# Word2Vec vs DBnary ou comment (ré)concilier représentations distribuées et réseaux lexico-sémantiques ? Le cas de l'évaluation en traduction automatique

Untel Trucmuche & Unetelle Machinchose
Lab, adresse, CP Ville, Pays
prénom.nom@univ.fr

RÉSUMÉ

Cet article présente une approche associant réseaux lexico-sémantiques et représentations distribuées de mots appliquée à l'évaluation de la traduction automatique. Cette étude est faite à travers l'enrichissement d'une métrique bien connue pour évaluer la traduction automatique (TA) : METEOR. METEOR permet un appariement approché (similarité morphologique ou synonymie) entre une sortie de système automatique et une traduction de référence. Nos expérimentations s'appuient sur la tâche *Metrics* de la campagne d'évaluation WMT 2014 et montrent que les représentations distribuées restent moins performantes que les ressources lexico-sémantiques pour l'évaluation en TA mais peuvent néammoins apporter un complément d'information intéressant à ces dernières.

ABSTRACT

**Word2Vec vs DBnary or how to bring back together vector representations and lexical resources ? A case study for machine translation evaluation.**

This paper presents an approach combining lexical-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. METEOR 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 less efficient than lexical-semantic resources for MT evaluation but they can nonetheless bring interesting additional information.

MOTS-CLÉS : Traduction automatique, évaluation, similarité entre mots, représentations distribuées de mots, ressources lexico-sémantiques.

KEYWORDS: Machine translation, evaluation metrics, word similarity, word embeddings, lexical-semantic resources.

## 1 Introduction

Les approches d'apprentissage de représentations distribuées de mots utilisant des réseaux de neurones ont récemment généré un fort enthousiasme dans la communauté de recherche en traitement automatique du langage naturel. En particulier, de nombreuses contributions se sont appuyées sur les travaux de (Mikolov *et al.*, 2013a,b,c) sur l'apprentissage de représentations distribuées de mots (ou *« word embeddings »*). Une des raisons de cet engouement est la proposition d'une architecture

neuronale simple, la distribution d'un outil *Word2Vec* [1] et la structuration rapide d'une communauté d'utilisateurs [2]. Un certain nombre de contributions ont étendu ce travail à des fragments (ou séquences de mots) (Mikolov *et al.*, 2013b; Le & Mikolov, 2014) et à des représentation bilingues (Luong *et al.*, 2015). Ces représentations distribuées permettent de capturer des similitudes entre mots, fragments ou phrases à différents niveaux, et notamment aux niveaux morphologique et sémantique.

Bien que ces représentations soient sémantiquement informatives, elles ne tiennent pas compte des informations fines présentes dans les ressources lexico-sémantiques telles que *WordNet* (Fellbaum, 1998) ou *DBnary* (Sérasset, 2012), par exemple. On peut aussi remarquer qu'à ce jour, peu de travaux ont tenté d'associer *beautiful data* et *big data*, c'est à dire associer ressources lexicales et représentations distribuées entraînées sur de gros corpus. Cependant, citons les travaux récents de (Faruqui *et al.*, 2014) qui propose de raffiner l'apprentissage de représentations en utilisant des ressources lexico-sémantiques. L'approche consiste à forcer les mots connectés dans un réseau lexical à avoir une représentation proche, par exemple à travers un lien de synonymie. La technique proposée est évaluée sur plusieurs *benchmarks* (similarité de mots, analyse de sentiments, sélection de synonymes).

**Contribution** Cet article tente de faire le point sur l'apport des représentations distribuées pour mesurer la similarité entre phrases et leur capacité à prendre en compte des relations de synonymie ou de proximité morphologique entre mots. Nous comparons ces approches de représentations distribuées, apprises automatiquement, avec les mesures de similarité fondées sur des réseaux lexico-sémantiques. Une tâche nous a semblé particulièrement bien appropriée pour faire cette comparaison. Il s'agit de l'évaluation des systèmes de traduction automatique (TA). Plus précisément, nous proposons des ajouts à la métrique METEOR (Banerjee & Lavie, 2005) qui permet un appariement approché (similarité morphologique ou synonymie) entre une sortie de système automatique et une traduction de référence.

**Plan** Dans la partie 2, nous détaillons la métrique d'évaluation (METEOR) dans laquelle seront intégrés nos ajouts en terme de ressources lexicales et de représentations distribuées de mots (également présentées dans cette partie). La partie 3 présente notre méthodologie d'évaluation d'un score METEOR enrichi, fondée sur les données issues de la tâche *Metrics* de la campagne WMT 2014. Enfin, la partie 4 présente les résultats obtenus tandis que la partie 5 est consacrée à la conclusion et aux perspectives.

# 2 Etat de l'Art

## 2.1 Une métrique d'évaluation en TA utilisant des appariements non exacts : METEOR

### 2.1.1 METEOR : les origines

Les travaux de Banerjee & Lavie (2005) menant à METEOR avaient pour but de pallier à quelques faiblesses des métriques BLEU (Papineni *et al.*, 2002) et NIST (Doddington, 2002). Les faiblesses identifiées étaient notamment le fait que BLEU n'est pas une métrique orientée « rappel », qu'elle n'apporte pas d'indications du point de vue de l'adéquation grammaticale et qu'elle ne permet pas d'appariement approché au niveau mot entre l'hypothèse et la référence.

Le but du développement de METEOR était d'obtenir une métrique mieux corrélée avec les jugements

1. http ://word2vec.googlecode.com/svn/trunk/
2. https ://groups.google.com/d/forum/word2vec-toolkit

humains en utilisant plus que les alignements mot-à-mot entre une hypothèse et une ou plusieurs références. L'alignement est produit de manière incrémentale sur trois niveaux : la correspondance exacte au niveau de la forme de surface du mot (niveau 1 ou module *Exact*), la correspondance approchée en utilisant la racine du mot (niveau 2 ou module *Stem*) et enfin la correspondance approchée synonymique (niveau 3 ou module *Synonyme*) en utilisant les « synsets » de WordNet. Le second niveau utilise un « stemmer » qui permet d'obtenir automatiquement la racine des mots et le troisième niveau utilise les ressources lexico-sémantiques de WordNet en version anglaise.

Cet article va proposer des alternatives pour les niveaux 2 et 3 via l'utilisation d'une ressource lexico-sémantique autre que Wordnet (ici DBnary) pour évaluer la TA sur des couples de langues dont la cible est différente de l'anglais. En effet, la synonymie n'est pas traitée dans la version actuelle de METEOR pour des langues cibles autres que l'anglais car, malheureusement, toutes les versions de WordNet ne sont pas disponibles en libre accès. Par exemple, c'est le cas des versions française, espagnole ou allemande. Par ailleurs, les niveaux 2 et 3 de METEOR peuvent être réalisés via l'utilisation de la similarité entre deux représentations distribuées (sous la forme de deux vecteurs de nombres flottants) pour réaliser un appariement approché.

### 2.1.2 METEOR : extensions récentes

Proposé par Denkowski & Lavie (2010a), METEOR-NEXT vise à mieux corréler le jugement humain avec le score HTER (Human-targeted Translation Edit Rate – *HTER* (Snover *et al.*, 2006)). Le HTER est une mesure semi-automatique de la post-edition qui mesure la distance d'édition entre une hypothèse de traduction et le résultat de la post-édition humaine. Le but de cette métrique est d'aller plus loin que l'alignement mot-à-mot en proposant des alignements segments-à-segments. Les travaux de Snover *et al.* (2009) ont permis la création de la base de données pour l'anglais tandis que Denkowski & Lavie (2010b) ont proposé des bases de données pour d'autres langues comme l'allemand, le français ou le tchèque.

En 2014, les mêmes auteurs proposent une nouvelle version appelée *METEOR Universal* qui permet l'utilisation de corpus parallèles pour créer un dictionnaire de paraphrases (Denkowski & Lavie, 2014). Apidianaki & Marie (2015) ont proposé de prévenir les correspondances n'ayant pas le même sens sémantique en utilisant la désambigüisation sémantique au niveau du mot. Les références anglaises sont annotées et désambigüisées à travers *Babelfly* (Moro *et al.*, 2014) pour différentes paires de langues (français, hindi, allemand, tchèque et russe vers l'anglais). Les expériences montrent une meilleure corrélation avec le jugement humain au niveau du segment.

Enfin, pour pallier les lacunes du niveau 3 de METEOR (indisponibilité de Wordnet pour une langue cible autre que l'anglais), Elloumi *et al.* (2015) ont récemment proposé d'utiliser la base de données lexicale DBnary pour les langues cibles suivantes : français, allemand, espagnol, russe et anglais.

## 2.2 Ressources lexico-sémantiques

### 2.2.1 WordNet

WordNet est une l'une des plus célèbres bases de données lexicale pour l'anglais. Développé par les linguistes de l'Université de Princeton (Fellbaum, 1998), elle est devenue un élément important et une ressource très utile pour le Traitement Automatique de la Langue Naturelle, notamment, dans des domaines tels que la traduction automatique, la désambiguïsation lexicale, la recherche d'information multilingue, etc. WordNet lie les noms, verbes, adjectifs et adverbes à des ensembles de synonymes (appelés « synsets ») dont chacun d'eux représente un concept spécifique.

Les synsets sont connectés par des relations sémantiques, conceptuelles et lexicales. Les mots avec des significations multiples appartiennent à plusieurs synsets et leurs sens sont classés par fréquence. Il existe différentes versions de WordNet dans des langues autres que l'anglais, comme l'arabe, le français, etc. Cependant, ces ressources lexicales ne sont pas librement disponibles. Comme dit précédemment, METEOR utilise WordNet pour tenter de faire correspondre les mots des hypothèses de traduction avec les mots de la référence. La dernière version de WordNet 3.0 contient plus de 117 000 synsets.

Pour les formes lemmatisées, METEOR utilise la fonction Morphy-7WN1 de WordNet. Cette fonction utilise un processus en deux étapes pour trouver le lemme d'un mot particulier *W*. Tout d'abord, Morphy vérifie des exceptions dans une liste (transformations morphologiques qui ne sont pas régulières). Si *W* n'est pas dans la liste d'exceptions, Morphy utilise les règles de détachement pour les noms, verbes et catégories des adjectifs (pas de règles appliquées aux adverbes). Après chaque transformation, WordNet est utilisé pour extraire les informations de la catégorie syntaxique correspondante.

### 2.2.2 DBnary

DBnary est une ressource lexicale multilingue au format RDF (Klyne & Carroll, 2004) recueillie au Laboratoire d'Informatique de Grenoble (Sérasset, 2012). Les données lexicales sont représentées en utilisant un vocabulaire standard. La structure est définie en utilisant le vocabulaire LEMON (McCrae *et al.*, 2011). La plupart des informations liées aux parties-de-discours (Part-of-Speech ou POS) sont mises en correspondance avec les vocabulaires *Lexinfo* (Cimiano *et al.*, 2011) ou *Olia standards* (Chiarcos & Sukhareva, 2015), ce qui rend les données réutilisables dans de nombreux contextes.

DBnary est disponible soit sous forme de fichiers téléchargeables soit en ligne via un navigateur ou une application utilisant un point d'accès SPARQL public. Les données lexicales disponibles sont automatiquement extraites à partir de Wiktionary, le dictionnaire de Wikipedia, pour 21 langues différentes [3].

| | anglais | français | russe | allemand |
|---|---|---|---|---|
| Nombre d'entrées | 620 K | 322 K | 185 K | 104 K |
| Nombre de sens | 498 K | 416 K | 176 K | 116 K |
| Nombre de synsets | 35 K | 36 K | 31 K | 33 K |

TABLE 1 – Détail des données existantes dans DBnary pour les langues cibles considérées.

Parmi les données disponibles, il existe 2,9 millions d'entrées comportant des catégories lexicales, des formes canoniques, avec des prononciations (lorsque les formes fléchies sont disponibles), etc. Les entrées lexicales sont sous-divisées en sens lexicaux avec leurs définitions et quelques exemples d'utilisations (environs 2,5 millions). DBnary contient également plus de 4,6 millions de traductions à partir des 21 langues extraites vers plus de 1500 langues cibles différentes. De plus, DBnary contient des relations lexico-sémantiques (syno / antonymie, hypo / hyperonymie, mero / holonymie et troponymie).

Le tableau 1 présente les données concernant les langues utilisées dans les expériences présentées dans cet article. D'autres données sont disponibles sur le site Internet de DBnary [4].

3. Bulgare, néerlandais, anglais, finnois, français, allemand, grec (moderne), indonésien, italien, japonais, latin, lituanien, malgache, norvégien, polonais, portugais, russe, serbe, croate, espagnol, suédois et turc

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

L'extraction des formes lemmatisées pour l'utilisation de DBnary est fondée sur le module de *TreeTagger* (Schmid, 1995). Ce dernier nous permet de trouver les synsets correspondant aux entrées lemmatisées de la ressource lexico-sémantique.

## 2.3 Représentations distribuées de mots

### 2.3.1 Aperçu

Les *« word embeddings »* ou représentations distribuées de mots sont une représentation des mots dans un espace continu et un sujet d'étude particulièrement actif ces dernières années (Bengio *et al.*, 2003; Turian *et al.*, 2010; Collobert *et al.*, 2011; Huang *et al.*, 2012).

L'idée principale est que la représentation d'un mot peut être obtenue en fonction de son contexte (les mots qui l'entourent) (Baroni & Zamparelli, 2010). Les mots sont projetés vers un espace continu de dimension prédéfinie et les mots ayant des contextes similaires sont, de fait, proches dans cet espace continu. Comparées à une représentation discrète, les représentations distribuées permettent d'induire des relations syntaxiques et sémantiques (Blacoe & Lapata, 2012; Mikolov *et al.*, 2013b). Par ailleurs, des similarités entre les mots peuvent être calculées en utilisant les représentations vectorielles via, par exemple, une distance de type *cosinus*.

Dans nos expériences, nous proposons d'utiliser une des représentations les plus employée : celle proposée par Mikolov *et al.* (2013a). Dans cette représentation, deux modèles (CBOW et SKIP-GRAM) sont proposés pour apprendre les relations entre les mots selon leur contexte (Mikolov *et al.*, 2013b).

### 2.3.2 Application à l'évaluation des systèmes en TA

L'utilisation des représentations distribuées de mots en Traitement Automatique de la Langue Naturelle (TALN) s'est répandue depuis les travaux de Collobert *et al.* (2011) et plus récemment de Mikolov *et al.* (2013a,b). Des applications comme la traduction automatique statistique (Cho *et al.*, 2014), la recherche d'informations (Shen *et al.*, 2014), les systèmes de questions-réponses (Belinkov *et al.*, 2015) et bien d'autres utilisent ces représentations distribuées.

Concernant l'évaluation en traduction automatique, Gupta *et al.* (2015) a proposé une métrique qui utilise des modèles de languages neuronaux sous forme d'arbres de dépendances pour mettre en relation la référence et l'hypothèse de traduction. Le score est prédit en utilisant une similarité calculée à l'aide d'un réseau de neurones. Dans le même temps, Vela & Tan (2015) ont utilisé une représentation distribuée de documents (« document embeddings ») pour prédire la qualité de traduction avec la notion d'adéquation. Ces travaux sont proches de ceux présentés dans cet article mais ils sont liés à un apprentissage de la métrique d'évaluation. Cet apprentissage remet en cause la portabilité de la métrique vers d'autres domaines ou tâches. Dans nos travaux, nous utilisont des représentations apprises sur des corpus du même type que l'évaluation (discours et informations), cependant celles-ci peuvent être apprises sur des données plus généralistes comme le corpus Wikipédia.

D'autres travaux proche des nôtres ont été proposés par Zou *et al.* (2013), ces derniers utilisent des représentation vetorielles bilingues des mots pour détecter des similarités pour l'alignement. Cette information est utilisée comme paramètre supplémentaire dans un système de traduction automatique fondé sur les paires de segments. Enfin, Banchs *et al.* (2015) utilisent une représentation continue de la proximité sémantique entre deux phrases dans une métrique mais cette représentation est fondée sur l'analyse sémantique latente (Salton *et al.*, 1975). Enfin, les travaux de Ng & Abrecht (2015) utilisent des représentations distribuées afin d'étendre une métrique pour le résumé automatique : ROUGE.

Au vu de ces multiples travaux, notre proposition est d'enrichir la métrique d'évaluation METEOR (l'une des métriques automatiques correlant le mieux avec les jugements humains), avec un module utilisant les représentations distribuées de mots afin de détecter des correspondances approchées (floues) entre les mots. Cette concordance floue sera fondée sur un score dérivé de la similarité cosinus (*cosine similarity*) et normé entre 0 et 1. Un seuil de décision est nécessaire pour décider si un mot de l'hypothèse peut être apparié (considéré comme similaire) à un mot de la référence. Un seuil par défaut a été fixé à 0,80 afin de n'apparier que des mots suffisament proches selon leur représentation distribuée.

# 3 Evaluation d'un METEOR enrichi

## 3.1 Données et protocole

L'évaluation de la corrélation de notre METEOR avec les jugements humains est faite en utilisant les données de la tâche *Metrics* de WMT14 (Machacek & Bojar, 2014). Ces données permettent d'évaluer la corrélation entre une métrique automatique et le jugement humain pour les tâches de traductions de l'anglais vers le français, l'allemand, le russe et vice-versa. Dans nos expériences, la corrélation est mesurée au nivau du segment uniquement en utilisant le coefficient $\tau$ de corrélation de Kendall. Cette corrélation prend en compte le classement du système, au niveau de la phrase, sur la base de jugements humains et le score de METEOR (nous allons utiliser nos métriques enrichies et les comparer au *METEOR Baseline*).

L'amélioration de METEOR est faite suivant deux axes. Le premier est l'utilisation des représentations distribuées de mots *à la place* des données lexicales (i.e. : le module *Vecteurs* remplace les modules *Stem* et *Synonymes*, soit respectivement les modules 2 et 3). Le second est l'utilisation des représentations distribuées de mots *en complément* des données lexicales.

Le module *Synonymes* du score que nous nommerons *METEOR Baseline* utilise des données de WordNet uniquement pour l'anglais. Nous proposons de l'étendre aussi à l'utilisation de DBnary pour l'anglais, le français, l'allemand et le russe. Par ailleurs, le module *Vecteurs* utilise des représentations distribuées de mots appris sur les corpus présentés dans le tableau 3. Dans le score METEOR enrichi, tous ces modules utilisent le même poids que le module *Synonyme* dans chaque langue.

En résumé, nos résultats sont obtenus avec plusieurs configurations différentes de METEOR chacune correspondant à une combinaison des modules :

- *METEOR Baseline* : calcul du score de METEOR avec les modules *Exact*, *Stem*, *Synonymes* et *Paraphrase* classiques pour l'anglais et utilisation des modules *Exact*, *Stem* et *Paraphrase* pour les autres langues ;
- *METEOR DBnary* : calcul du score de *METEOR Baseline* avec le module *Synonymes* utilisant les données issues de DBnary plutôt que celles issues de WordNet ;
- *METEOR Vecteurs* : les modules *Stem* et *Synonymes* sont remplacés par le module *Vecteurs* ;
- *METEOR Baseline* + *Vecteurs* : le *METEOR Baseline* est enrichi avec le module *Vecteurs* ;
- *METEOR DBnary* + *Vecteurs* : le *METEOR DBnary* est enrichi avec le module *Vecteurs*.

## 3.2 *METEOR DBnary*

Comme mentionné dans la partie 2.1, le module *Synonyme* de METEOR utilise le dictionnaire de synsets de WordNet, qui est une ressource riche de 117K entrées pour l'anglais. Les ressources externes pour cette version de METEOR enrichie ont été extraites à partir de la base de données DBnary (Sérasset, 2012).

Utilisant des requêtes SPARQL sur le serveur de DBnary [5], les relations de synonymie dans la base de données en anglais, français, russe et allemand ont été extraites. Le résultat de l'extraction est dans le format suivant : $lemme \rightarrow Synonyme$. Ensuite, ces données ont été mises au même format que WordNet (qui est déjà compatible avec METEOR). Ce traitement consiste à attribuer un identifiant pour chaque lemme et à construire une liste de synonymes pour chaque lemme sous format ID. Par exemple : $lemme \rightarrow ID\_Syn_1,\ ID\_Syn_2,\ ID\_Syn_3$.

Afin de mesurer la différence entre l'utilisation des synsets issus de DBnary (noté *METEOR DBnary*) et ceux issus de WordNet (noté *METEOR Baseline*), nous avons utilisé les sorties de cinq des systèmes de la campagne d'évaluation de la tâche de traduction de WMT14 (Bojar *et al.*, 2014) sur la traduction du français vers l'anglais.

| Métrique | Systèmes : | | | | |
|---|---|---|---|---|---|
| | 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 |

TABLE 2 – Tableau présentant les scores de METEOR obtenus sur le corpus *newstest* de l'évaluation WMT14 pour la tâche de traduction du français vers l'anglais. Sont comparées les données lexicales issues de WordNet (*METEOR Baseline*) avec celles de DBnary sur l'anglais (*METEOR DBnary*).

Les résultats du tableau 2 présentent les scores obtenus avec *METEOR DBnary* et *METEOR Baseline*. Les scores obtenus par *METEOR DBnary* sont au dessus de ceux obtenus avec le *METEOR Baseline* d'environs 0,7 points. En d'autres termes, la ressource lexico-sémantique DBnary permet d'obtenir plus de correspondances que WordNet, malgré le fait que le nombre de synsets de cette dernière est 3,3 fois plus grand que celui de DBnary. Ceci est notamment dû au fait que Wordnet ne propose des synonymes que pour les noms, les verbes, les adjectifs et les adverbes, tandis que DBnary les propose pour un ensemble plus large de catégories morpho-syntaxiques.

## 3.3 *METEOR Vecteurs*

Comme proposé dans la partie 2.3.2, nous proposons de remplacer les données lexicales par des données issues des représentations distribuées de mots. Dans cette approche, les mots sont représentés sous forme de vecteurs dont les valeurs sont extraites automatiquement à partir d'un corpus (voir partie 2.3).

| Langue | corpus | # de lignes | # de mots |
|---|---|---|---|
| anglais | Europarl V7 + news commentary V10 | 2,2 M | 60 M |
| français | Europarl V7 + news commentary V10 | 2,2 M | 67 M |
| allemand | Europarl V7 + news commentary V10 | 2,1 M | 57 M |
| russe | Common Crawl + news commentary V10 + News Crawl 2013 (10%) | 2,4 M | 50 M |

TABLE 3 – Corpus utilisés pour apprendre les représentations distribuées de mots pour chaque langue.

Contrairement aux modules lexicaux classiques, nous utilisons les scores de similarité (dérivés de la similarité *cosinus*) pour déterminer s'il y a correspondance ou non entre deux mots. Ce score de similarité est normé et est seuillé : s'il est supérieur au seuil nous considérons qu'il y a appariement (*match*), tandis que dans le cas contraire, nous considérons que les mots sont différents. Le seuil de

5. http://kaiko.getalp.org/about-dbnary/online-access/

décision oracle est trouvé de manière empirique. Dans nos expériences, nous proposons d'utiliser un seuil par défaut fixé arbitrairement à 0,80 ainsi qu'un seuil oracle trouvé sur les données de l'évaluation *metrics* de WMT14 (Machacek & Bojar, 2014).

Le tableau 3 présente les corpus utilisés pour chaque langue. Concernant le russe, nous avons décidé d'ajouter 10% du corpus monolingue russe News Crawl 2013 (choisis aléatoirement) afin d'atteindre une quantité de corpus d'apprentissage équivalent aux autres langues. Les représentations distribuées de mots utilisés ont été apprises avec un modèle CBOW et sont d'une de taille 50, grâce à la boite à outil *Word2Vec* (Mikolov *et al.*, 2013b).

| Métrique | Systèmes : | | | | |
|---|---|---|---|---|---|
| | 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 Vecteurs* | 37,00 | 37,34 | 31,87 | 33,67 | 32,34 |
| *METEOR Baseline + Vecteurs* | 37,08 | 37,40 | 31,96 | 33,75 | 32,45 |
| *METEOR DBnary + Vecteurs* | 37,53 | 37,88 | 32,60 | 34,32 | 33,05 |

TABLE 4 – Tableau présentant les scores de METEOR obtenus sur le corpus *newstest* de l'évaluation WMT14 (*fr-en*) et comparant les données lexicales issues de WordNet et de DBnary avec les représentations distribuées de mots. Le module *Vecteurs* utilise le seuil de décision par défaut (0,80).

Dans nos expériences, nous remplaçons donc le module *Synonymie* et le module *Stem* par le module *Vecteurs*. Ce dernier permet d'approximer les relations syntaxiques et sémantiques grace aux représentations de mots (voir partie 2.3). Les résultats présentés dans le tableau 4 montrent que l'utilisation des représentations distribuées de mots permet d'augmenter mécaniquement le score de METEOR et donc les appariements entre mots de la sortie de TA et la référence. Nous allons voir dans la section suivante si cette augmentation du score METEOR s'accompagne d'une meilleure corrélation avec les jugements humains.

# 4 Corrélations entre METEOR enrichi et jugement humain

## 4.1 Résultats expérimentaux

Lors de ces expériences, nous présentons les résultats obtenus avec le module *Vecteurs* selon deux seuils : le seuil par défaut (0,80) et les seuils oracles qui maximisent les scores de corrélation avec le jugement humain selon les différentes configurations et les paires de langues utilisées.

Le tableau 5 présente les résultats de la corrélation avec le jugement humain sur les données d'évaluation issues de la campagne WMT2014 (Machacek & Bojar, 2014). L'évaluation est sur trois paires de langues : français–anglais (fr–en), allemand–anglais (de–en) et russe–anglais (ru–en) dans les deux sens de traduction. Le français, l'allemand et le russe ont été choisis pour simuler une difficulté croissante sur les variabilités des formes de surfaces d'un même mot. L'anglais en langue cible sert ici de référence dans la comparaison entre les deux ressources lexicales.

On observe que le *METEOR Baseline* et la nouvelle version de *METEOR Vecteurs* avec le seuil de décision par défaut à $0,80$ obtiennent des résultats relativement proches en moyenne. *METEOR DBnary* obtient un meilleur score de corrélation que la version *Baseline* ($+0,003$ points). Lorsque nous combinons les informations lexico-sémantiques et les représentations distribuées de mots (module

*Vecteurs*), le score moyen obtenu avec l'anglais comme langue cible par *METEOR Baseline+Vecteurs* est amélioré de $0,005$ points par rapport au score de référence. Si la combinaison est faite avec les données lexico-sémantiques de DBnary (*METEOR DBnary+Vecteurs* ), l'amélioration du score de référence atteint $0,006$ points.

Dans un deuxième temps, nous proposons de trouver le seuil de décision oracle du module *Vecteurs* pour la tâche de corrélation avec le jugement humain. Pour la version *METEOR Vecteurs*, cette optimisation améliore de $0,001$ points par rapport au seuil par défaut. La combinaison *METEOR Baseline + Vecteurs* reste identique en terme de performance moyenne, même si les scores pour les paires de langues sont différents. Enfin la dernière combinaison (*METEOR DBnary + Vecteurs*) améliore le coefficient de corrélation de $0,002$ points.

| Langues | fr–en | | de–en | | ru–en | | Moyenne | |
|---|---|---|---|---|---|---|---|---|
| Métrique | Seuil | $\tau$ | Seuil | $\tau$ | Seuil | $\tau$ | Seuil | $\tau$ |
| *METEOR Baseline* | – | 0,411 | – | 0,323 | – | 0,329 | – | 0,354 |
| *METEOR DBnary* | – | 0,408 | – | 0,334 | – | 0,328 | – | 0,357 |
| *METEOR Vecteurs* | 0,80 | 0,409 | 0,80 | 0,332 | 0,80 | 0,325 | 0,80 | 0,355 |
| *METEOR Baseline + Vecteurs* | 0,80 | 0,414 | 0,80 | 0,334 | 0,80 | 0,327 | 0,80 | 0,359 |
| *METEOR DBnary + Vecteurs* | 0,80 | 0,409 | 0,80 | 0,339 | 0,80 | 0,333 | 0,80 | 0,360 |
| *METEOR Vecteurs* | 0,90 | 0,410 | 0,78 | 0,335 | 0,70 | 0,328 | 0,79 | 0,357 |
| *METEOR Baseline + Vecteurs* | 0,80 | 0,414 | 0,78 | 0,338 | 0,89 | 0,330 | 0,82 | 0,360 |
| *METEOR DBnary + Vecteurs* | 0,78 | 0,412 | 0,80 | 0,339 | 0,78 | 0,335 | 0,78 | 0,362 |
| Langues | en-fr | | en-de | | en-ru | | Moyenne | |
| Métrique | Seuil | $\tau$ | Seuil | $\tau$ | Seuil | $\tau$ | Seuil | $\tau$ |
| *METEOR Baseline* | – | 0,280 | – | 0,238 | – | 0,427 | – | 0,315 |
| *METEOR DBnary* | – | 0,284 | – | 0,239 | – | 0,435 | – | 0,319 |
| *METEOR Vecteurs* | 0,80 | 0,282 | 0,80 | 0,235 | 0,80 | 0,420 | 0,80 | 0,312 |
| *METEOR Baseline + Vecteurs* | 0,80 | 0,280 | 0,80 | 0,238 | 0,80 | 0,428 | 0,80 | 0,315 |
| *METEOR DBnary + Vecteurs* | 0,80 | 0,281 | 0,80 | 0,238 | 0,80 | 0,436 | 0,80 | 0,318 |
| *METEOR Vecteurs* | 0,72 | 0,287 | 0,70 | 0,236 | 0,90 | 0,421 | 0,76 | 0,314 |
| *METEOR Baseline + Vecteurs* | 0,85 | 0,287 | 0,88 | 0,240 | 0,83 | 0,429 | 0,85 | 0,318 |
| *METEOR DBnary + Vecteurs* | 0,91 | 0,285 | 0,77 | 0,241 | 0,84 | 0,437 | 0,82 | 0,321 |

TABLE 5 – Corrélations au niveau segment entre nos différentes versions de METEOR et des jugements humains (données WMT14). Les scores obtenus avec le module *vecteur* sont présentés d'abord avec un seuil de décision par défaut (0,80) puis viennent en dessous de la ligne en pointillé, les scores obtenus avec les seuils oracles.

Dans ce même tableau, sont présentés les résultats de la corrélations des scores de traduction avec le jugement humain des différentes combinaisons dans le sens anglais vers français, allemand et russe. Lors de ces expériences, le module *Synonyme* de *METEOR Baseline* est désactivé, car il n'y a pas de données lexico-sémantiques pour ces langues dans cette version. *METEOR DBnary* obtient les résultats attendus : une meilleure corrélation avec le jugement humain pour les trois langues en comparaison avec *METEOR Baseline*. Utilisant le seuil par défaut, *METEOR Vecteurs* améliore les résultats par rapport à *METEOR Baseline* seulement sur le français. Concernant la combinaison entre représentations distribuées (seuil par défaut) et *METEOR Baseline* ou *METEOR DBnary*, les résultats sont plus mitigés sauf lorsque la langue cible est le russe. Cependant, les scores de corrélation obtenus par ces combinaisons sont en moyenne améliorés de $0,003$ points lorsque le module *Vecteurs* utilise le seuil oracle.

## 4.2 Analyse et discussion

Les scores de corrélation obtenus avec les métriques enrichies tendent à suggérer que les représentations distribuées restent moins performantes que les ressources lexico-sémantiques pour l'évaluation en traduction automatique mais peuvent néanmoins apporter un complément d'information intéressant

à ces dernières, ou être utiles lorsqu'aucune ressource lexicale en langue cible n'est disponible.

Si on considère les scores de corrélations moyens obtenus sur l'anglais et sur le français, les configurations *METEOR Vecteurs* et *METEOR DBnary* sont comparables. Sur les autres langues cibles, les corrélations obtenues par *METEOR Vecteurs* ne sont pas au niveau de *METEOR DBnary*. Les représentations distribuées pour les mots semblent apparier péniblement les mots de l'hypothèse avec ceux de la référence.

En revanche, lorsqu'on combine les données lexicales avec le module *Vecteurs* (*METEOR DBnary + Vecteurs*), on observe une légère augmentation du score de corrélation, notamment lorsque le seuil est optimal comparé à l'utilisation des données lexicales ou des représentations distribuées seules.

### 4.2.1 Exemples d'alignements avec notre métrique enrichie

Afin d'illustrer les correspondances obtenues, nous présentons deux exemples représentatifs issus de l'évaluation. Dans ces exemples, nous présentons les alignements obtenus avec notre métrique enrichie : *METEOR DBnary + Vecteurs*.

L'exemple présenté tableau 6 montre une sortie issue du système *rbmt 1* soumis lors de l'évaluation WMT14. *METEOR baseline* n'a trouvé que les alignements avec les formes de surface identiques (les mots et signes : *« qu' », « il », « est »*, etc.). Le module *Synonyme* utilisant les données de DBnary permet de trouver une correspondance entre les couples de mots *« employées » – « utilisées »* et *« pour » – « dans »*. Enfin, le module *Vecteurs* indique que les mots *« pense »* et *« estime »* sont contextuellement proches, de même que les mots *« je »* et *« j' »*.

| | |
|---|---|
| Hypothèse : | je pense qu' il est concevable que ces données soient employées pour le bénéfice mutuel . |
| Référence : | j' estime qu' il est concevable que ces données soient utilisées dans leur intérêt mutuel . |

TABLE 6 – Premier exemple tiré du système *rbmt 1* évalué avec la combinaison *METEOR DBnary + Vecteurs*. Les relations détectées avec la ressource lexicale DBnary sont encadrées en trait continu tandis que celles obtenues grâce aux représentations distribuées sont en pointillé.

Lorsque l'exemple est seulement évalué avec *METEOR Vecteurs*, les mots *« employées »* et *« utilisées »* sont également appariés avec le seuil de décision par défaut (0,80). En revanche, les mots *« bénéfice »* et *« intérêt »* ne sont appariés par le module *Vecteurs* que si le seuil de décision est abaissé à 0,75.

| | |
|---|---|
| Hypothèse : | le créateur de SAS disait il faisait un genre du feuilleton géopolitique . |
| Référence : | le père de SAS disait faire un genre de feuilleton géopolitique . |

TABLE 7 – Second exemple évalué avec la combinaison *METEOR DBnary + Vecteurs*.

Dans le second exemple présenté tableau 7, l'hypothèse est fournie par le système *rbmt 4*. Comme dans le précédent exemple, les correspondances trouvées avec le module *Synonymie* utilisant les données lexicales de DBnary (encadrées par un trait continu) sont complétées par celles trouvées par le module *Vecteurs* (encadrées par un trait en pointillé). Les formes de surfaces identiques trouvées grâce au module *Exact* ne sont pas signalées ici. Les correspondances trouvées à l'aide du module *Synonymes* utilisant DBnary sont les couples *« créateur » – « père »* et *« faisait » – « faire »*. Les

mots *« du »* et *« de »* sont appariés grâce au module *Vecteurs*. Comme dans l'exemple précédent, les alignements trouvés avec le module *Synonymes* peuvent être trouvés avec le module *Vecteurs* mais moyennant un abaissement du seuil de détection (0,60).

Ces exemples illustrent la complémentarité entre les ressources lexicales et les représentations distribuées de mots. Ces dernières peuvent permettre d'apparier des mots importants (comme *« pense »* et *« estime »* dans notre premier exemple), mais également des mots outils (comme *« du »* et *« de »* dans notre second exemple). Pour ce type de métrique automatique, l'importance linguistique des mots et leurs rôles dans la phrase n'étant pas pris en compte, tout appariement pouvant être correctement établi étant « bon à prendre ».

### 4.2.2 Limitations des représentations distribuées de mots

Intuitivement, nous pouvons dénombrer deux limitations pour ces représentations. La première est liée au mots inconnus et la seconde à la qualité du modèle appris (qui rejoint le premier point).

D'une part, la limitation liée au mots inconnus semble être un frein potentiellement important. En effet, l'espace des représentations distribuées repose sur un apprentissage à partir d'un corpus monolingue. Au cours de l'évaluation, tout mot de l'hypothèse inconnu de ce corpus initial n'aura pas de représentation et ne pourra donc pas être apparié avec les mots de la référence.

D'autre part, la qualité du modèle dépend du corpus d'apprentissage. Il est donc recommandé que le corpus d'apprentissage des représentations distribuées soit relativement proche des données d'évaluation ou tout au moins, que sa couverture lexicale soit la plus importante possible.

# 5 Conclusion & Perspectives

Dans cet article, nous avons proposé de comparer les mesures de similarité entre phrases fondées sur des représentations distribuées avec des mesures de similarité fondées sur des réseaux lexico-sémantiques. Plus précisément, nous avons proposé d'étendre la métrique METEOR utilisée pour l'évaluation de la traduction automatique. Malgré les réserves qu'on peut émettre sur les fondations linguistiques de ces représentations distribuées, nous avons montré qu'elles peuvent apporter un complément non-négligeable aux données lexico-sémantiques. D'après nos expériences, l'association entre ces deux types d'information est prometteuse pour l'évaluation automatique en traduction automatique. De plus, les observations faites sur le français se retrouvent sur l'anglais, l'allemand et le russe. La complémentarité des approches semble donc exister sur plusieurs langues.

Nous prévoyons d'étendre ces techniques d'appariement au-dela du mot et de prendre en compte un appariement au niveau du segment, voire de la phrase. Nous envisageons également d'étudier le comportement des représentations distribuées lorsqu'on applique des pré-traitements classiques comme la suppresion des mots vides, par exemple. De plus, nous étudierons l'adaptation de ces approches sur d'autres métriques automatiques utilisant des ressources lexicales (e.g. : TER-Plus). Enfin, nous envigeons d'utiliser les représentations distribuées de mots dans d'autres domaines comme l'estimation de mesures de confiance. Par exemple, en traduction automatique, nous souhaiterions mesurer les corrélations existantes entre des mesures de confiance enrichies avec des temps de post-édition de traductions.

L'outil, les données et les modèles présentés dans cet article sont mis en ligne [6] afin de faciliter la reproductibilité des expériences réalisées.

6. `http://site-anonyme.com`

# Remerciements

Anonymisés

# Références

APIDIANAKI M. & MARIE B. (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)*.

BANCHS R. E., D'HARO L. F. & LI H. (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.

BANERJEE S. & LAVIE A. (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*.

BARONI M. & ZAMPARELLI R. (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*, p. 1183–1193.

BELINKOV Y., MOHTARAMI M., CYPHERS S. & GLASS J. (2015). VectorSLU : A Continuous Word Vector Approach to Answer Selection in Community Question Answering Systems. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval*, volume 15.

BENGIO Y., DUCHARME R., VINCENT P. & JANVIN C. (2003). A neural probabilistic language model. *The Journal of Machine Learning Research*, **3**, 1137–1155.

BLACOE W. & LAPATA M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, p. 546–556.

BOJAR O., BUCK C., FEDERMANN C., HADDOW B., KOEHN P., LEVELING J., MONZ C., PECINA P., POST M., SAINT-AMAND H., SORICUT R., SPECIA L. & TAMCHYNA A. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, p. 12–58, Baltimore, Maryland, USA.

CHIARCOS C. & SUKHAREVA M. (2015). Olia – ontologies of linguistic annotation. *Web Semantics : Science, Services and Agents on the World Wide Web*, **6**(4), 379–386.

CHO K., VAN MERRIENBOER B., GULCEHRE C., BAHDANAU D., BOUGARES F., SCHWENK H. & BENGIO Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1724–1734, Doha, Qatar.

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.

COLLOBERT R., WESTON J., BOTTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, **12**, 2493–2537.

DENKOWSKI M. & LAVIE A. (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*, p. 250–253.

DENKOWSKI M. & LAVIE A. (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*, p. 339–342.

DENKOWSKI M. & LAVIE A. (2014). METEOR Universal : Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

DODDINGTON G. (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*.

ELLOUMI Z., BLANCHON H., SERASSET G. & BESACIER L. (2015). METEOR for Multiple Target Languages using DBnary. In *Proceedings of MT Summit 2015*, Miami, Florida, USA.

FARUQUI M., DODGE J., JAUHAR S. K., DYER C., HOVY E. H. & SMITH N. A. (2014). Retrofitting word vectors to semantic lexicons. *CoRR*.

FELLBAUM C. (1998). *WordNet : An Electronic Lexical Database*. MA : MIT Press.

GUPTA R., ORASAN C. & GENABITH J. V. (2015). Machine Translation Evaluation using Recurrent Neural Networks. In *Proceedings Workshop on Machine Translation (WMT), Metrics Shared Task*, Lisbonne, Portugal.

HUANG E. H., SOCHER R., MANNING C. D. & NG A. Y. (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*, p. 873–882.

KLYNE G. & CARROLL J. J. (2004). *Resource Description Framework (RDF) : Concepts and Abstract Syntax*. Rapport interne.

LE Q. V. & MIKOLOV T. (2014). Distributed Representations of Sentences and Documents. In *Proceedings of The 31st International Conference on Machine Learning*.

LUONG T., PHAM H. & MANNING C. D. (2015). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, p. 151–159.

MACHACEK M. & BOJAR O. (2014). Results of the WMT14 Metrics Shared Task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, p. 293–301, Baltimore, Maryland, USA.

MCCRAE J., SPOHR D. & CIMIANO P. (2011). *The Semantic Web : Research and Applications : 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 2011, Proceedings, Part I*, chapter Linking Lexical Resources and Ontologies on the Semantic Web with Lemon, p. 245–259.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient Estimation of Word Representations in Vector Space. In *The Workshop Proceedings of the International Conference on Learning Representations (ICLR) 2013*, Scottsdale, Arizona, USA.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, p. 3111–3119. Curran Associates, Inc.

MIKOLOV T., YIH W.-T. & ZWEIG G. (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*, p. 746–751, Atlanta, Georgia.

MORO A., RAGANATO A. & NAVIGLI R. (2014). Entity linking meets word sense disambiguation : a unified approach. *Transactions of the Association for Computational Linguistics*, **2**, 231–244.

NG J. & ABRECHT V. (2015). Better Summarization Evaluation with Word Embeddings for ROUGE. In *In The Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a Method for Automatic Evaluation of Machine Translation. In *ACL*.

SALTON G., WONG A. & YANG C. S. (1975). A Vector Space Model for Automatic Indexing. *Commun. ACM*, **18**(11), 613–620.

SCHMID H. (1995). Treetagger| a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, **43**, 28.

SÉRASSET G. (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.

SHEN Y., HE X., GAO J., DENG L. & MESNIL G. (2014). A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, p. 101–110.

SNOVER M., DORR B., SCHWARTZ R., MICCIULLA L. & MAKHOUL J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, p. 223–231.

SNOVER M., MADNANI N., DORR B. J. & SCHWARTZ R. (2009). Fluency, adequacy, or HTER ? : exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, p. 259–268.

TURIAN J., RATINOV L. & BENGIO Y. (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*, p. 384–394.

VELA M. & TAN L. (2015). Predicting Machine Translation Adequacy with Document Embeddings. In *Proceedings Workshop on Machine Translation (WMT), Metrics Shared Task*, Lisbonne, Portugal.

ZOU W. Y., SOCHER R., CER D. M. & MANNING C. D. (2013). Bilingual Word Embeddings for Phrase-Based Machine Translation. In *EMNLP*, p. 1393–1398.