. Finally, these word embeddings were used to enrich ROUGE, a metric for evaluating automatic summarization (Ng and Abrecht, 2015).

As far as MT evaluation is concerned, Gupta et al. (2015) proposed a metric based on neural network language models jointly with dependency trees to link an hypothesis to a reference. Meanwhile, Vela and Tan (2015) proposed an approach to model document embeddings to predict translation adequacy.

These works are close to ours but they propose metrics which need to be learned and optimized to a specific task or domain. In our work, we use word embeddings trained once and for all on a (large) general corpus. Our detailed methodology to augment METEOR metric is presented in the next section.

3 Augmented METEOR

3.1 Data and protocol

We evaluate our augmented METEOR through WMT14 framework (*metrics* task (Machacek and Bojar, 2014)). This framework enables us to estimate the correlation of proposed evaluation metric with human judgements for several machine translation outputs and several language pairs (English-French, English-German, English-Russian, and vice versa). In our experiments, we use segment level Kendall's τ correlation coefficient, as proposed in WMT14 (based on systems ranking at sentence level by humans, compared to automatic metric ranking).

We augment METEOR in two ways: firstly, we replace the use of lexical resources by the use of word embeddings. In other words, we replace *Stem* and *Synonym* modules by our new *Vector* module. Secondly, we combine lexical resources and word embeddings by using jointly *Stem*, *Synonym* and our *Vector* module.

To summarize, the following variants of METEOR are evaluated:

- *METEOR Baseline*: the METEOR score is estimated using *Exact*, *Stem*, *Synonym* and *Paraphrase* modules for English as a target language and *Exact*, *Stem* and *Paraphrase* modules for other target languages,
- *METEOR DBnary*: similar to *METEOR Baseline* but *Synonym* module is available for any target language since it uses DBnary resource instead of Wordnet,
- *METEOR Vector*: the *Stem* and *Synonym* modules are replaced by the *Vector* module ;
- *METEOR Baseline + Vector*: the *METEOR Baseline* configuration is augmented with the *Vector* module ;
- *METEOR DBnary + Vector*: the *METEOR DBnary* configuration is augmented with the *Vector* module.

3.2 *METEOR DBnary*

As mentioned in section 2.1, the *Synonym* module of METEOR uses WordNet's synsets (117K entries for English). As an alternative, we use another lexical resource: DBnary (Sérasset, 2012), as proposed recently by Elloumi et al. (2015). This allows us to use *Synonym* module for any target language: French, German, Spanish, Russian and English.

More precisely, synonym relations are extracted from DBnary using SPARQL request on the DBnary server[6]. We extract data for English, French, Russian and German languages. The extraction process outputs relations in the following format: $lemma \rightarrow Synonym$. Then, these data are projected to the

[6] http://kaiko.getalp.org/about-dbnary/online-access/

WordNet format used in METEOR code. This process gives an identifier (ID) for each lemma and builds a list of synonym IDs for each lemma such as: $lemma \rightarrow ID\_Syn_1,\ ID\_Syn_2,\ ID\_Syn_3$.

The first two lines of Table 3 compare *METEOR DBnary* and *METEOR Baseline* for several French-English MT systems submitted to WMT14 (Bojar et al., 2014).

METEOR DBnary improved the score by 0.7 points from *METEOR Baseline*. In other words, DBnary seems to match more synonyms than WordNet, despite the fact that WordNet is 3.3 time bigger than DBnary in English. This could be due to the fact that WordNet has only 4 morpho-syntactic categories (Noun, Verbs, Adjectives and Adverbs) while DBnary has more morpho-syntactic categories.

3.3 *METEOR Vector*

As mentioned in section 2.3.2, we propose to replace lexical resources by word embeddings. Word embeddings capture the context of the words. Consequently, similar word vectors may correspond to synonyms or morphological variants (see section 2.3).

Language	Corpora	# of lines	# of source words	# of target words
French–English	Europarl V7 + news commentary V10	2.2 M	67.2 M	60.7 M
German–English	Europarl V7 + news commentary V10	2.1 M	57.2 M	59.7 M
Russian–English	Common Crawl + news commentary V10 + Yandex	2.0 M	47.2 M	50.3 M

Table 2: Bilingual corpora used to train the word embeddings for each language pair.

In our *Vector* module, the matching between two words is done using a similarity score derived from the cosine similarity. If the similarity score is higher than a threshold, the words are considered as matched (potential synonyms or morphological proximity). In our experiments, we evaluate using: (a) a default threshold fixed to 0.80 (b) an oracle threshold obtained empirically on the WMT14 data set (Machacek and Bojar, 2014).

Table 2 summarizes data used to train monolingual word embeddings and bilingual word embeddings. These word embeddings were trained with a CBOW model, a vector size of 50 and a windows size ± 5 words, thanks to the MultiVec toolkit (Bérard et al., 2016).

Metrics	Systems: online A	online B	online C	rbmt 1	rbmt 4
METEOR Baseline	36.33	36.71	31.19	33.00	31.65
METEOR DBnary	36.93	37.33	32.01	33.69	32.42
METEOR Vector	37.00	37.34	31.87	33.67	32.34
METEOR Baseline + Vector	37.08	37.40	31.96	33.75	32.45
METEOR DBnary + Vector	37.53	37.88	32.60	34.32	33.05

Table 3: METEOR scores (all configurations) on the *newstest* corpus of the WMT14 translation evaluation task from French to English.

The results presented in table 3 show that word embeddings (*Vector* module) can efficiently replace lexical resources (*Synonym* and *Stem* modules) to match words in the translation hypothesis with those in the reference. In addition, their combination shows a good potential to match even more words between hypothesis and reference. In the next section, we evaluate if the proposed versions of *augmented* METEOR better correlate with human judgements.

4 Correlations of Augmented METEOR with Human Judgements

4.1 Results of different METEOR configurations

In these experiments, we present results obtained with the *Vector* module based on two threshold values: a default one (0.80) and an oracle one which maximizes the correlation with human judgement.

Table 4 presents the correlation scores obtained within the framework of WMT14 metrics task (Machacek and Bojar, 2014)[7]. The evaluation is done according to several translation tasks: from English to French (en–fr), German (en–de) and Russian (en–ru), and vice versa. French, German and Russian

[7]For better readability, we do not add standard deviations in the tables. These numbers will be, however, provided in supplementary material put on the paper web page (`https://github.com/cservan/METEOR-E/paper`).

as target languages represent a growing difficulty due to their morphology. English as target language allows to compare the lexical databases (Wordnet *vs* DBnary).

To English. Firstly, when the translation direction is to English, we can observe that *METEOR Baseline* and *METEOR Vector* get equivalent results in average. *METEOR DBnary* also obtains similar results to *METEOR Baseline*. When we combine WordNet lexical resource and word embeddings (*METEOR Baseline + Vector*), the reference score is increased by $0,005$ points. If the combination is done with DBnary's lexical data (*METEOR DBnary + Vector*), the improvement is similar.

For *Vector* module, optimization of the threshold slightly improves the average correlation. Combination of *METEOR Baseline + Vector* or *METEOR DBnary + Vector* improves by $0,002$ points when the threshold is optimized.

From English. Secondly, when the translation direction is from English, we can observe an improvement of the correlation score obtained with *METEOR DBnary*, compared with *METEOR Baseline*. This is due to the fact that for French, German and Russian as target languages, *METEOR Baseline* does not use any *Synonym* module. Our *METEOR Vector* with the default threshold also gets better correlation scores compared to *METEOR DBnary* ($+0.003$ points in average). The combinations *METEOR Baseline + Vector* and *METEOR DBnary + Vector* further improve correlations with human judgements ($+0.001$ points in average). Finally, when we use an oracle threshold for *Vector* module, improvements are bigger and can reach 0.013 points in average, compared to *METEOR Baseline*.

Language pairs	fr–en		de–en		ru–en		Average	
Metric	Threshold	τ	Threshold	τ	Threshold	τ	Threshold	τ
METEOR Baseline	–	0.406	–	0.334	–	0.329	–	0.356
METEOR DBnary	–	0.408	–	0.334	–	0.328	–	0.357
METEOR Vector	0.80	0.407	0.80	0.332	0.80	0.328	0.80	0.356
METEOR Baseline + Vector	0.80	0.407	0.80	0.343	0.80	0.332	0.80	0.361
METEOR DBnary + Vector	0.80	0.407	0.80	0.337	0.80	0.338	0.80	0.361
METEOR Vector	0.89	0.411	0.78	0.333	0.80	0.328	0.82	0.357
METEOR Baseline + Vector	0.73	0.412	0.80	0.343	0.88	0.333	0.80	0.363
METEOR DBnary + Vector	0.73	0.413	0.79	0.338	0.80	0.338	0.77	0.363
Language pairs	en-fr		en-de		en-ru		Average	
Metric	Threshold	τ	Threshold	τ	Threshold	τ	Threshold	τ
METEOR Baseline	–	0.280	–	0.238	–	0.427	–	0.315
METEOR DBnary	–	0.284	–	0.239	–	0.435	–	0.319
METEOR Vector	0.80	0.290	0.80	0.241	0.80	0.436	0.80	0.322
METEOR Baseline + Vector	0.80	0.288	0.80	0.241	0.80	0.440	0.80	0.323
METEOR DBnary + Vector	0.80	0.289	0.80	0.242	0.80	0.439	0.80	0.323
METEOR Vector	0.72	0.295	0.79	0.241	0.72	0.439	0.74	0.325
METEOR Baseline + Vector	0.86	0.296	0.79	0.242	0.79	0.445	0.81	0.328
METEOR DBnary + Vector	0.88	0.294	0.75	0.245	0.79	0.443	0.81	0.327

Table 4: Correlation score at segment level between several METEOR configurations and human judgements (WMT14 framework). Scores obtained with the *Vector* module are presented firstly with the default threshold (0.80) and secondly with the oracle threshold (under the dashed line).

4.2 Investigating more embeddings configurations

In the previous section, *METEOR Vector* used a simple and monolingual word embedding configuration. This section investigates more configurations (monolingual and bilingual) to improve METEOR.

In this experiment, we focus only on *METEOR Vector*. Indeed, the *monolingual (baseline)* shown in table 6 corresponds to the line *METEOR Vector* in Table 4. Firstly, we propose to train our embeddings on bitexts (Table 2) using *bivec* approach (Luong et al., 2015). We also try to train pseudo-bilingual embeddings on a pseudo bitext with target language text and POS tags (see an example in Table 5). The main idea is to strongly link words with their syntactic class when learning word embeddings. We

madam president , on a point of order .	$\Leftrightarrow$	NOUN NOUN PUNCT ADP DET NOUN ADP NOUN PUNCT

Table 5: Example of bitext where the target side is replaced by POS.

call this kind of model *pseudo-bilingual with POS*. In the same way, we train bilingual models called *pseudo-bilingual with lemmas*, where the POS tags are replaced by lemmas. In addition, we also learn word embeddings with lemmas only and bilingual models with lemmas only.

Models:	*monolingual (baseline)*	*bilingual*	*pseudo-bilingual with POS*	*pseudo-bilingual with lemmas*	*monolingual (lemmas)*	*bilingual (lemmas)*
To English	0.356	0.354	0.355	0.354	0.357	0.357
From English	0.322	0.322	0.320	0.325	0.324	0.318

Table 6: Average correlation score at segment level for *METEOR Vector* with several training configurations of word embeddings with the default threshold (0.80).

In the Table 6, we compare several training configuration of the word embeddings through the same protocol as previous section (only average correlations are reported while the detailed results will be provided as supplementary material on the paper web page). When we observe the average results, the *bilingual* embeddings seem not to be as efficient as the monolingual baseline. The pseudo-bilingual approaches with POS and Lemmas obtained slightly the same results as the monolingual baseline regarding all the configurations we have. Finally, the monolingual model learned on lemmas (instead of words) tends to be slightly better when the translation direction is to English. However, this trend should be confirmed in a future investigation.

4.3 Discussion

The correlation scores obtained with the enriched metric tend to suggest that distributed representations are as powerful as lexico-semantic resources for automatic MT evaluation. Furthermore, vector representations can bring additional information, and they are definitely useful when no lexical resource is available in the target language.

Considering the average correlation scores obtained, the configurations *METEOR Vector* and *METEOR DBnary* are comparable, except on German language, for which *METEOR Vector* obtained a better correlation score. On the other hand, when we combine lexical data with *Vector* module (*METEOR DBnary + Vector*), we observe a small increase of the correlation score, in particular when threshold is tuned, which suggests a tunable version of METEOR.

Finally, several embeddings variants were trained but it seems that monolingual models are efficient enough for the specific task (MT evaluation) considered here.

4.3.1 Examples

To illustrate the word matching obtained by our versions of METEOR, we analyze two examples from the evaluation data set. In these examples, we present the alignments obtained with *METEOR DBnary + Vector*.

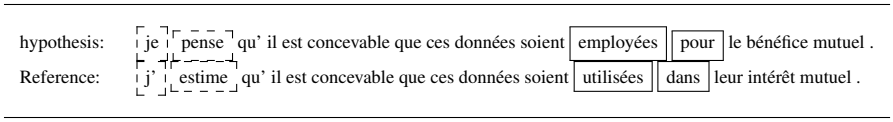

Table 7: First example from the system *rbmt 1* evaluated with the combination *METEOR DBnary + Vector*. The relations detected with the lexical resource DBnary are framed in continuous line while those obtained thanks to the distributed representations are framed in dotted line.

The example presented in table 7 shows *rbmt 1* system output submitted during the WMT14 translation task. *METEOR baseline* found only alignments for words with the same surface forms (*"qu' ", "il", "est"*, etc. – these forms are found identical thanks to the *Exact* module and are not highlighted here). The *Synonym* module based on DBnary makes it possible to find a correspondence between words *"employées" – "utilisées"* and *"pour" – "dans"*. Lastly, *Vector* module indicates that words *"pense"* and *"estime"* are contextually closed, just as the words *"je"* and *"j"'*. When the example is only evaluated with *METEOR Vector*, words *"employées"* and *"utilisées"* are also paired with the default threshold (0.80). On the other hand, the words *"bénéfice"* and *"intérêt"* are paired by the module *Vector* only if the decision threshold is lowered to 0.75.

In the second example presented in table 8, the hypothesis is provided by *rbmt 4* system. As in the previous example, the correspondences found with *Synonym* module based on DBnary (framed by one

hypothesis:	le créateur de SAS disait il faisait un genre du feuilleton géopolitique .
Reference:	le père de SAS disait faire un genre de feuilleton géopolitique .

Table 8: Another example scored with the combination *METEOR DBnary + Vector*.

continuous line) are supplemented by those found by *Vector* module (dotted line): *Synonym* module found *"créateur"* – *"père"* and *"faisait"* – *"faire"*; while *"du"* and *"de"* are aligned thanks to *Vector* module.

These examples illustrate the complementarity between lexical resources and word embeddings for sentence similarity detection. Word vectors can enable to match important words (like *"pense"* and *"estime"* in our first example), but also empty words (like *"du"* et *"de"* in our second example).

4.3.2 Limitations of Word Embeddings

So far, we did not deal with Out-Of-Vocabulary (OOV) words in *METEOR Vector*. By OOV we mean words that do not have a vector representation because they were not found in the training corpus for word embeddings. In that case, no matching can occur between the word in the hypothesis and words in reference. Consequently, it might be interesting to carefully select the training corpus for word vectors so that it will be close enough to the machine translation outputs to evaluate. This could be considered in future works.

5 Conclusion and Perspectives

In this paper, we proposed to compare text similarity measures based on vector representations with similarity measures based on lexico-semantic resources. Our work was applied to machine translation evaluation and we extended an existing evaluation metric called METEOR. Our experiments have shown that word vector representations can be useful when no lexical resource is available in the target language. Moreover, it seems that these representations can bring complementary information in addition to lexical resources (experiments done for French, English, German and Russian as target languages).

Our future works on this topic will focus on the use of phrase embeddings to complement the *Paraphrase* module of METEOR. We also plan to introduce a *syntax flavor* in our *Vector* module by weighting the cosine distances differently according to the morpho-syntactic category of the words. Finally, we will study the adaptation of our approach to other metrics such as TER-Plus, for instance.

The tool, the data and the models presented in this paper will be put online[8] to facilitate reproducibility of the experiments we carried out.

Acknowledgements

This work was supported by the KEHATH project funded by the French National Agency for Research (ANR) under the grant number ANR-14-CE24-0016-03.

References

Marianna Apidianaki and Benjamin Marie. 2015. METEOR-WSD: Improved Sense Matching in MT Evaluation. In *the Proceedings of the 9th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST'9)*.

Raphael E. Banchs, Luis F. D'Haro, and Haizhou Li. 2015. Adequacy–Fluency Metrics: Evaluating MT in the Continuous Space Model Framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482, March.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics*.

[8] `https://github.com/cservan/METEOR-E`

Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155.

Alexandre Bérard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. 2016. MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP. In *The 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Christian Chiarcos and Maria Sukhareva. 2015. Olia – ontologies of linguistic annotation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):379–386.

P. Cimiano, P. Buitelaar, J. McCrae, and M. Sintek. 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.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Michael Denkowski and Alon Lavie. 2010a. Extending the meteor machine translation evaluation metric to the phrase level. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Michael Denkowski and Alon Lavie. 2010b. METEOR-NEXT and the METEOR paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*.

Michael Denkowski and Alon Lavie. 2014. METEOR Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*.

Zied Elloumi, Hervé Blanchon, Gilles Serasset, and Laurent Besacier. 2015. METEOR for Multiple Target Languages using DBnary. In *Proceedings of MT Summit 2015*.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard H. Hovy, and Noah A. Smith. 2014. Retrofitting word vectors to semantic lexicons. *CoRR*.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MA: MIT Press.

Rohit Gupta, Constantin Orasan, and Josef Van Genabith. 2015. Machine Translation Evaluation using Recurrent Neural Networks. In *Proceedings Workshop on Machine Translation (WMT), Metrics Shared Task*.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*.

Graham Klyne and Jeremy J. Carroll. 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax. Technical report.

Quoc V. Le and Tomas Mikolov. 2014a. Distributed Representations of Sentences and Documents. In *Proceedings of The 31st International Conference on Machine Learning*.

Quoc V. Le and Tomas Mikolov. 2014b. Distributed Representations of Sentences and Documents. In *Proceedings of the 31th International Conference on Machine Learning (ICML'14)*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*.

Matous Machacek and Ondrej Bojar. 2014. Results of the WMT14 Metrics Shared Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

John McCrae, Dennis Spohr, and Philipp Cimiano, 2011. *Linking Lexical Resources and Ontologies on the Semantic Web with Lemon*, pages 245–259. Springer Berlin Heidelberg, Berlin, Heidelberg.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *The Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better Summarization Evaluation with Word Embeddings for ROUGE. In *In The Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Alexander Panchenko. 2016. Best of both worlds: Making word sense embeddings interpretable. In *the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL*.

Sascha Rothe and Hinrich Schütze. 2015. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.

G. Salton, A. Wong, and C. S. Yang. 1975. A Vector Space Model for Automatic Indexing. *ACM*, 18(11):613–620, November.

Helmut Schmid. 1995. Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

Gilles Sérasset. 2012. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web Journal-Special issue on Multilingual Linked Open Data*, 6(4):355–361.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*.

Matthew Snover, Nitin Madnani, Bonnie J Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER?: exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*.

Mihaela Vela and Liling Tan. 2015. Predicting Machine Translation Adequacy with Document Embeddings. In *Proceedings Workshop on Machine Translation (WMT), Metrics Shared Task*.

Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *EMNLP*.