République Algérienne Démocratique Et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université d'Adrar
Faculté des Sciences et de la Technologie
Département des Mathématiques et Informatique

Mémoire de Fin d'Etude

Pour l'Obtention du Diplôme de Master en Informatique

Option: Réseaux et Systèmes Intelligents

Thème :

Implémentation et étude comparative des Métriques d'évaluation dont le cas d'une traduction Automatique d'un texte de l'Arabe vers l'Anglais.

Présenté Par:

BENCHIKH MEBARKA

Encadré par :

MR. CHERAGUI MOHAMED AMINE

Année Universitaire 2013/2014

Résumé

L’une des tâches les plus difficile dans le domaine du traitement automatique du langage naturel est l’évaluation d’une application TALN (traducteur automatique, correcteur orthographique, résumeur automatique, ….etc.), et comme toute science il est nécessaire de le démontrer par valeur numérique, afin de vérifier mais aussi prouver de combien cette nouvelle application est mieux par rapport à ces prédécesseurs.

Le problème avec la langue naturelle, est que cette dernière n'est pas exact de la même façon que sont les modèles mathématiques et les théories de la physique. La langue naturelle a un certain degré d'imprécision (ambiguïté), ce qui rend difficile de mettre des chiffres objectifs à une application TALN.

Considéré comme étant l’axe de recherche vital du traitement automatique du langage naturel, la traduction automatique est toujours confrontée à ce problème d’évaluation, qui reste un défi scientifique à relever afin de juger la qualité de n’importe quel traducteur automatique, pour améliorer la performance du traducteur mais aussi pour guider les consommateurs dans leur prise de décision en termes d’investissement.

Notre but à travers ce travail est de proposer une boite à outils d’évaluation pour la traduction automatique (Arabe / Anglais), en utilisant des métriques d’évaluation automatique.

Abstract

One of the most difficult task in the field of Natural Language Processing (NLP) is evaluating a software application (machine translator, spell checker, Automatic Summarizer, etc). Like any scientist is necessary to demonstrate it by numerical value, but also to verify how this new application is better compared to its predecessors.

The problem with natural language is not the same way as they are in mathematical models and theories of physics. Natural language has a degree of imprecision (ambiguity), which makes it difficult to provide objective figures to a Software application.

Considered as being a vital research axis of an automatic processing of natural language, machine translation is still facing this problem of evaluation, which remains a scientific challenge to be reached before judging the quality of any machine translator, to improve the performance of the translator but also to guide consumers in their decision making in terms of investment.

Our goal in this work is to propose à tool box evaluation for machine translation (from Arabic to English) using automatic metrics

Remerciements

Tout d'abord nous remercions Dieu Tout Puissant de nous avoir donné la force, la volonté, et le privilège d'étudier et de suivre le chemin de la science;

ensuite nous remercions les membres du jury, d'avoir accepté de porter un jugement sur ce travail

Nous tenons également à remercier plus particulièrement :

Mon encadreur Mr CHERAGUI M. Amine.

Tous les professeurs du département d'informatique, à qui l'on doit tout le respect.

Surtout Mr. Omari Mohamed

pour l'aide précieuse.

Sans oublier les deux professeurs BAADID Abednasser et KHLEIFI Moustapha

Nos remerciements vont également à toutes les personnes qui ont contribué de près ou de loin à l'élaboration de ce mémoire.

Merci à tous

Dédicaces

Je dédie ce modeste travail à :

Mes parents pour leur soutien tout au long de mon cursus universitaire.

Mon encadreur Mr.CHERAGUI Mohamed Amine

Mes frères et sœurs

Ma famille

Mes amis

Et à tous mes camarades de promotion.

BencheikhMebarka

Table des matières

Listes des Figures

Liste des Tableaux

Introduction générale

Après la réalisation et le développement d'un traducteur, il faut passer à l'étape d'évaluation de ce traducteur qui n'est pas une tâche facile à réaliser. A ce niveau on doit faire face à un problème épineux qui est l'ambigüité de la langue naturelle, celle-ci signifie qu'il peut exister plusieurs traductions possibles d'une phrase source donnée. Ces possibilités de traduction se différent selon le choix des mots utilisés et l'ordre choisi de ces même mots. Ce qui nous mène à se poser deux questions :

1. C'est quoi une bonne ou une mauvaise traduction ?
2. Comment distinguer la bonne traduction de la mauvaise ?

Traditionnellement, l'examen de la qualité de la traduction est fait par un être humain. Nous l'appelons l'évaluation humaine ou l'évaluation subjective. Ce type d'évaluation ce fait par une personne expert métrisant bien la langue source et la langue cible, et pour déterminer la qualité de la traduction il doit effectuer la comparaison entre le texte traduit avec le texte original en respectant les deux normes suivantes :

1. la norme de la portée de la concordance grammaticale du texte avec les règles de la langue souhaitée.

2. le degré de conservation du sens des mots entre la langue originale et de la langue souhaitée.

L'évaluation humaine est coûteuse cependant. De plus, elle a le défaut d'être non reproductible puisque il existe une variabilité entre les annotateurs évaluateurs. C'est pourquoi plusieurs méthodes de mesure automatique ont vues les jours, afin de simuler l'être humain dans son évaluation de la qualité de la traduction, c'est ce qu'on appelle une évaluation automatique ou l'évaluation objective. Ce type d'évaluation ne nécessite quasiment aucune intervention humaine a posteriori (mais il faut un corpus de test bilingue a priori).

Dans cette optique, ce mémoire porte sur l'implémentation et l'étude comparative des Métriques d'évaluation dont le cas d'une traduction d'un texte de l'Arabe vers l'Anglais.

Les systèmes de traduction qui sont évalué sont : Systran, Babylon, Microsoft, Reverso, et Google traduction. On cherche à savoir comment évaluer un système de TA à l'aide des métrique automatique.

Ce mémoire s'articule sur les points suivants :

- ❖ Chapitre1 : présente la traduction automatique en générale avec une brève historique, et les approche utilisées puis en se concentrant particulièrement sur la traduction Arabe vers Anglais.

- ❖ Chapitre 2 : Présente les déférentes approches d'évaluation : entre évaluation humaine qui est fait par les êtres humains (points forts et faiblesses) et l'évaluation automatique adoptant des mesures automatiques (Principes et techniques).

- ❖ Chapitre 3 : sera consacré à la description de l'architecture générale de notre boite à outils d'évaluation « EvaMT » en précisant les différents modules.

- ❖ Chapitre 4 : est vu comme le point d'ancrage de notre travail, où on va constater les résultats obtenus de l'implémentation des cinq métriques d'évaluation BLEU, NIST, METEOR, TER, et WER. En utilisant à la fois plusieurs traducteurs automatiques et un Corpus d'évaluation de référence.

Chapitre 1

Introduction à la Traduction automatique

1. Introduction

Si par le passé, faire traduire un texte par une machine relevait de l'utopie, les avancées technologiques et les efforts constants de nombreux chercheurs en Traitement Automatique du Langage Naturel (TALN) font que cela paraît aujourd'hui possible. Il faut dire que la Traduction Automatique (T.A) est considérée comme étant le champ d'étude phare du TALN vu les opportunités qui se présentent que se soient sur le plans scientifique et même socio économique (l'émergence d'un nombre important de produit et de sociétés spécialisées).La TA désigne le fait de faire traduire un texte d'une langue source vers une langue cible par une machine sans aucune intervention humaine.

Le but de ce chapitre est de donner un aperçu général sur la traduction automatique à travers les points suivants : son historique, les différentes approches développées et bref détour sur la traduction automatique de la langue arabe.

2. Historique

2.1. Première Période (1948-1960)

- **1949 :** Warren Weaver dans son *Memorandum*de 1949 propose les premières idées sur l'utilisation de l'ordinateur dans la traduction, en adoptant l'expression *computer translation*.
- **1952 :** premier colloque de Traduction Automatique, intitulé *Conference on Mechanical Translation*, qui a lieu en juin 1952 au MIT sous la direction de Yehoshua Bar-Hillel.
- **1954 :** La mise au point du premier traducteur automatique (très rudimentaire) par un groupe de chercheurs de l'université de Georgetown en collaboration avec IBM, qui traduit plus de soixante phrases russes en anglais. Les auteurs prétendaient que dans un délai de trois ou cinq ans, la traduction automatique ne serait plus un problème.
- **1954 :** Dans la même année Victor Yngve publie la premire revue sur la traduction automatique intitule « *Mechanical translation – devoted to the translation of languages by the aid of machines* ».

2.2. Deuxième Période (1960-1966)

- **Début des années 1960** C'est l'analyse syntaxique qui est mise en avant comme la seule voie de recherche possible pour faire avancer la traduction automatique. Ainsi il existe déjà de nombreux analyseurs syntaxiques développés à partir de différents modèles de grammaires, comme par exemple la grammaire de dépendance de Tesnière ou la grammaire stratificationnelle de Lamb.
- **1961 :** c'est en février de cette année que la *computational linguistics* voit le jour, grâce aux conférences hebdomadaires organisées par David G. Hays à la Rand Corporation de Los Angeles. Ces conférences seront reprises comme communications lors de la *First International Conference on Machine Translation of Languages and Applied Language Analysis* de Teddington en septembre 1961 avec la participation de linguistes et informaticiens acteur de la traduction comme : Paul Garvin, Sydney M. Lamb, Kenneth E. Harper, Charles Hockett, Martin Kay et Bernard Vauquois.

- **1964 :** la création du comité ALPAC (Automatic Language Processing Advisory Committee) par gouvernement américain pour étudier les perspectives et les chances de la traduction automatique.
- **1966 :** ALPAC publie son célèbre rapport dans lequel elle conclu que les travaux sur la traduction automatique est une perte de temps et d'agent ; les conclusions de ce rapport on eu un impact négatif sur la recherche en Traduction Automatiquepour uncertain nombred'années.

2.3. Troisième Période (1966-1980)

- **1970 :** Lancement du projet REVERSO par un groupe de chercheurs russes [ANN , 2003].
- **1970 :** Développement du Système SYSTRAN[1] (Russe-Anglais) par Peter Toma, qui été à cette époque un membre du groupe de recherche de Georgetown [HUT, 1992].
- **1976 :** Création du système METEO dans le cadre du projet TAUM (Traduction Automatique à l'Université de Montréal) sous la direction d'Alain Colmerauer pour la traduction automatique des prévisions météorologiques destinées au grand public, ce système a été créé par un groupe de chercheurs [1].
- **1978 :** Création du système ATLAS[2] par la firme japonaise FUJITSU, ce traducteur à base de règles est capable de traduire du Coréen au japonais et l'inverse.

2.4. Quatrième Période (1980-1990)

- **1982 :** La firme Japonaise SHARP commercialise son Traducteur automatique DUET (Anglais - Japonais), ce traducteur à base de règle adopte une approche de traduction par transfert [JOH, 1995].
- **1983 :** En tant que géant de l'informatique, NEC développe bien entendu son propre système de traduction, basé sur un algorithme baptisé PIVOT. Commercialisé sous le nom de *HonyakuAdaptor II*, la version grand public du système de traduction de NEC est également basée sur la méthode du pivot, qui consiste à utiliser *Interlingua*.
- **1986 :** Développement du système PENSEE par OKI[3] , qui est un traducteur (Japonais-Anglais) à base de règles [SAY, 1999]**.**

[1]Ce même traducteur à été adopté par la commission européen 1976 pour la traduction (Anglais-français)[HUT, 1992].
[2] Actuellement nous somme dans la version 14 de ce Traducteur.
[3]OKI : Fondé en 1881, Oki Electric Industry Co, est un fabricant japonais de télécommunications où le siège socialest àTokyo, Avec plus de20.000employés dans le monde, Oki est aujourd'hui une entreprise qui fournit aux clients des produitshaut de gammeen terme de technologiespourles systèmesde télécommunication, systèmesd'informations, et les dispositifs électroniques.

- **1986 :** Le groupe Hitachi développe son propre système de traduction à base de règles (dont l'approche adoptée est par transfert), HICATS (Hitachi Computer Aided Translation System / japonais- Anglais) [8].

2.5. Cinquième Période (Depuis 1990)

- **1993 :** Le projet C-STAR (Consortium for Speech Translation Advanced Research) est une coopération internationale. Le thème du projet est la traduction automatique de la parole dans le domaine du tourisme (dialogue client-agent de voyage), en vidéoconférence. Ce projet A donné naissance au système C-STAR I qui traitait trois (03) langues[4] (anglais, allemand et japonais) et a effectué les premières démonstrations transatlantiques trilingues en janvier 1993.
- **1998 :** Commercialisation du traducteur REVERSO par la société Softissimo [ANN, 03].
- **2000 :** Le Développement du système ALPH par le laboratoire Japonais ATR, ce Traducteur (Japonais-Anglais et Chinois - Anglais) adopte une approche à base d'exemples [YVE, 07].
- **2005 :** l'apparition des premiers traducteurs automatiques sur le web (en ligne), comme Google **(http://translate.google.fr/).**

3. Approchesde basede la traduction automatique

Différentes stratégiesont été adoptées pardifférents chercheursàdes moments différents dans l'histoire de traduction automatique. Le choix de la stratégiereflète d'un côté laprofondeurde la diversité linguistiqueetla grandeurdel'ambition d'un autre côté. Le but de cette section est présenter les différentes stratégies ou approches qui ont été mises en place pour accomplir à bien le processus de traduction.AinsiSelon le triangle de Vauquois (Figure 1.1), il existe trois types de stratégie classiquede traduction [DAN, 1999], qui sont:

[4]C-STAR II a pris le relais, de 1993 à 1999, en s'étendant à 3 autres langues (coréen, italien, français).

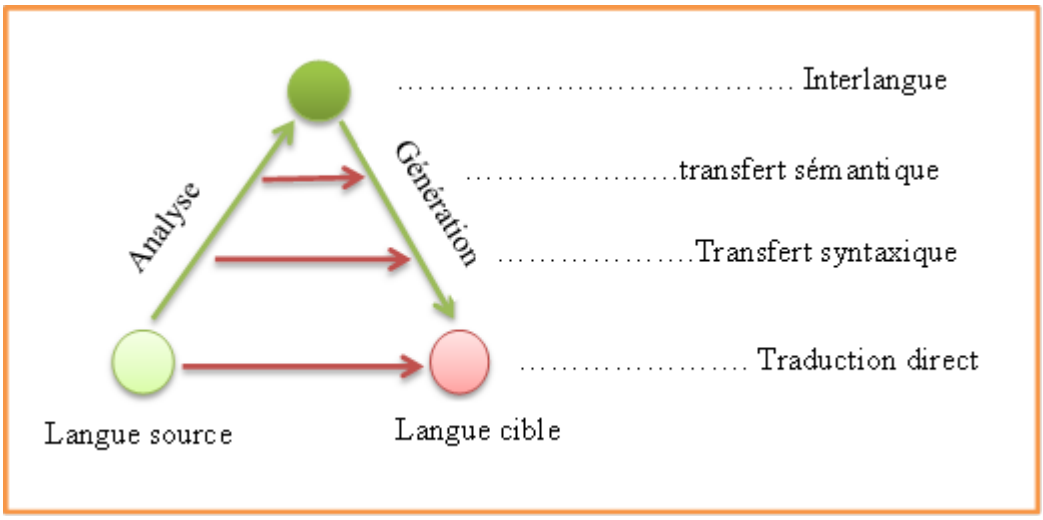


Figure 1.1:Triangle de Vauquois

3.1. Approche Directe

La stratégie directe permet de traduire un texte saisie en langue source «directement » à un texte en langue cible en sortie, sans aucune analyse intermédiaire. Cette stratégie a été adoptée par presque tous les systèmes de traduction mis au point dans les années 50 et 60, Un système de traduction directe est conçu, dès le départ, pour une paire de langues (source et cible) spécifique. Ainsi aucune théorie linguistique générale ou les principes d'analyse ne sont pas nécessairement présents. [HUT, 1992]

Adoptée une stratégie directe cela consiste à suivre trois(03) étapes principales, à savoir:

a. Une analyse morphologique du texte en langue source. A ce stade, le système identifie les mots en réduisant les formes fléchies à leur forme de base sans inflexion.

b. Dictionnaire bilingue de recherche. En fonction d'un dictionnaire bilingue normale système décide le remplacement correct pour des mots de la langue source avec des mots équivalents dans la langue cible.

c. Réorganisation locale du texte en langue cible. Après que le remplacement des mots soit fait, le système de traduction va effectuer l'ajustement du texte en langue cible en appliquant des règles pour mettre les mots dans le bon ordre.

3.2. Approche Pivot (Interlingua / interlangue)

L'approche l'interlangue se base sur l'utilisation d'une langue intermédiaire pour traduire un texte en langue source vers un texte en langue cible. Cette langue intermédiaire est neutre dans le sens ou elle ne possède pas des caractéristiques de la langue source ou cible. "Dans le passé, l'intention ou l'espoir était de développer une langue intermédiaire, qui était vraiment

«universelle »et pouvait donc être intermédiaire entre les langues naturelles. À l'heure actuelle, les systèmes adoptant l'approche interlinguasont moins ambitieux."

L'idée de base derrière l'approche interlangue, est que la traduction est effectuée en deux étapes. Dans la première étape le texte en langue source est analysée afin de produire une représentation en langue intermédiaire et la deuxième étape, le texte en langue cible est générée à partir de cette langue intermédiaire. Cette dernière devrait être entièrement indépendant de la langue" dans le sens où elle ne porte pas de caractéristiques de la langue source et n'est pas conçu avec n'importe quelle langue cible spécifique à l'esprit. [HUT, 1992]

3.3. Approche par Transfert

La méthode de transfert se positionne au milieu entre la stratégie de traduction directe et interlangue. La stratégie de transfert peut être considérée comme un compromis pratique entre l'utilisation efficace des ressources du système adoptant l'approche interlangue, et la facilité de mise en œuvre d'un système adoptant l'approche directe. Le texte en langue source est analysé afin de générer une représentation qui porte les caractéristiques de la langue source. Ensuite, un ensemble de règles de transfert sont appliqués pour transformer cette représentation en une autre qui porte les caractéristiques de la langue cible. À la fin,un module de génération est utilisépour produirele texte en langue cible. [HUT, 1992]

4. La traduction automatique arabe

Comparée au développement qu'a connu le domaine de traduction automatique aux Etats Unis, Europe et même en Asie en terme de performance, la traduction automatique concernant la langue arabe reste en retard. Néanmoins beaucoup de travaux aboutissant à des produits commerciaux ont pu voir le jour pendant ces deux (02) dernières décennies, citant à titre d'exemple :

4.1. Traducteur « Transphere »

En 1990 la firme Apptek[5] commence le développement d'un traducteur automatique arabe, ce qui a donné naissance à un système de traducteur « Anglais- Arabe » baptisé « Transphere », ce système peut être utilisé pour traduire à la fois les textes d'ordre général comme des textes spécialisés par domaine. Le système permet l'utilisation d'un lexique de plus de 100.000 entiers. Il peut être utilisé pour la traduction des documents et, en conjonction avec le traitement de

[5]Applications Technology, Inc(AppTek), basée à McLean, en Virginie, est une société américaine spécialisée dans le développement de logiciels pour les technologies du langage humain

texte, comme un outil pour l'apprentissage assisté par ordinateur de l'arabe. Il aurait été utilisé comme un moteur de traduction dans de grands produits complexes liés à la circulation aérienne, de transport, de communication et de contrôle. Le logiciel fonctionne actuellement sous UNIX et Windows.

4.2. Traducteur « Al- Nakeel »

CIMOS produit le logiciel « Al- Nakel » destinés à la traduction entre différenteslangues. Actuellement il traduit de l'anglais vers l'arabe et le français et du français vers l'arabe et l'anglais. Il est destiné à aider, (non à remplacer) les traducteurs humains dans un large éventail de domaines (science, technologie, commerce, banque, informatique et pétrole). Al Nakels'exécute sous MS Windows.

4.3.Traducteur « Al ArabyMutarjim »

ATA[6] produit Al Araby Mutarjim, qui est annoncé comme le premier système de traduction pour PC et Macintosh assurant traduction anglais - arabe. Ce système a été introduit en 1995 avec une vitesse de traduction minimale de 1000 mots par minute. Il dispose d'un dictionnaire, qui cible des domaines spécifiques tels que : Science, Technologie, Commerce, Finance, Droit, Industrie pétrolière, Agriculture, Médecine, Militaire, etc... Ce système affiche également des significations alternatives, quand elles existent. En Mars 1999 Al - Mutarjim on- line a été présenté (sous le nom ArabeyNet) comme le premier logiciel de traduction l'arabe sur internet.

5. Pour quoi la traduction automatique l'Arabe vers l'Anglais est plus difficile ?

- La morphologie Arabe : La morphologie Arabe est plus complexe et riche si on la compare à la morphologie anglaise.
- Morphologie flexionnelle : L'arabe est une langue flexionnelle. Elle emploie, pour la conjugaison du verbe et la déclinaison du nom, des indices d'aspect, de mode, de temps, de personne, de genre, de nombre et de cas, qui sont en général des suffixes et des préfixes. [SLI, 2008]
- Problème de l'ordre des mots dans la phrase : En arabe on a toujours le libre choix de terme qu'on veut mettre en valeur, en tête de la phrase. Cet ordre relativement libre des mots, provoque des ambiguïtés syntaxiques artificielles dans la mesure où il faut prévoir

[6]Editeur de logiciels basé à Londres, spécialisée dans les logiciels d'affaires arabes.

- dans la grammaire toutes les combinaisons possibles d'inversion de l'ordre des mots dans la phrase.
- L'ambigüité syntaxique de l'arabe : Au niveau syntaxique, nous trouvons dans la langue arabe des phrases simples et des phrases complexes, les phrases complexes sont formées de deux ou plusieurs propositions et la proposition sont souvent des phrases simples.
- L'absence ou l'existence d'un verbe permet de diviser les phrases simples en deux catégories : les phrases nominales et les phrase verbale. Certain de ces caractéristique de la langue peuvent être source d'ambigüité pour l'analyse automatique.
- Problème d'absence des voyelles : Les signes de « voyellation », qui sont positionnés, lorsqu'ils sont notés, sous la forme de signes diacritique placés au-dessus ou au-dessus des lettres, apparaissent dans certain texte (coran hadith) ou littéraires (poésie classique, notamment). Le non vocalisation génère plusieurs cas d'ambiguïtés lexicales et morphologique. [HHA, 04]

- La difficulté d'alignement entre l'anglais et l'arabe: il faut dire que dans la langue arabe un peut signifier toute une phrase en anglais ainsi l'alignement devient une tâche ardue dû au fait de la différence en nombre de mots entre l'anglais et l'arabe [SOH,2011].

Chapitre 2

Evaluation d'un Traducteur Automatique

1. Introduction

Depuis les années 2000, l'évaluation des systèmes de traduction automatique (TA) a pris une importance considérable au sein de la communauté des chercheurs en TA. Cela concerne surtout la TA statistique de l'écrit et de l'oral, parce que les méthodes d'évaluation utilisées sont dérivées de celles qui ont fait leurs preuves en reconnaissance de la parole.

De nombreuses campagnes d'évaluation compétitives ont été organisées, soit directement par les bailleurs de fonds (DARPA, UE, Académie des Sciences Chinoise), soit à leur incitation et avec leur support (NIST, ELDA/ELRA pour la campagne CESTA), soit par des consortiums dans le cadre de projets coopératifs

(CSTAR en 1999, projet NESPOLE! en 2002 et 2003, CSTAR de nouveau avec IWSLT depuis 2004, TC-STAR en 2006). [HER, 2007].

2. Évaluation manuelle des traductions

Initialement, l'évaluation de la qualité des traductions faisait exclusivement appel aux compétences d'experts humains. En effet, chaque évaluateur se voit confier un ensemble de paires de traduction. Chaque paire de traduction fait l'objet de différents types d'évaluation à l'issue desquelles des scores lui sont attribués. Le schéma ci-dessous représente les différents scores proposés pour une évaluation d'une TA:

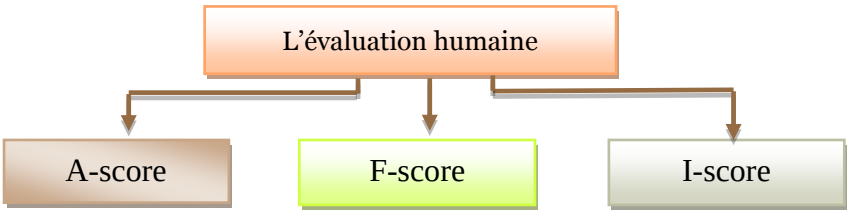


Figure2.1 : Les scores d'évaluation humain

2.1. Le A-score (Adequacy score)

Il indique si le sens de la phrase source est correctement retranscrit dans sa traduction. Chaque juge donne son avis sur la question et attribue à chaque traduction un score allant de 1 (intégralité du sens conservé) à 5 (rien à voir entre la phrase source et sa traduction). L'attribution du A-score ne doit pas tenir compte de la bonne construction de la traduction dans la langue cible.Toutefois, une mauvaise construction peut influencer sa valeur. Il est plus aisé d'attribuer un bon A-score si la phrase est bien construite plutôt que s'il lui manque un verbe ou si les mots ne sont pas correctement ordonnés. [CAR,2010]

2.2. F-score (Fluency score)

Le F-score est une mesure del'intelligibilité qui rend compte du degré de bonne construction de la phrase dans la langue cible. Chaque juge doit évaluer si la traduction est correctement écrite et si elle a un sens sur une échelle de 1 (la traduction est comparable à une phrase écrite par un natif de la langue cible) à 5 (la phrase est incompréhensible). [CAR,2010]

2.3. I-score (Informativeness score)

Tout comme l'A-score, le I-score est une mesure de fidélité. Seul le protocole d'évaluation change. Les juges répondent cette fois à une série de questions à choix multiples sur le contenu des traductions afin de juger si les informations apportées sont équivalentes à celles de la phrase source. Ces trois scores sont assez subjectifs. Une même phrase peut être jugée de façon totalement différente par deux juges. De plus, il semble inapproprié de dissocier le critère d'intelligibilité du critère d'adéquation pour évaluer la qualité d'une traduction, l'un des critères pouvant largement influencer l'autre.[CAR,2010]

3. Problème de l'évaluation humaine

Ces critères de qualité constituent la vraie mesure de l'adéquation du système de traduction à la tâche visée, mais requièrent une coûteuse intervention humaine. Par ailleurs, toute évaluation subjective souffre des problèmes de non-reproductibilité et de variabilité inter-annotateur. C'est en particulier le cas des critères *fluency* et *adequacy* cités plus haut, dont l'évaluation sur une échelle absolue de 1 à 5 est longue et difficile (prend de temps). C'est pourquoi plusieurs mesures automatiques ont été développées au fil des années. Leur objectif est d'être corrélé avec les scores que produirait une évaluation manuelle. Ceci est un problème difficile, car une même phrase peut être traduite de nombreuses façons possibles et également acceptables. Les mesures automatiques doivent autoriser les variations légitimes et pénaliser les erreurs. [DEC,2007]

4. Évaluation automatique

Il s'agit essentiellement de méthodes fondées sur des « références », c'est-à-dire sur des traductions produites par des experts humains : elles consistent à faire des calculs de distance ou de similarité entre les traductions de référence et les traductions « candidates » produites par TA. Ce pour quoi plusieurs métriques d'évaluation sont développées, la figure suivant représente les déferrent mesures les plus utilisée dans l'évaluation automatique.

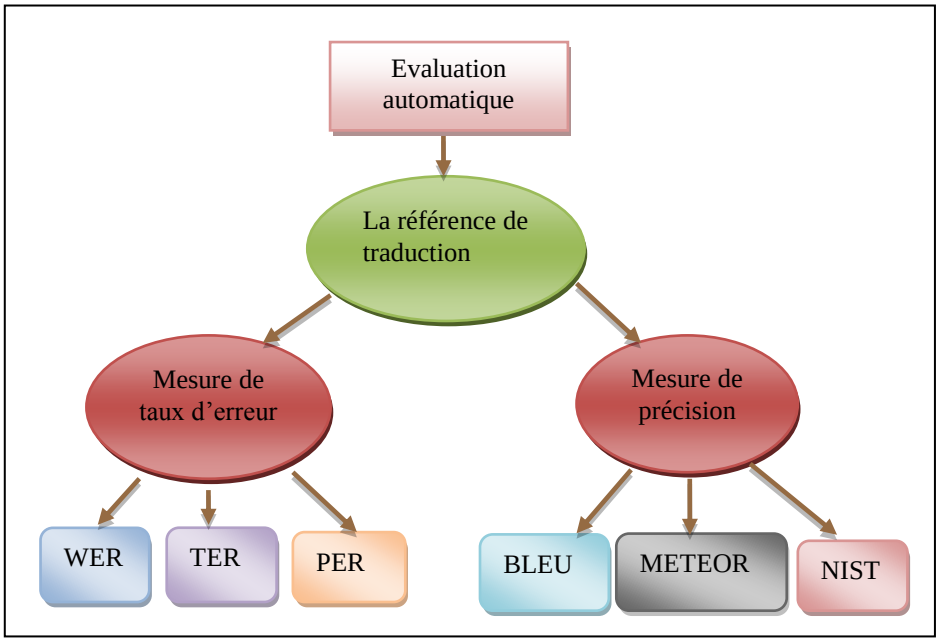


Figure 2.2 : Les mesures automatiques d'évaluation

3.1. La métrique d'évaluation BLEU

Depuis ces dernières années, la métrique la plus souvent utilisée est le score BLEU (en anglais : *BiLingual Evaluation Understudy*). Le score BLEU est proposé par KishorePapineni en 2001. Il ne considère pas seulement la ressemblance au niveau des mots mais aussi la ressemblance au niveau des *n-grammes* entre l'hypothèse et les références. La tâche principale est de comparer les n-grammes de l'hypothèse avec les n-grammes de la référence et de compter le nombre d'équivalences. Les correspondances sont indépendantes de la position. Plus il y a de correspondances, meilleure est l'hypothèse. Tout d'abord, les précisions modifiées de n-gramme (*pn*) avec l'ordre de *1* à *N* (*n=1..N*) sont calculées pour chaque paire d'hypothèses et sa référence (ou ses références lorsque plusieurs références sont utilisées).

$$P_{n\ chaque\ paire} = \frac{\sum_{n-gram \in e_h} Compte_{clip}(n-gram)}{\sum_{n-gram \in e_h} Compte_{e_h}(n-gram)} \qquad (1,1)$$

Pour un*n-gram* donné, soient *Compte*$_{eh}$*(n-gram)* le nombre de fois que ce *n-gram* apparaît dans e_h. Si nous notons *c* le nombre de mots de l'hypothèse e_h, $e_h$contient *c-n+1* n-grammes.Le dénominateur devient *c-n+1*.

Compte$_{clip}$*(n-gram)* est le nombre d'appariements de ce *n-gram* entre $e_h$et e_r, donc il est calculé par : *min(Compte*$_{eh}$*(n-gram), max{*e_r*}(Compte*$_{er}$*(n-gram)))* où

max{e_r}(Compte$_{er}$(n-gram)) est le nombre maximal de fois que ce *n-gram* apparaît dans une référence, parmi toutes les références disponibles.

Pour calculer la précision n-gramme modifiée sur le corpus de test entier, nous accumulons simplement les comptes pour chaque paire d'hypothèses et sa référence.

$$pn = \frac{\sum_{c\in\{\text{Candidates}\}} \sum_{n-gram\in C} Count_{\text{clip}}(n-gram)}{\sum_{C\in\{candidat\}} \sum_{n-gram\ \in C} count(n-gram)} \qquad (1,2)$$

Pour combiner les *N* précisions n-grammes modifiées, le score BLEU utilise le logarithme moyen pondéré, ce qui est équivalent à une moyenne géométrique, et pour pénaliser les hypothèses plus courtes que leurs références, une pénalité de brièveté *BP* est introduite. Le score BLEU est finalement calculé comme suit

$$BLEU = BP.exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \qquad (1,3)$$

*Wn*sont les poids positifs tels que $\sum_{n=1}^{n} w_n = 1$ souvent nous utilisons des poids uniformes $w_n=1/N$ et N=4.

$$BP = \begin{cases} 1 & c > r_p \\ e^{(1-r_p/c)} & c < r_p \end{cases} \qquad (1,4)$$

Pour une paire de phrases, *c* est la longueur de l'hypothèse *eh*, et *rp*est la longueur de la référence la plus proche de *eh*parmi les références. Pour le corpus entier, la somme totale de *c* et la somme totale de *rp*de toutes les hypothèses du corpus sont calculées.

Dans le domaine des logarithmes,

$$\log BLEU = min\left(1-\frac{r_p}{c},0\right) + \sum_{n=1}^{N} w_n \log p_n \qquad (1,5)$$

BLEU est un score de précision, sa valeur varie de 0 à 1. Plus le score est élevé, meilleure est la traduction. Une hypothèse se voit attribuer un score BLEU de « *1* » lorsqu'elle est identique à une des références ; au contraire, elle aura un score BLEU de « *0* » si aucun de ses n-grammes n'est présent dans une référence [THI,2011].

La métrique BLEU est représenté par le shéma suivant :

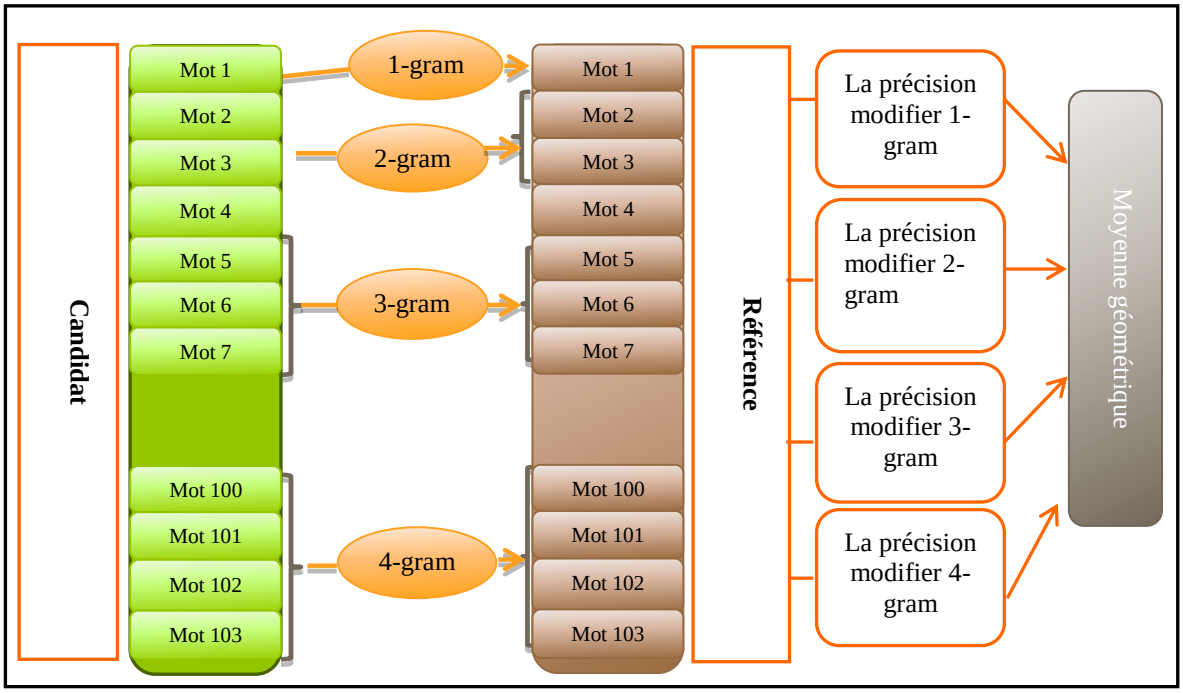


Figure 2.3 : Représentation de la métrique BLEU

3.2. La métrique d'évaluation METEOR

La métrique METEOR proposée par Lavie et Agarwal en 2007 (Metric for Evaluation of Translation with Explicit Ordering) évalué la traduction en calculant un score basé sur l'alignement direct et explicite mot-par-mot entre la traduction du système et la traduction de référence donnée. Lorsqu'on a plusieurs traductions des références sont disponibles, on calcule le score de cette traduction pour chaque traduction des références indépendamment, et on choisit le pair qui donne le meilleur score.

Avec une paire du Chaine de caractèredonnée à comparer, METEOR génère un alignement des mots entre les deux chaine de caractère : l'alignement est élaboré entre des mots,de façon que chaque mot dans chaque chaine de caractère est relie à un mot au plus dans l'autre chaine de caractère.

Cet alignement est progressivement produit par une séquence des modules de mot-alignement

- Le module « exact » relier deux mot s'ils sont identiques.
- Le module « porter stem » relier deux mots s'ils ont le même lemme après la lemmatisation on utilise le Porter Stemmer.

- ❖ Le module « WN synonymy » relier deux mots s'ils sont considéré comme synonymes, en se basant sur le fait queles deux mots appartiennent à un même « synset » dans le Word Net.[ABA,2008]

Une fois l'alignement final a été produit entre la traduction du système et la traduction de référence, le score du METEOR est calculé par rapport à cette paire comme suite :

On se base sur le nombre d'uni-grammes attachés qu'on obtient entre les deux chaine de caractères (m), le nombre d'uni-gramme total dans la traduction (t), et le nombre d'uni-gramme dans la référence (r). On calcule la précision de l'uni-gramme P= m/t ; et le rappel d'uni-gramme R= m/r. alors on calcule le paramètreharmonicmean de P et R.

$$Fmean = \frac{P.R}{\alpha.P + (1-\alpha).R} \qquad (2,1)$$

La précision, Rappel et Fmean sont basées sur l'alignement d'un seul mot. Pour tenir compte àquel point l'alignement d'uni-gramme dans les chaine de caractères sont dans le même ordre des mots, METEOR calcule une pénalité pour un alignement donnée comme suite :

Primairement, laséquence d'alignement uni-gramme entre les deux chaine de caractère est divisé au nombre minimum de « chunks » de façon que l'alignement uni-gramme dans chaque chunk sont adjacent (dans le deux chaine de caractère) et le même ordre de mots. Le nombre du Chunks et le nombre d'alignement (m) sont utilisé pour calculer la fraction de fragmentation frag = Ch/m. la pénalité est donc calculé comme suite :

$$Pen = \gamma.frag^{\beta} \qquad (2,2)$$

La valeur du y détermine le maximum de pénalité (0<y<1), la valeur de b détermine la relation fonctionnelle entre la fragmentation et la pénalité.

Le score du METEOR d'alignement entre les deux chaine de caractère est calculé comme suite :

$$score = (1 - Pen).\text{Fmean} \qquad (2,3)$$

Le mauvais score de METEOR est 0, est attribuéorsqu'on ne trouve aucun alignement. Après on applique les trois module d'alignement par exemple le module lemmatisation, et METEOR donne le score 1 lorsque on trouve l'alignement parfait entre la référence et l'hypothèse.[BIR, 2011] on donne une résumée de cette métrique avec le schéma suivant :

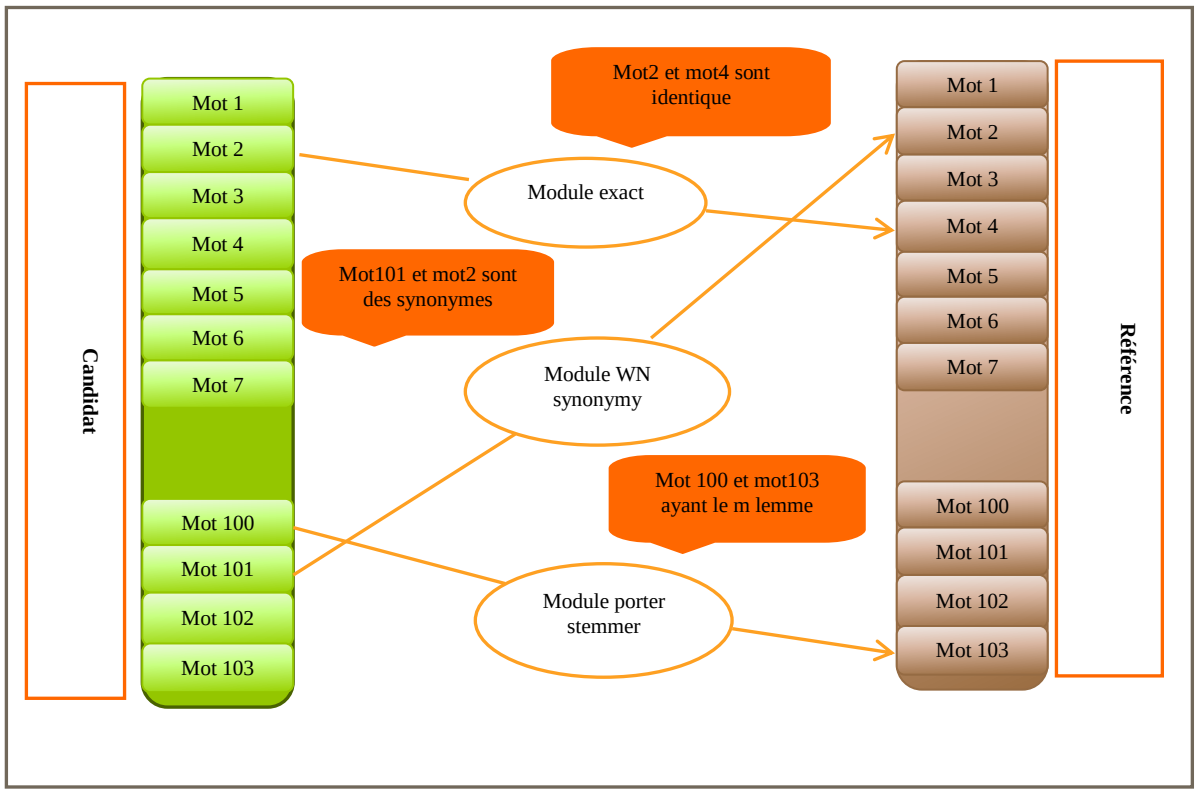


Figure 2.4: Représentation de la métrique METEOR

3.3. La métrique d'évaluation NIST

Le score NIST (National Institute of Standards and Technology)proposé par Doddingtonen 2002, tout comme le score BLEU, repose sur la précision n-gramme.Cependant, il considère, non pas la moyenne géométrique des n-grammes communs à la traductionautomatique et à la référence comme le fait BLEU, mais la moyenne arithmétique. Une deslacunes du score BLEU réside dans le fait que tous les n-grammes dans le calcul de la précisionont la même importance. Selon Doddington il paraît plus approprié d'accorder davantage de poidsaux n-grammes importants, c'est-à-dire ceux qui sont porteurs d'information. Par conséquent, ilpondère le compte des n-grammes dans le calcul de la moyenne par leur importance, comme lemontre la formule suivante :

$$info(n-gram) = info(m1, m2, \dots m_n) = \log_2 \left(\frac{compte(m_1 \dots m_{n-1})}{compte(m_1 \dots m_n)} \right) \qquad (3,1)$$

N est l'ordre des n-grammes considérés. BP est, comme dans BLEU, une pénalité destinée aux traductions automatiques courtes. M est l'ensemble des n-grammes communs à la traduction hypothèse et à la référence. hypest le nombre de n-grammes dans la traduction évaluée. Enfin, Info(ngram) est une fonction dépendante du nombre d'occurrences du n-gramme passé en paramètre dans les traductions références. Le score NIST considère les n-grammes fréquents moins importants que les n-grammes qui apparaissent peu de fois.

$$NIST = \sum_{n=1}^{N} \left\{ \frac{\sum_{n-gram \in e_h \cap n-gram \in e_r} info(n-gram)}{\sum_{n-gram \in e_h} compte(n-gram)} \right\} . exp\left\{ log^2 \left[\min \left(\frac{c}{r}, 1 \right) \right] \right\} \qquad (3,2)$$

Les deux métrique BLEU et NIST sont des mesures de précision plus leur valeurs sont grandes,meilleurs est la traduction.

Le score de BLEU et NIST ne sont pas additifs pour des déférenténonces. C.-à-d. le score du documente ne peut pas obtenu par l'addition des scores des énoncesindépendant. [CAR,2010]

3.4. La métrique d'évaluation TER

TER (Translation Edit Rate) proposé par Matthew Snover et Bonnie Dorr en 2006,elle permet de mesurer le nombre minimum d'opérations qu'une personne doit apporter à une traduction automatique pour que celle-ci soit identique à une des références humaines correspondantes. Les opérations considérées sont : l'insertion, la suppression et le remplacement d'un mot, tout comme le score WER, mais aussi le déplacement d'une suite de mots. Ce score est défini par :

$$TER = \frac{Nb\ (op)}{AvregN_{\ Ref}} \qquad (4,1)$$

- « Nb(op) » est le nombre minimum d'opérations, calculé par programmation dynamique.
- « Nref » est la taille moyenne en mots des références.

Une fois de plus, cette mesure rend compte d'une certaine distance entre une traduction automatique et une référence humaine. Cependant la mesure TER ne permet pas vraiment de juger si une traduction est acceptable ou non du fait

qu'aucune reformulation de la traduction de référence n'est acceptée. D'autres variantes du score TER ont été proposées, il s'agit des mesures TERp et HTER.[SNO,2006]

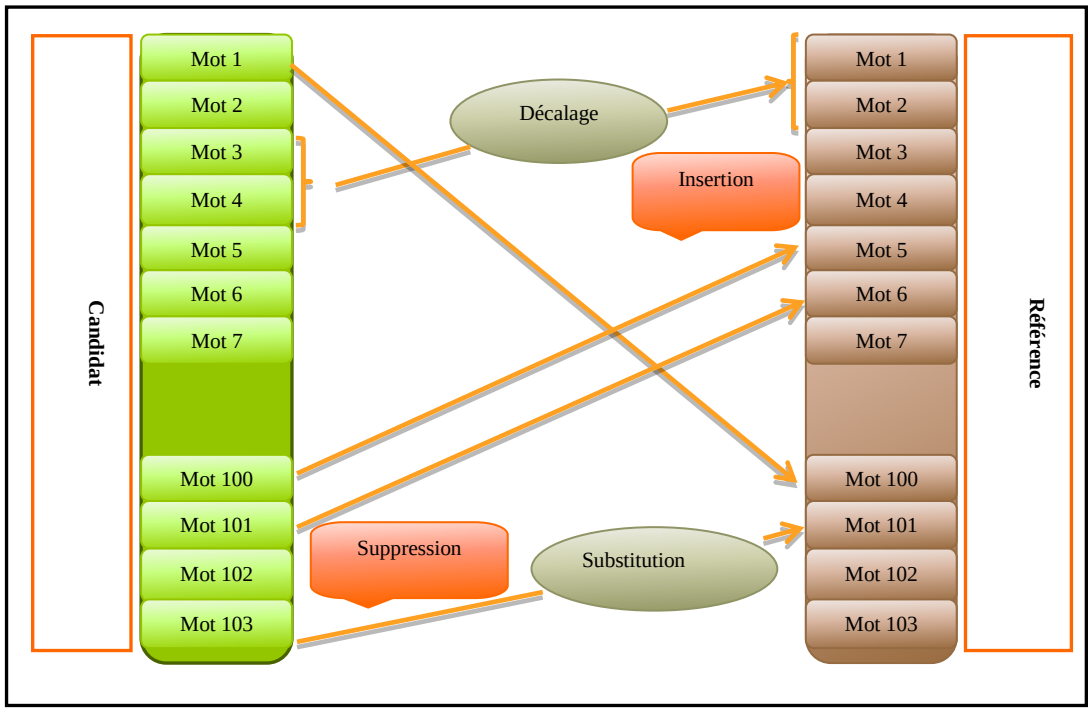


Figure 2.5: représentation de la métrique TER

3.5. La métrique d'évaluation WER

La métrique WER (Word Error Rate) proposé par Tillmann en 1997est calculée à l'aide de la distance d'édition séparant une traduction candidate de sa référence. La distance d'édition est le nombre minimal d'opérations de suppression, d'insertion ou de remplacement à effectuer pour transformer une chaîne de caractères en une autre. Le score WERest obtenu en normalisant la distance d'édition par la taille de la référence, en termes de caractères. Le score WERtraduit le degré de non-conformité ou (non-authenticité) de la traduction et de la référence. Autrement dit lorsque WER augmente la conformité ou la similarité de la traduction à la référence diminue.[CHI,2012]

3.6. La métrique d'évaluation SER

La métrique SER(Phrase Error Rate) mesure le taux de phrases traduites non identiques aux phrases de référence. Comme pour la métrique WER**,** on note que le

degré de similarité ou de conformité de la traduction à la référence varie en sens inverse avec la mesure SER.

Les deux indicateurs de sélection et de performance WERet SERont la même importance que le score BLEU. Néanmoins, WERet SERont l'inconvénient et le risque de pouvoir omettre ou éliminer à tort des traductions potentiellement correctes. Ces dernières sont considérées par WERet SERdes traductions erronées, car ces métriques ne tiennent compte que de la ressemblance entre la référence et la traduction candidate.[DEC,2007]

3.7. La métrique d'évaluationPER

Le score WER est utilisé dans beaucoup de domaines, comme par exemple la reconnaissance de la parole. Il est là particulièrement adapté du fait de la monotonicité entre le signal audio et les mots transcrits. Pour la traduction, en revanche, le scoreWER peut pénaliser injustement des traductions correctes si elles organisent les mots différemment des traductions de référence. Cette inflexibilité du score WER a motivé la création du score PER (*Position-independentwordError Rate*, taux de mots erronés indépendamment des positions). Appelons « sc » le « sac de mots » contenant les mots de la traduction candidate ; les mots apparaissant plusieurs fois dans « ec » apparaissent exactement autant de fois dans « sc », et en particulier |sc| = |ec|. De même, soit sr le sac de mots de la traduction de référence er. Le score PER de ecest calculé comme suit :

$$per_{(es)} = \frac{\max\left(|e_s \backslash s_r, |s_r \backslash s_r|\right)}{|e_r|} \quad (7,1)$$

Où « sc\sr » est le sac de mots scprivé des mots également présents dans sr. Par exemple, si « sc » contient quatre fois le mot « w » et que « sr » ne le contient que deux fois, « sc\ sr » le contient encore deux fois et sr \ scne le contient pas du tout.

Comme précédemment pour le score WER, lorsque l'on dispose de plusieurs références, on utilise le plus petit numérateur constaté sur l'ensemble des références, et la moyenne des longueurs des références au dénominateur. [DEC,2007]

3.8. La métrique d'évaluationHTER

Est une version modifiée du score TER avec l'intervention des traducteurs humains. Après avoir lu les références, les traducteurs humain éditent l'hypothèse du système pour générer une nouvelle phrase qui a la même signification que les références

originales. Cette phrase est considérée comme une nouvelle référence humaine de l'hypothèse. Puis, le score HTER(*Human-targeted Translation Edit Rate*)est le score TER minimum calculé entre l'hypothèse et les références originales plus la nouvelle référence humaine.

Le score HTER est moins subjectif que les jugements humains, mais il est encore coûteux, en ce que le traducteur perd environ de 3 à 7 minutes pour éditer chaque phrase. Et le score HTER n'est pas adapté pour être utilisé dans le cycle de développement d'un système de TA.

3.9. La métrique d'évaluation TER-Plus (TERp)

Est une autre extension du score TER avec des paramètres ajustables et l'incorporation avec la morphologie, la synonymie et des paraphrases. TERp aligne un mot de l'hypothèse avec un mot de la référence non seulement quand deux mots sont les correspondances exactes, mais aussi quand ils possèdent la même racine ou ils sont les synonymes. Plus, TERp utilise aussi la substitution de groupe de mots (des paraphrases) pour aligner deux phrases. Il utilise donc toutes les opérations de TER : l'insertion, la suppression, la substitution de mots, le déplacement d'une suite de mots ; et trois nouvelles opérations : la correspondance en racine, la correspondance en synonyme et la substitution de paraphrases. Le coût de toutes les opérations est optimisé afin de maximiser la corrélation avec les jugements humains.

Le score TERp permet de mieux aligner l'hypothèse et les références, mais le calcul dépend du dictionnaire de synonymes, de la liste de mots possédant la même racine, de la liste des paraphrases, qui ne sont pas toujours disponible pour toutes les langues. [THI,2011]

3.10. La métrique d'évaluation WNM

Babych et Hartley ont également proposé une extension de la méthode BLEU appelée WNM pour Weighted N-gram en partant du même constat qu'une équivalence entre mots clés qui apportent beaucoup d'information pour le sens de la phrase est beaucoup plus significative qu'une équivalence entre mots outils de la langue qui ne sont nécessaires que pour la bonne structure de la phrase. Ils proposent donc de pondérer les n-grammes à l'aide de deux scores : la mesure standard tf.idf proposé par Salton (1968) et la mesure S-Score proposé par Babych (2003). Ces deux mesures permettent d'accorder plus de poids aux mots ou suite de mots porteurs de sens qu'aux mots outils de la langue. Elles rendent compte de la prépondérance d'un mot dans un corpus. Ces deux scores sont calculés pour chaque n-gramme du corpus de

référence. Ils sont ensuite intégrés dans le calcul de la précision, du rappel et de la F-mesure en pondérant chaque équivalence de n-grammes par les scores qui lui sont associés. Rappelons que la précision est le nombre d'équivalences den-grammes divisé par le nombre de n-grammes de la traduction candidate et que le rappel est le nombre d'équivalences de n-grammes divisé par le nombre de n-grammes de la référence. La F-mesure est une moyenne harmonique de la précision et du rappel. Le score WNM prend donc en compte la précision et le rappel des n-grammes, contrairement au score BLEU qui ne considère pas directement le rappel. En effet, pour le score BLEU, c'est seulement par l'intermédiaire de la pénalité BP que le score tient compte d'un problème éventuel de recouvrement en défavorisant les traductions automatiques trop courtes par rapport aux références.[CAR,2010]

4. Conclusion

On peut constater qu'il existe dans d'évaluation automatique plusieurs métriques ayant pour rôle d'évaluer automatiquement la qualité d'un traducteur, ces métriques sont variés dans le principe et la manière d'évaluation, par exemple BLEU utilise la précision modifié n-gram, par contre TER calcule le nombre d'opération indispensable. La question qui dispose qu'elle est la métrique d'évaluation qui donne le bon résultat c'est-à-dire. Quelle est la métrique qui permit de conserve les deux points suivants : l'adéquation et la fluidité ?

Chapitre 4

Tests, Résultats et Analyse

1. Introduction

Une fois l'architecture générale de notre boite à outils implémentée et arrivée au bon fonctionnement de chaque métrique, on a procédé à la phase de test, pour définir d'un coté l'efficacité de notre boite à outils, mais aussi la possibilité d'avoir une indication sur le meilleur traducteur dans le cas d'une traduction automatique Arabe / Anglais.

Le but de ce chapitre, est de présenter les résultats obtenus, de l'évaluation de la qualité des traducteurs : Microsoft, Google, Reverso, Systan et Babylon avec différentes métriques (BLEU, NIST, METEOR, TER, WER) en utilisant le corpus d'évaluation des Nations unis.

2. Outils et environnement de développement

Avant de commencer l'implémentation de l'application, il y a lieu d'abord de spécifier les outils utilisés, qu'on a suggéré être le bon choix vu les avantages qu'ils offrent.

2.1. Environnements de développement et matériels

Nous avons développé notre application sur une machine Intel ® Core™ i3-370, avec une vitesse de 2,4 GHz, doté d'une capacité mémoire de 3,00 GB de RAM.

Concernant les ressources logicielles un Microsoft Windows 7 Edition Intégrale est installée sur cet ordinateur, avec un l'éditeur java NetBeans qui est utilisée pour l'implémentation du travail.

2.1.1. Pour quoi on choisir le langage java ?

Java est un langage de programmation à usage général, évolué et orienté objet dont la syntaxe est proche du C. Ses caractéristiques ainsi que la richesse de son écosystème et de sa communauté lui ont permis d'être très largement utilisé pour le développement d'applications de types très disparates. Java est notamment largement utilisé pour le développement d'applications d'entreprise et mobiles.

2.1.2. Les caractéristiques du langage java

Java possède un certain nombre de caractéristiques qui ont largement contribué à son énorme succès :

- **Java est interprétée** : le source est compilé en pseudo code ou byte code puis exécuté par un interpréteur Java: la Java Virtual Machine (JVM). Ce concept est à la base du slogan de Sun pour Java : WORA.

- **Java est portable** : il est indépendant de toute plate-forme : il n'y a pas de compilation spécifique pour chaque plate forme. Le code reste indépendant de la machine sur laquelle il s'exécute. Il est possible d'exécuter des programmes Java sur tous les environnements qui possèdent une Java Virtual Machine.

- **Java est orienté objet :** comme la plupart des langages récents, Java est orienté objet. Chaque fichier source contient la définition d'une ou plusieurs classes qui sont utilisées les unes avec les autres pour former une application.

- **Java est simple :** le choix de ses auteurs a été d'abandonner des éléments mal compris ou mal exploités des autres langages tels que la notion de pointeurs (pour éviter les incidents en manipulant directement la mémoire), l'héritage multiple et la surcharge des opérateurs, ...

❖ **Java est sûr :** La sécurité fait partie intégrante du système d'exécution et du compilateur. Un programme Java planté ne menace pas le système d'exploitation. Il ne peut pas y avoir d'accès direct à la mémoire. L'accès au disque dur est réglementé dans une applet.

Les applets fonctionnant sur le Web sont soumises aux restrictions suivantes dans la version 1.0 de Java :

Aucun programme ne peut ouvrir, lire, écrire ou effacer un fichier sur le système de l'utilisateur

- ❖ aucun programme ne peut lancer un autre programme sur le système de l'utilisateur
- ❖ toute fenêtre créée par le programme est clairement identifiée comme étant une fenêtre Java, ce qui interdit par exemple la création d'une fausse fenêtre demandant un mot de passe
- ❖ les programmes ne peuvent pas se connecter à d'autres sites Web que celui dont ils proviennent.

❖ Java est multitâche : Il permet l'utilisation de threads qui sont des unités d'exécutions isolées. La JVM, elle-même, utilise plusieurs threads. [JEA, 2008]

2.2. Les Traducteurs automatiques on line

Un système de Traduction Automatique est un système informatique qui a : pour entrée un texte "t1", ou texte source écrit dans une langue "L1" ou langue d'origine, et n'ayant pas subi d'aménagements spéciaux préalables au traitement automatique qu'il va subir, et pour sortie un texte "t2" ou texte traduit écrit dans une langue "L2" ou langue cible, tel qu'il n'ait pas à subir de transformations pour être reconnu par les utilisateurs comme une traduction du texte " t1".

2.2.1. Le traducteur Systran

Fondée par le Dr Peter Toma en 1968, est une des plus anciennes sociétés de traduction automatique. SYSTRAN a beaucoup travaillé pour le ministère américain de la Défense et de la Commission européenne[7].

Depuis sa création, SYSTRAN n'a jamais cessé d'innover pour remplir sa mission : « permettre la communication dans différentes langues ». SYSTRAN est à la pointe de la recherche dans le domaine du traitement automatique des langues tant linguistique que statistique. Des nouvelles voies sont explorées en permanence pour améliorer les logiciels en termes de qualité, de performance et d'intégration.

[7] http://en.wikipedia.org/wiki/Systran. (Dernier Accès : le 25/5/2013 à 20 : 14).

Le moteur de traduction hybride mis au point par SYSTRAN combine les qualités de la technologie à base de règles ("rule-based") et du traitement "statistique". Le moteur hybride permet d'atteindre les objectifs des entreprises en termes de qualité de traduction, d'investissement et de productivité. Par conséquent, la taille des modèles statistiques générés puis mis en œuvre est également réduite, ce qui constitue un avantage en termes de performance et de configuration requise.[8]

2.2.2. Le traducteur Babylon

Est un dictionnaire informatique et logiciel de traduction, développé par Babylon Ltd, une société publique israélienne (TASE: BBYL). basé à Or Yehuda. La société a été créée en 1997 par l'entrepreneur israélien Amnon Ovadia. Babylon comprend des dictionnaires et des glossaires créés par la communauté. Il est un outil utilisé pour la traduction et la conversion des devises, les mesures et le temps, et pour obtenir d'autres informations contextuelles. Le programme permet aux utilisateurs d'entendre la prononciation correcte des mots et du texte[9]. Le moteur de traduction pour ce traducteur est base sur l'approche statistique.

2.2.3. Le traducteur Google traduction (en anglais : Google Translate)

Est un service fourni par Google qui permet de traduire un texte ou une page Web dans une autre langue. Contrairement à d'autres services de traduction comme Babel Fish, AOL et Yahoo qui utilisent SYSTRAN, Google utilise son propre logiciel de traduction[10].

Lorsque Google Traduction génère une traduction, il recherche des modèles dans des centaines de millions de documents afin de déterminer quelle est la meilleure traduction. En recherchant des modèles dans des documents traduits par des traducteurs humains, Google Traduction peut identifier la traduction la plus appropriée. Ce processus de recherche dans d'importants volumes de texte est appelé "traduction automatique statistique". Les traductions étant générées par des machines, elles peuvent présenter des imperfections. Plus Google Traduction peut analyser de documents traduits par l'homme dans une langue donnée, meilleure est la traduction. C'est pour cette raison que la qualité des traductions peut varier d'une langue à l'autre[11].

[8] http://www.systran.fr/systran/entreprise/technologie (Dernier Accès : le 12/09/2013 à 19 : 00).
[9] http://en.wikipedia.org/wiki/Babylon_(software) (Dernier Accès : le 27/10/2013 à 20 : 00).
[10] http://en.wikipedia.org/wiki/Google_Translate (Dernier Accès : le 05/09/2013 à 14 : 00).
[11] http://translate.google.dz/about/intl/fr_ALL/ (Dernier Accès : le 27/09/2013 à 21 : 00).

2.2.4. Le traducteur Reverso

Est le modèle de traducteur de langue produit par Softissimo. La version de traduction du logiciel propose un modèle de traduction assez complexe et détaillé (de maximum 256 caractères pour un texte simple) dans plusieurs choix de langue.

2.2.5. Le traducteur Microsoft

La technologie de traduction automatique derrière Microsoft Translator est construite sur plus de dix années de travail au sein de Microsoft Research et offre un service de traduction automatique souple, instantanée et rentable pour toutes les destinations ; aide les entreprises, les développeurs et les utilisateurs à s'affranchir des barrières linguistiques. L'API de traduction, accessible et enrichie, permet aux développeurs d'applications et aux fournisseurs de solutions de livrer les services de traduction demandées par leurs clients.

Microsoft Translator offre une traduction automatique de texte dans une langue déterminée. C'est un système de traduction automatique statistique et de pointe, permettant de traduire dans toutes les langues prises en charge et d'améliorer des milliards de traductions tous les jours. L'API Microsoft Translator est disponible sur Windows Azure Marketplace[12].

2.3. Corpus d'évaluation (Nations Unis)

2.3.2. Présentation

Afin d'évaluer et valider notre Box (boite) d'évaluation nous avons opté pour un corpus d'évaluation parallèle (bilingue : Arabe / Anglais) de référence à savoir, Arabic- English Parallel Corpus of United Nations Texts qui fait partie intégrante d'un corpus multilingues MultiUN (Multilingual UN Parallel Text)[13].

Ce corpus à été développé dans le laboratoire « Language Technology Lab in DFKI GmbH (LT-DFKI) » en Allemagne par Andreas Eisele et Yu Chen, comportant des documents issus des interventions au sein des Nations Unis entre 2000 et 2009. Les documents sont convertis en format XML.

[12] https://ieonline.microsoft.com/#ieslice (Dernier Accès : le 27/09/2013 à 21 : 00).
[13] http://www.euromatrixplus.net/multi-un/ (Dernier Accès : le 27/7/2013 à 14 : 00).

2.4. Description

Les tableaux ci-dessus nous donnent une indication concernant le volume (la masse : documents[14], phrases et Mots) d'information que contient notre corpus d'évaluation.

Langue	Arabe	Anglais
Documents	65156	96240
Phrases	11050313	17098695
Mots	237412090	385894793

Tableau 4.1 : Description monolingue du Corpus d'évaluation

Langue	Arabe	Anglais
Documents	63257	
Phrases	8206568	
Mots	180759040	214681635

Tableau 4.2 : Description Parallèle du Corpus d'évaluation

3. L'interface Graphique principale

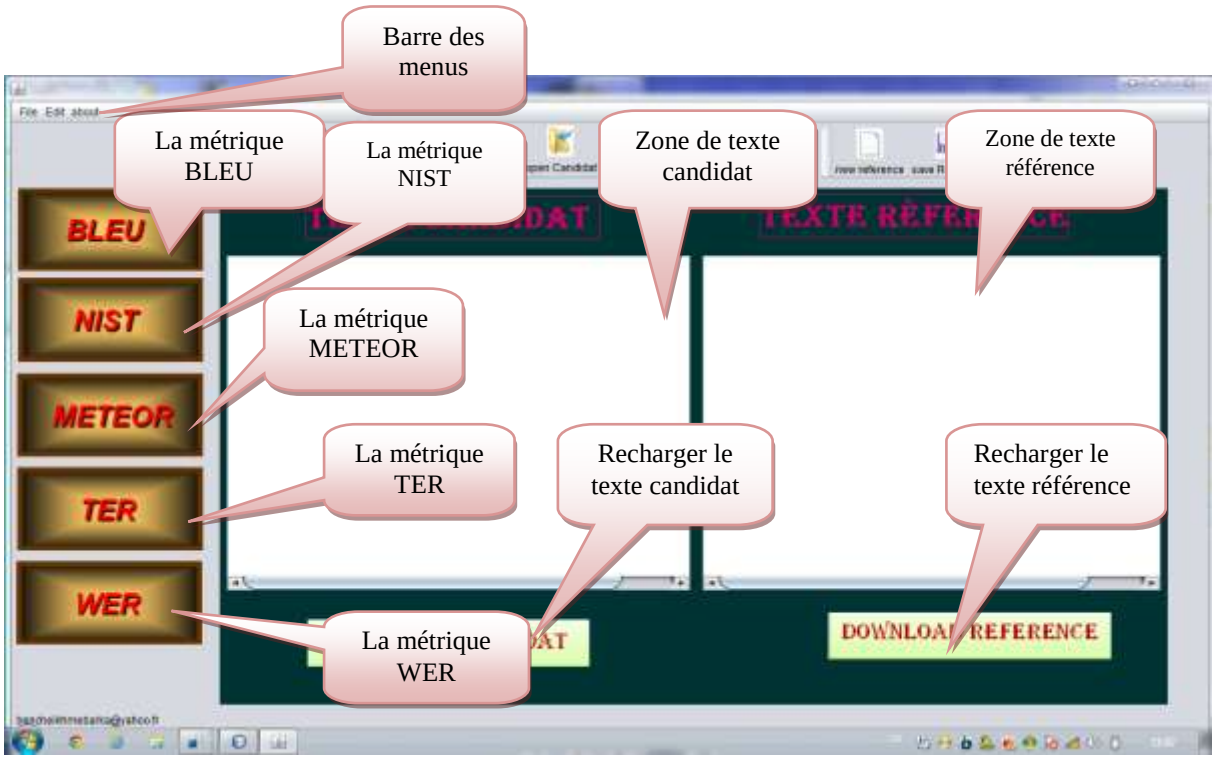
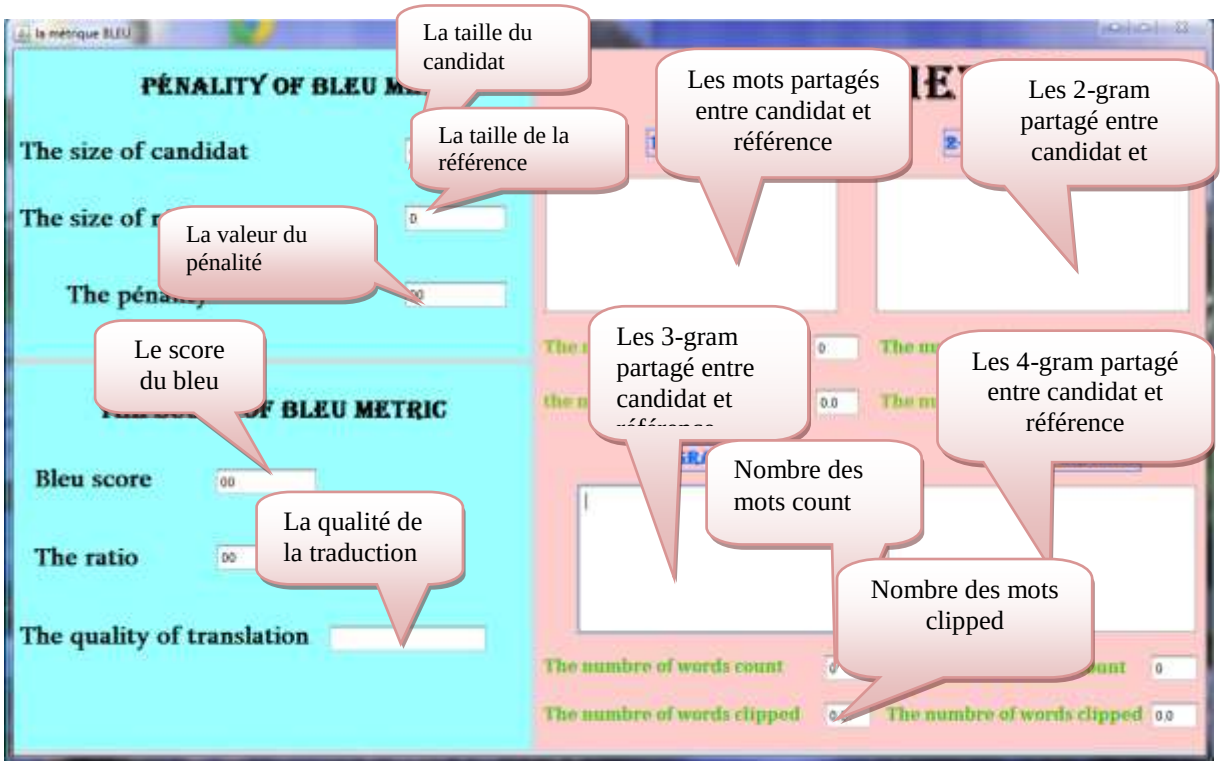


Figure 4.1 : l'interface principale du projet

[14] Les documents sont convertis en format XML

3.1. L'interface de la métrique BLEU



Figue 4.2: interface de la métrique BLEU

3.2. l'interface de la métrique NIST

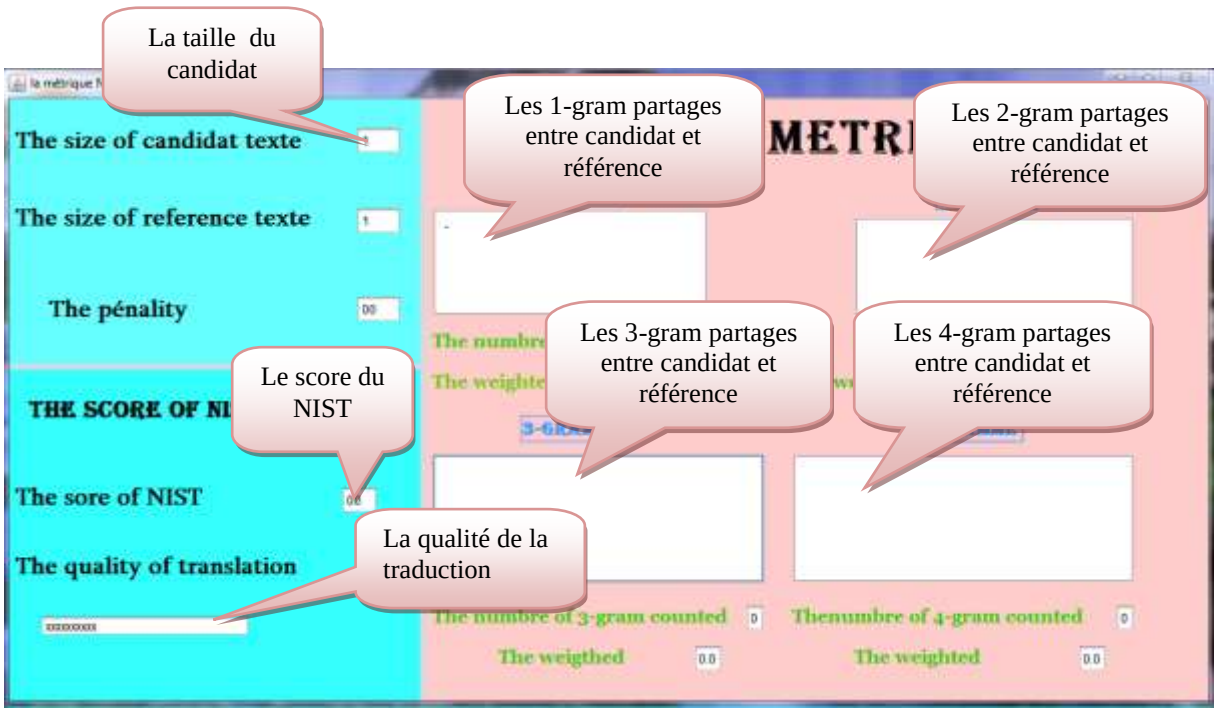


Figure 4.3 : l'interface du métrique NIST.

3.3. l'interface de la métrique METEOR

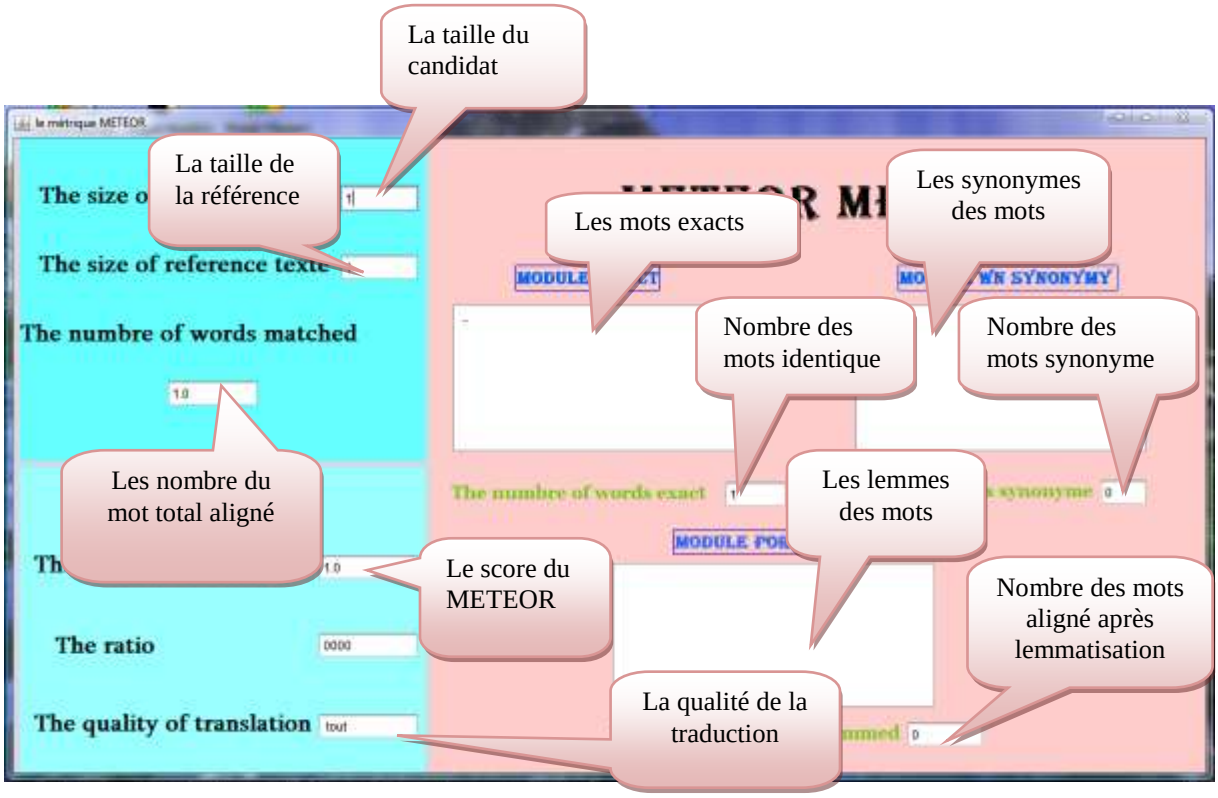


Figure 4.4 : l'interface du métrique METEOR

3.4. l'interface de la métrique TER

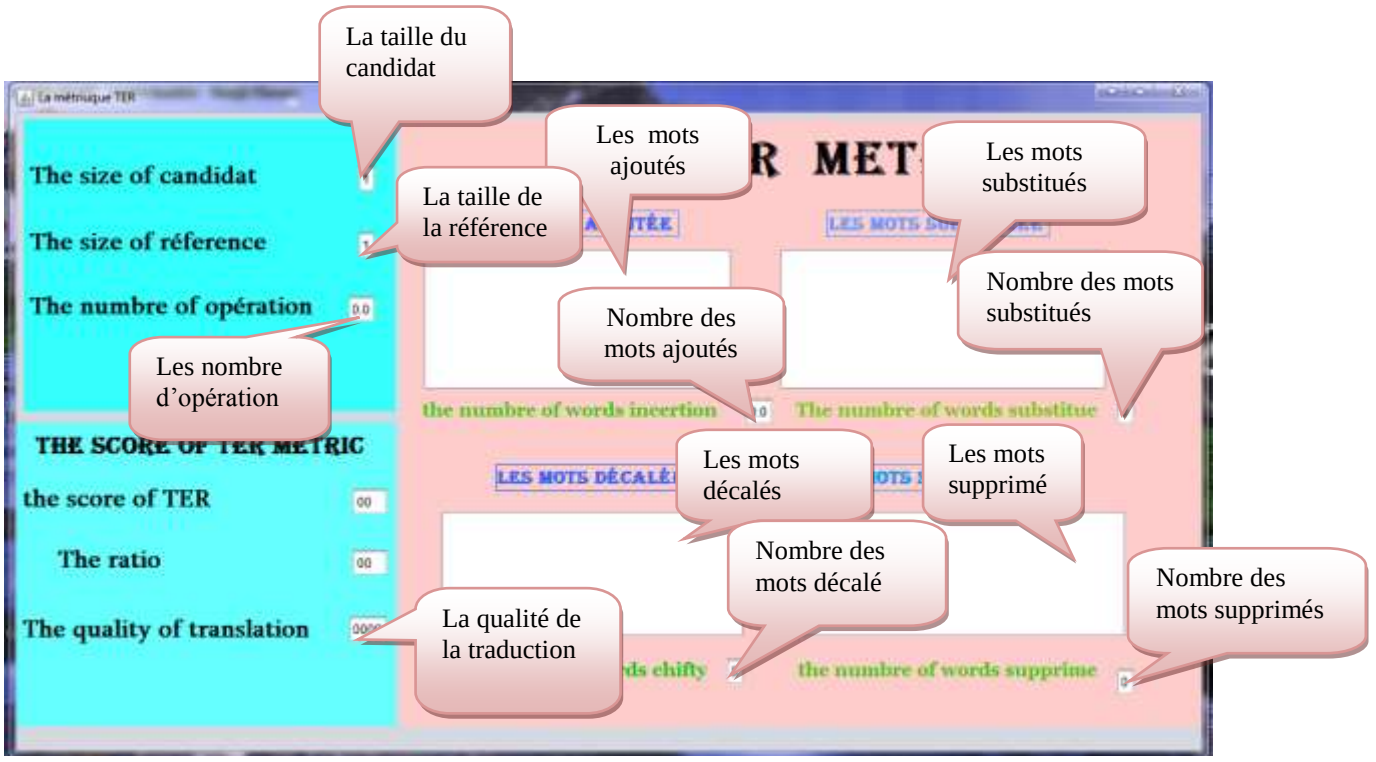


Figure 4.5: l'interface de la métrique TER

3.5. l'interface de la métrique WER

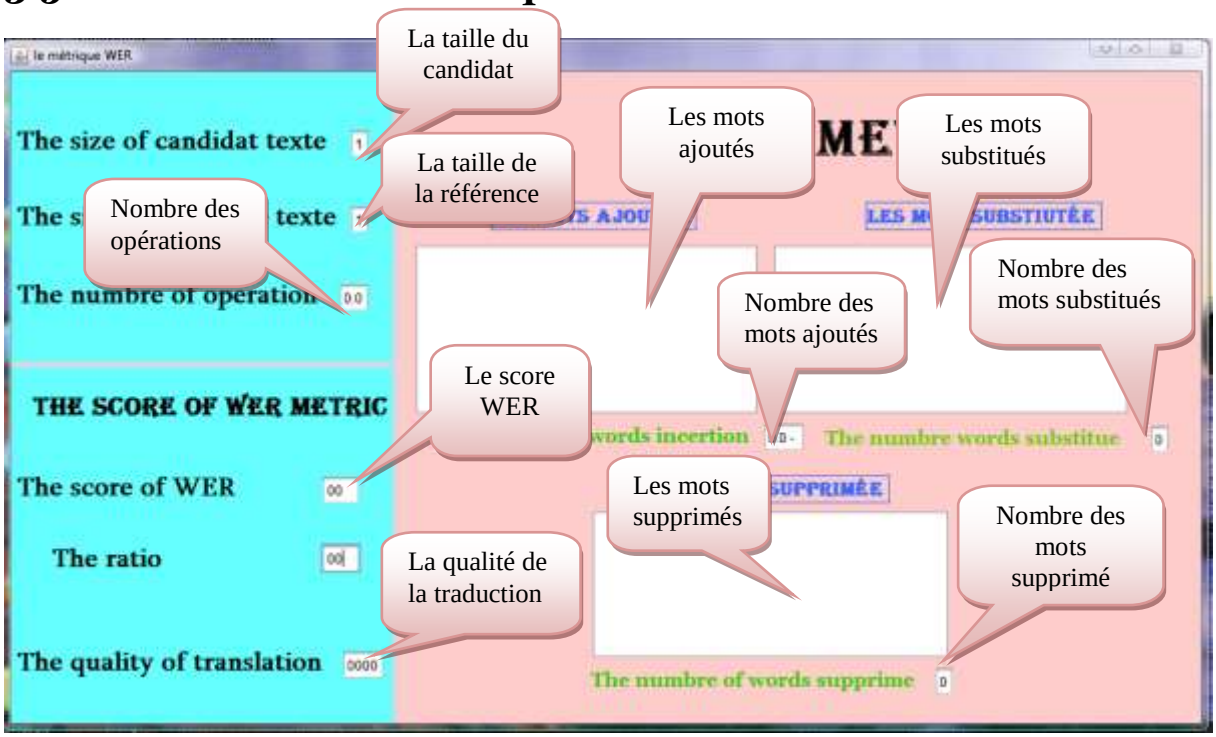


Figure 4.6 : l'interface du métrique WER

4. Résultat des tests

3.1. L'évaluation des traducteurs selon la métrique BLEU

Dans cette expérience nous évaluons chaque traducteur en utilisant la métrique BLEU, nous donnons une partie des résultats dans le tableau suivant :

3.1.1. Les résultats

	Babylon	Systran	Google	Microsoft	Reverso
1	0.8497	0	0.2589	0.9025	0.9025
2	0.6134	0	0.5343	0.6086	0.6463
3	0.8437	0	0.471	0.8713	0.496
4	0.5777	0.12	0.4593	0.5595	0.5421
5	0.6921	0	0.5469	0.7726	1
6	0.8364	0	0	0	0
7	0.797	0.0025	0.5811	0.8453	0.4693
8	0.995	0.1056	0.2082	0.4813	0.2338
9	0	0	0.6285	0.825	0.4662
10	0.7485	0.0469	0.4524	0.8753	0.7489
11	0.6144	0	0.6929	0.6929	0.325
12	1	0	0.8804	0.8447	0.8447

Tableau 4.3 : Le score BLEU pour les traducteurs (Babylon, Systran, Reverso, Microsoft, et Google traduction)

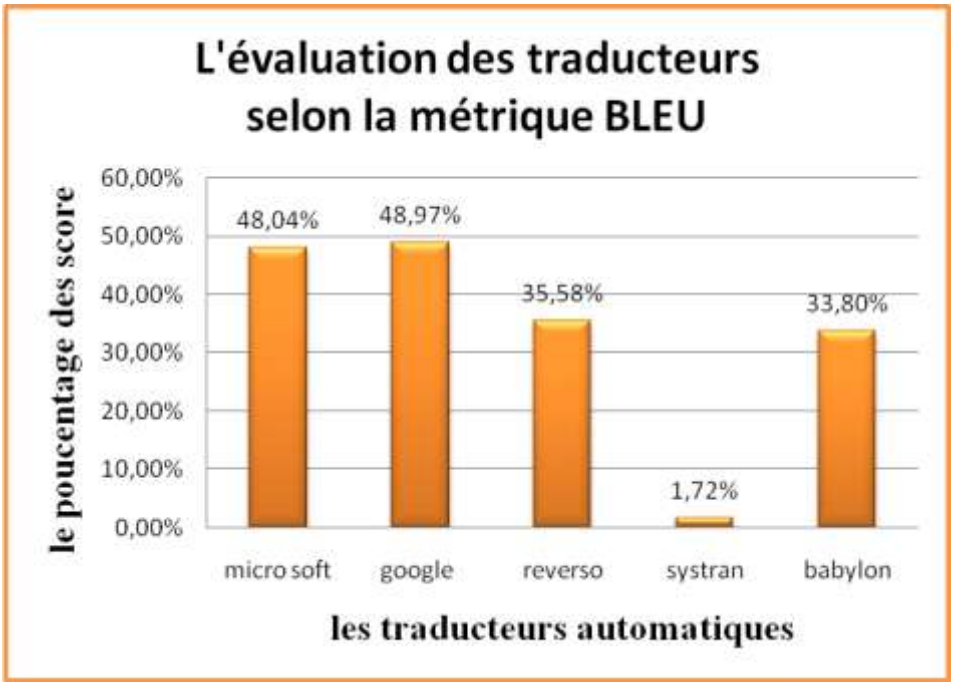


Figure 4.7 : l'évaluation des traducteurs en utilisant la métrique BLEU

3.1.2. Analyse critique

D'après les résultats obtenus de l'évaluation des cinq (05) traducteurs online en utilisant la métrique BLEU, on peut constater que en terme de performance les traducteurs ne donne pas une plein satisfaction vu que le meilleur score ne dépasse pas 0.5 (50%). On peut interpréter cette faiblesse par la stratégie de la métrique BLEU qui se base sur une tâche principale la comparaison des n-grammes de l'hypothèse avec les n-grammes de la référence et de compter le nombre d'équivalences.

Cependant, si on veut donner un ordre de classement des traducteurs, en première position on trouve Microsoft et Google ayant obtenus presque les mêmes score (respectivement, 48,04%, 48,97%), en s'adoptant une stratégie statistique.

En troisième position avec un score inferieure a 0.5 (50%) on a Reverso avec son architecture à base de règles Puis Babylon un score 0.31 (31%) qui adopte l'approche hybride entre la stratégie à base de règles ("rule-based") et l'approche statistique. et enfin en dernière position le traducteur Systran avec un score qui ne dépasse même pas le 0.05 (5%).

3.2. L'évaluation des traducteurs selon la métrique NIST

Dans cette expérience nous évaluons chaque traducteur en utilisant la métrique NIST, nous donnons une partie des résultats dans le tableau suivant :

3.2.1. Les résultats

	Babylon	Systran	Google	Microsoft	Reverso
1	3.68	0.97	2.35	3.68	3.68
2	3.45	1.13	3.33	3.39	3.67
3	4.43	1.24	3.16	4.6	6.27
4	3.06	0.5359	2.4	3.09	3.15
5	3.6	1.09	2.61	3.54	4.17
6	4.24	0.3187	0.66	0.52	0.93
7	3.85	1.27	3.27	4.07	4.45
8	4.8	0.88	2.32	3.72	3.45
9	0.22	0.995	3.01	3.6	5.54
10	4.06	0.8928	3	4.43	4.06
11	4.08	0.77	4.16	4.16	6.21
12	4.69	1.76	4.06	4.35	4.35

Tableau 4.4 : Le score NIST pour les traducteurs (Babylon, Systran, Reverso, Microsoft, et Google traduction)

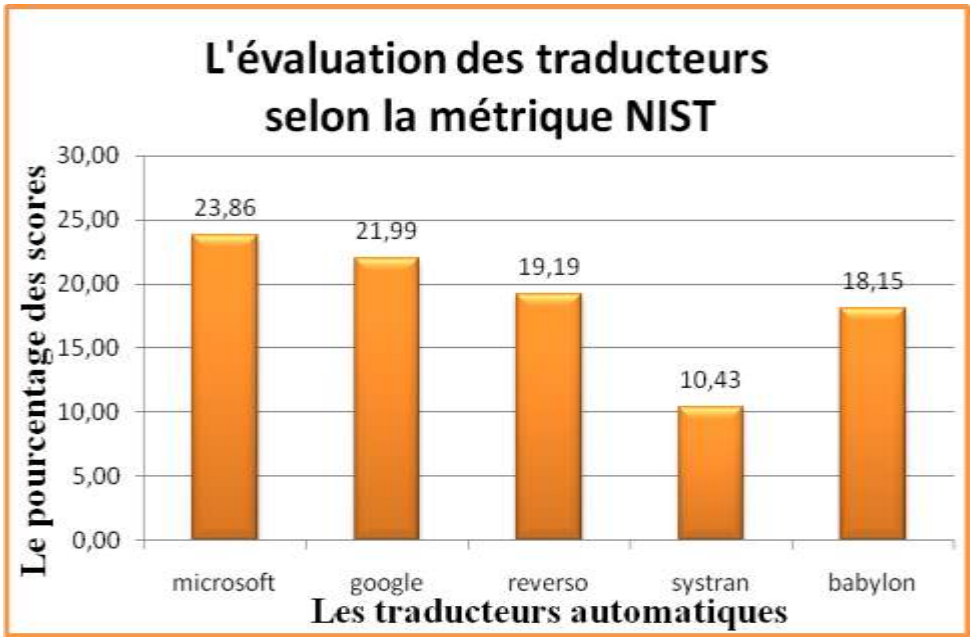


Figure 4.8 : l'évaluation les traducteurs en utilisant la métrique NIST

3.2.2. Analyse critique

Selon les résultats obtenus de l'évaluation des cinq (05) traducteurs online en utilisant la métrique NIST, on peut dire qu'en termes de performance les traducteurs ne donnent pas une pleine satisfaction vu que le meilleur score ne dépasse pas 0.3 (30%). On peut interpréter cette faiblesse, que le score NIST est tout comme le score BLEU, repose sur la précision n-gramme. Cependant, il considère, non pas la moyenne géométrique des n-grammes communs à la traduction automatique et à la référence comme le fait BLEU, mais la moyenne arithmétique.

Donc, si on veut donner un ordre de classement des traducteurs, en première position on trouve les deux traducteurs qui s'adoptent l'approche statistique Microsoft et Google ayant obtenus presque les mêmes scores.

Puis en deuxième position on trouve les traducteurs Reverso et Babylon avec a des valeurs ne dépasse (19%). et enfin en dernière position le traducteur Systran avec un score qui ne dépasse même pas le (10%)

3.3. L'évaluation des traducteurs selon la métrique METEOR

En appliquant la métrique METEOR sur les traducteurs et pour chaqu'un de ces dernier on fait un test. On trouve les résultats suivant :

3.3.1. Les résultats

	Babylon	Systran	Google	microsoft	Reverso
1	0.941	1	0.3779	1	1
2	0.8702	0.0693	0.8444	0.8768	0.8991
3	0.9626	0.06	0.6925	0.9638	0.6054
4	0.8268	0.1538	0.6011	0.8212	0.7476
5	0.8097	0.15	0.677	0.9612	1
6	0.9722	0.49	0.014	0.18	0.4964
7	0.8863	0.15	0.7956	0.9419	0.6174
8	1	0.18	0.5908	0.714	0.6052
9	0.47	0.14	0.8896	0.8758	0.6151
10	0.8993	0.236	0.7096	0.9762	0.8993
11	0.8153	0.1956	0.8446	0.8446	0.5071
12	1	0.0536	0.9486	0.9313	0.9313

Tableau 4.5 : le score METEOR pour les traducteurs automatiques

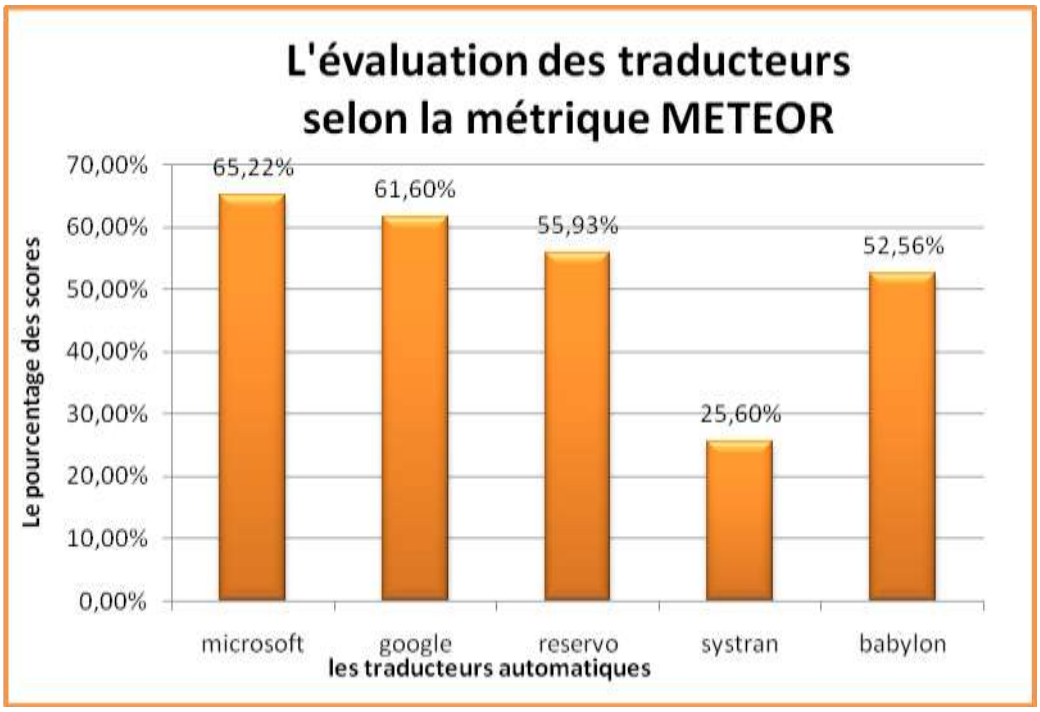


Figure 4.9 : l'évaluation les traducteurs en utilisant la métrique METEOR

3.3.2. Analyse critique

D'après les résultats obtenus de l'évaluation des cinq (05) traducteurs online en utilisant la métrique METEOR. On peut dire qu'en termes de performance les traducteurs donnent une meilleure satisfaction vu que les scores dépasse 50% sauf pour un seul. On peut interpréter ces résultats par la stratégie de la métrique METEOR qui s'utilise les synonymes et le lemme pour calculer leur score.

En suite l'ordre de classement des traducteurs, montre que Microsoft en première position, puis Google en deuxième position, avec des scores dépassant 0.6 (60%), après en troisième et quatrième position on a successivement Reverso avec son architecture a base de régle et Babylon qui adopte l'approche statistique avec des scores qui dépasse 0.5 (50%), en dernière position Systran avec un score inferieur à 0.3 (30%).

3.4. L'évaluation des traducteurs selon la métrique TER

En appliquant la métrique TER sur les traducteurs et pour chaqu'un de ces dernier on fait un test. On trouve les résultats suivant :

3.4.1. Les Résultat

	Babylon	Systran	Google	Microsoft	Reverso
1	0.0588	0.58	0.7058	0.0588	0.0588
2	0.317	0.1951	0.4634	0.3414	0.317
3	0.1	1	0.52	0.08	0.44
4	0.3636	0.10909	0.7555	0.3863	0.6888
5	0.2564	0.88	0.4411	0.147	0
6	0.16	0.3703	1	0.33	0.4637
7	0.1304	1	0.5942	0.1449	0.72
8	0	1	0.82	0.48	0.6151
9	1	1	0.3538	0.3484	0.1578
10	0.1578	1	0.5789	0.0789	0.5555
11	0.8153	1	0.2592	0.2592	0.1219
12	0	1	0.0975	0.1219	0.0526

Tableau 4.6 : le score TER pour les traducteurs automatiques.

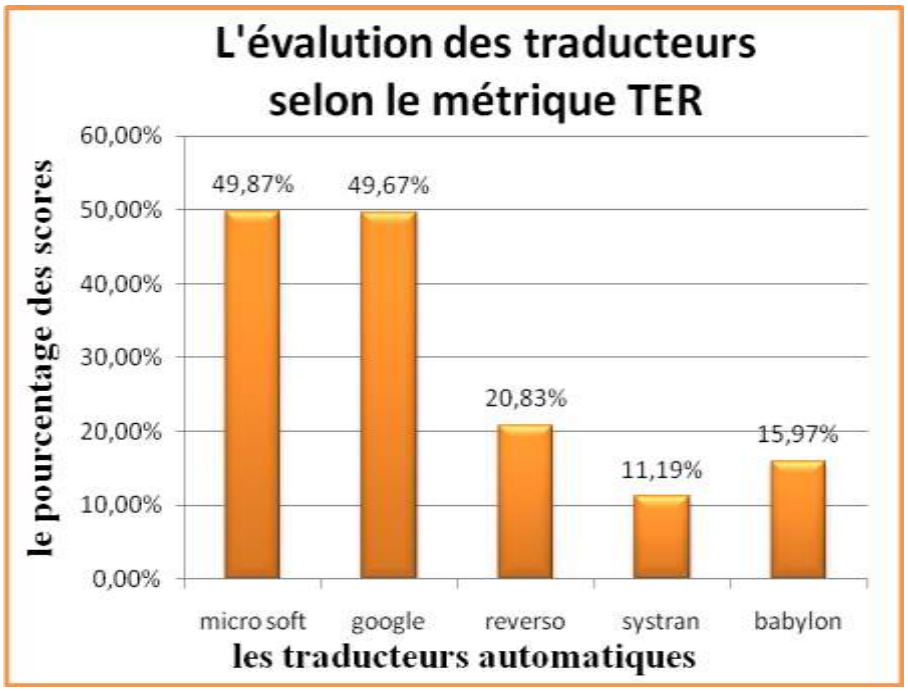


Figure 4.10 : l'évaluation des traducteurs en utilisant la TER

4.3.2. Analyse critique

D'après les résultats obtenus de l'évaluation des cinq (05) traducteurs online en utilisant la métrique TER, on remarque qu'en terme de performance les traducteurs donne une moyenne satisfaction vu que le meilleur score atteint 0.5 (50%). On peut interpréter cette amélioration par la technique de la métrique TER qui calcule la distance entre l'hypothèse et la référence.

En effet, si on veut donner un ordre de classement des traducteurs, en première position on trouve Microsoft et Google ayant obtenus presque les mêmes scores qui s'adaptent l'architecteur statistique.

Le score varie entre 0.2 (20%) et 0.1 (10%) pour les trois derniers traducteurs ; en troisième position Reverso aves son architecture a base de régle, en quatrième position Babylon qui adopte l'approche statistique, en cinquième et dernière position on a Systan.

3.5. L'évaluation des traducteurs selon la métrique WER

En appliquant la métrique WER sur les traducteurs et pour chaqu'un de ces dernier on fait un test. On trouve les résultats suivant :

3.5.1. Les résultats

	Babylon	Systran	Google	microsoft	reverso
1	0.0588	0.76	1	0.0588	0.0588
2	0.9268	0.2439	0.9512	0.856	0.9268
3	0.04	1	0.94	0.08	0.46
4	0.4545	0.1059	0.8	0.6136	0.6888
5	0.4615	1	0.911	0.2352	0
6	0.32	0.4074	1	0.33	0.48
7	0.1304	1	0.74	0.1449	0.4782
8	0	1	0.8	0.84	0.82
9	1	1	0.4923	0.5	0.4696
10	0.1578	1	0.5526	0.0789	0.1578
11	0.6543	1	0.2592	0.2592	0.6543
12	0	1	0.2926	0.1219	0.1219

Tableau 4.7 : le score WER pour les traducteurs (Babylon, Systran, Reverso, Microsoft, et Google traduction)

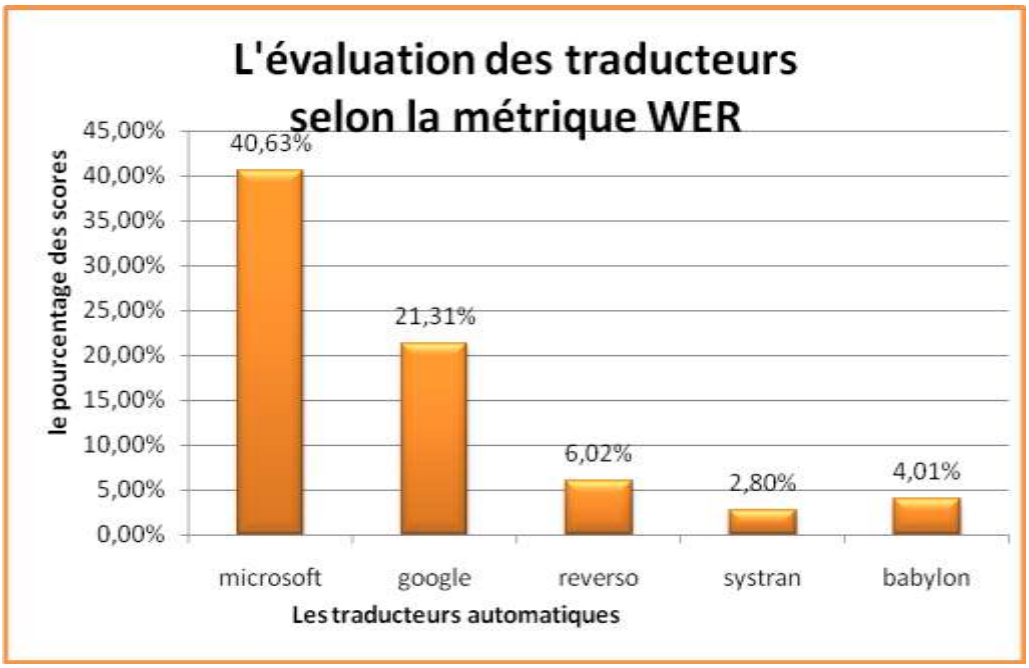


Figure 4.11 : l'évaluation des traducteurs en utilisant la métrique

3.5.2. Analyse critique

Grâce à des résultats obtenus obtenu de l'évaluation des cinq (05) traducteurs online en utilisant la métrique WER, on peut constater qu'en terme de performance les traducteurs ne donne pas une plein satisfaction vu que le meilleur score ne dépasse pas 0.5 (50%). On peut interpréter cette faiblesse la stratégie de cette métrique qui'est évalue la performance du système en termes de taux d'erreur au niveau des mots en utilisant une distance d'édition.

En effet, si on veut donner un ordre de classement des traducteurs, en première position on trouve microsoft avec un score 0.4 (40%) puis google traduction ayant un score 0.21 (21%)

En troiséme position en trouve reverso et babylon a des scores (respectivement, 0.6, 0.4), en dernière position on a systan.

3.6. L'évaluation du traducteur Google par les cinq métriques d'évaluation

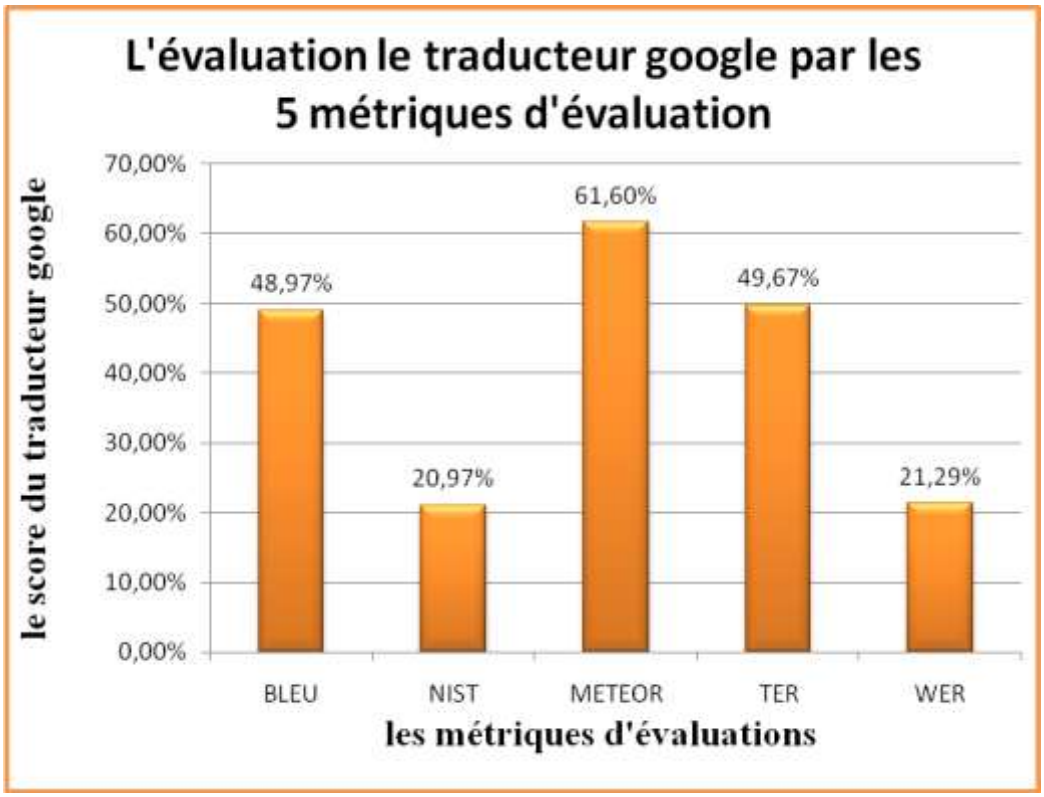


Figure 4.12 : l'évaluation du traducteur Google avec les 5 métriques

3.7. L'évaluation du traducteur Systran par les cinq métriques d'évaluation

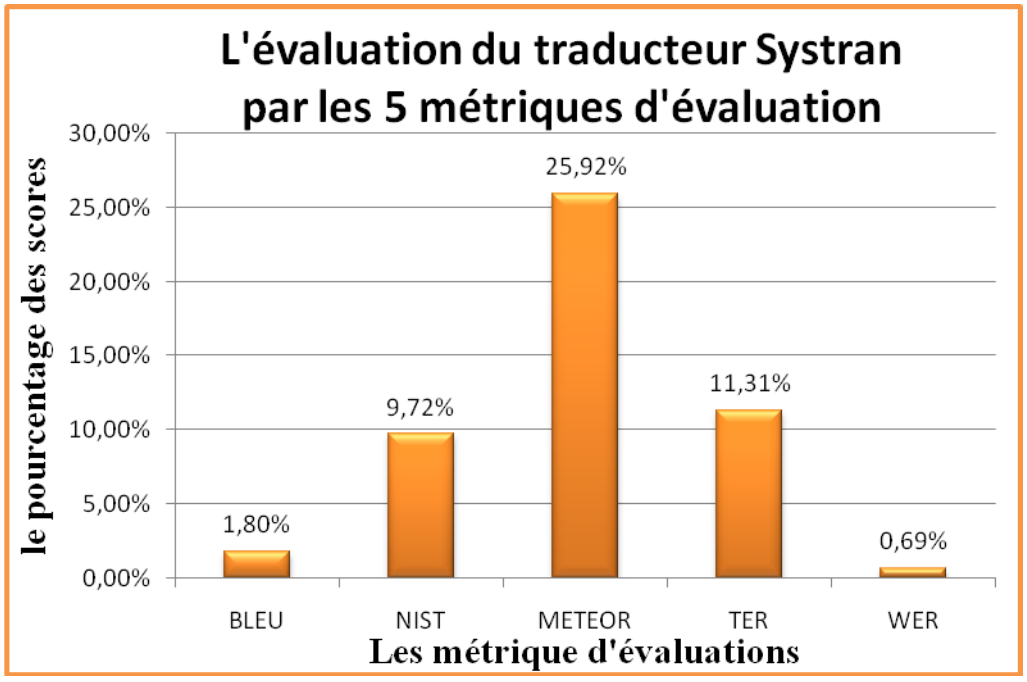


Figure 4.13 : l'évaluation du traducteur Systran par les 5 métrique

3.8. L'évaluation du traducteur Microsoft par les cinq métriques d'évaluation

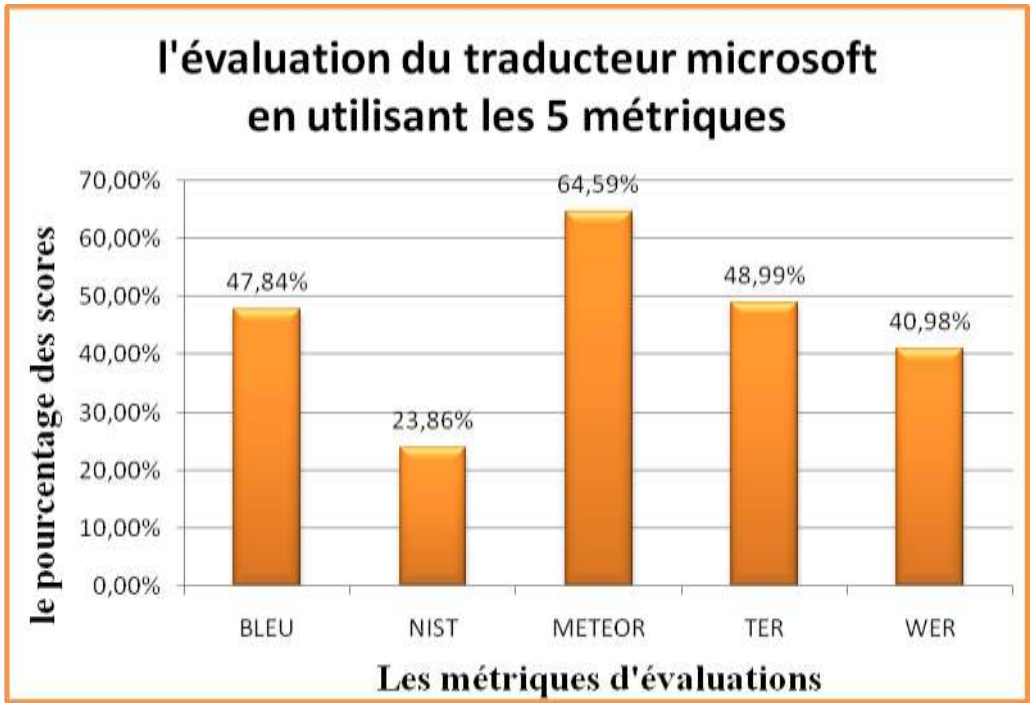


Figure 4.14 : l'évaluation du traducteur Microsoft par 5 métrique

3.9. L'évaluation du traducteur Babylon par les cinq métriques d'évaluation

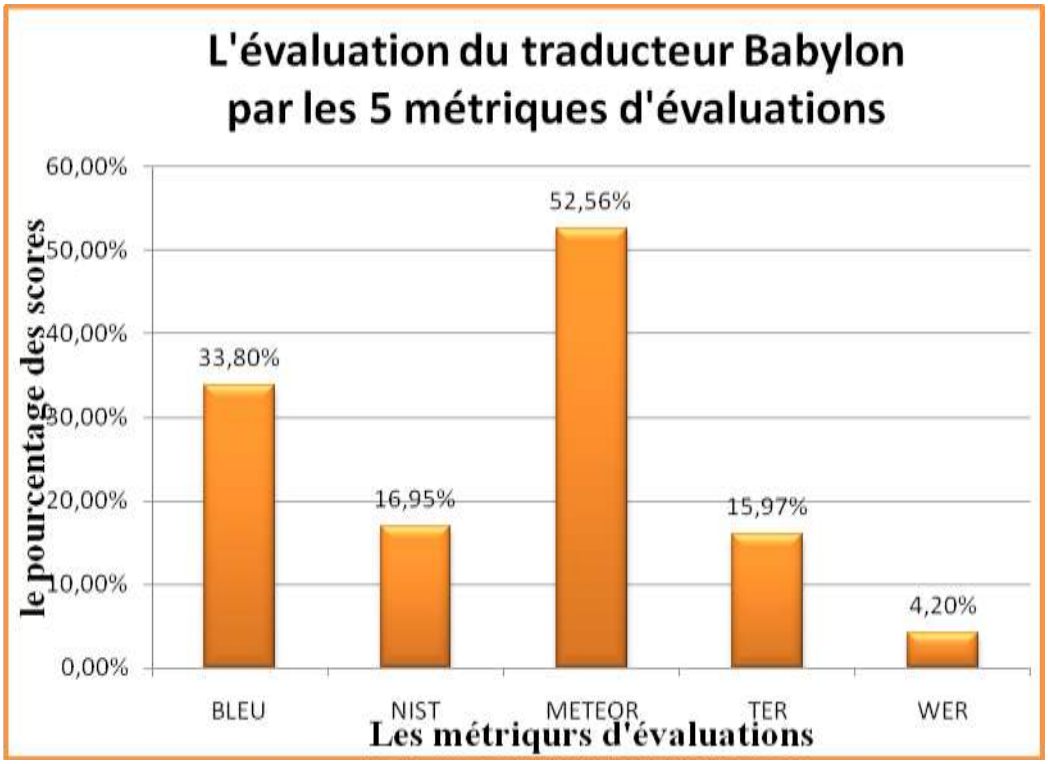


Figure 4.15 : l'évaluation du traducteur Babylon par les 5 métriques

3.10. L'évaluation du traducteur Reverso par les cinq métriques d'évaluation

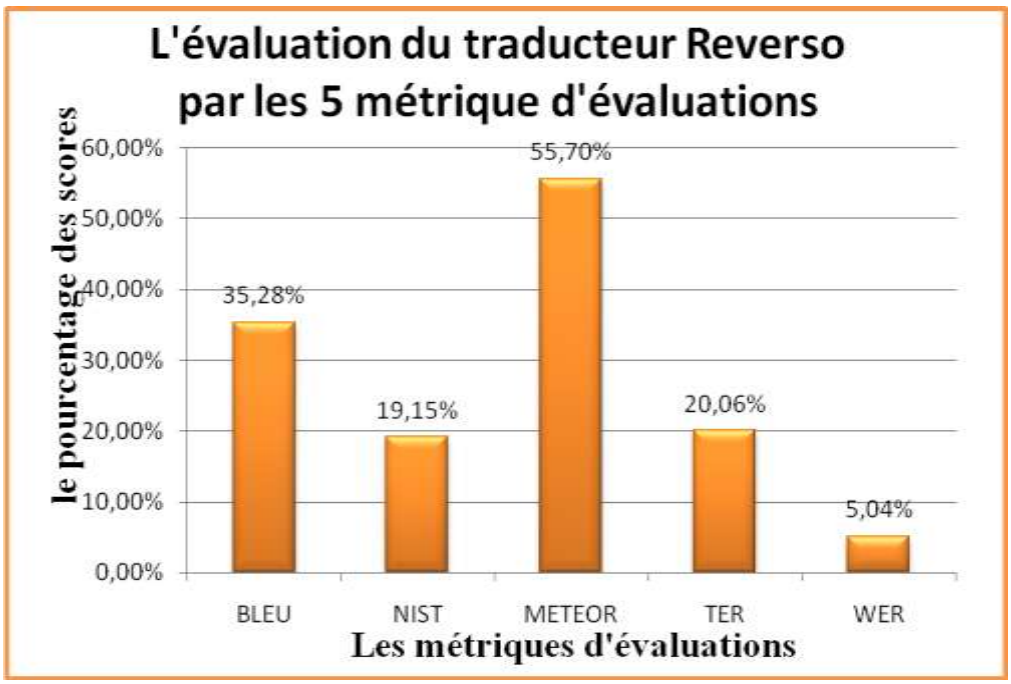


Figure 4.16 : l'évaluation le traducteur Reverso avec les 5 métrique.

4. Conclusion

D’après l’analyse et les résultats obtenus on peut dire que les traducteurs Microsoft et Google traduction sont les meilleurs traducteurs avec les cinq métriques d’évaluation si pour cela elles sont plus utilisés dans l’université et l’éducation.

Par contre le traducteur Systran est le moins bon, avec les cinq métriques d’évaluation utilisée dans le cas la traduction anglais vers arabe.

Conclusion générale (Bilan et Perspectives)

1. Bilan

Pou Conclure, on peut dire que les mesures automatiques ne remplacent pas le jugement humain pour évaluer la qualité d'une traduction. En effet, la langue est tellement riche et complexe que les scores automatiques sont incapables d'en discerner les nombreuses subtilités et en particulier en traduction. Mais confronter une traduction automatique à une ou plusieurs références limite considérablement le nombre important de façons différentes d'exprimer une idée sans en modifier le sens. Toutefois, le choix de bonne mesure reste toujours le meilleur moyen et rapide d'évaluer une traduction et la performance de cette traduction selon le score correspondant comme c'est aussi le meilleur moyen de différencier une bonne d'une mauvaise traduction.

D'après notre étude des ces critères et mesures on peut faire le bilan suivant : Le score obtenu en utilisant la métrique METEOR est le meilleur et plus performant car il se base sur les **synonymes** et **lemme**, cependant il faut retenir que le dictionnaire des synonymes n'est pas disponible pour toutes les langues.

Mais en ce qui concerne les métriques TER et WER de préférence il faut calculer la différence entre le texte original et le texte traduit. Mais ces mesures ne permettent pas vraiment de juger si une traduction est acceptable ou non du fait qu'aucune reformulation de la traduction de référence n'est acceptée.

Le poids utilisé dans le calcul du score NIST qui représente l'importance du **n-gram** dans la traduction a toujours besoin de beaucoup de références originales. Comme on a distingue qu'il ya une bonne concordance avec le score BLEU et l'évaluation humaine :

Le score de la métrique **BLEU** utilise le **N-gram** dans le calcul depuis **1-gram** qui donne les mots utilisés et le **4-gram** qui garantie en quelque sorte le sens du mot traduit.

Pour ce qui de notre évaluation du traducteur on a distingué que c'est celui qui utilise les cinq critères à savoir Google traduction et MICROSOFT qui se base dans la traduction sur une approche statistique.

2. Perspectives

- Apporter des modifications aux métriques d'évaluation :
  - Améliorer l'évaluation de la Métrique BLEU dans le N-gram en augmentant la valeur n pour atteindre n=6 afin de garantir la sauvegarde de la donnée et le sens en utilisant les synonymes et les origines des mots.
  - Introduire des concepts apportant plus de sémantique.
- Inverser le sens de la traduction (Anglais / Arabe au lieu de l'Arabe / Anglais).
- Travailler sur d'autre corpus d'évaluation pour constater les différences de résultats (en particulier les textes de spécialité)

Références Bibliographiques

[ANN, 2003] Annik Baumgartner-Bovier, « La traduction automatique, quel avenir ? Un exemple basé sur les mots composés », *Cahiers de Linguistique Française N°25 :* Temporalité et causalité, 2003.

[DAN, 1999] Daniel Jurafsky and James H. Martin, « Speech and Language Processing », edition :Prentice Hall, 1999.

[HUT, 1992] W. John Hutchins, L. Harold Somers, « An Introduction to Machine Translation », edition: ACADEMIC PRESS, 1992.

[SAY, 1999] SayoriShimohata, Toshiki Murata, Atsushi Ikeno, Tsuyoshi Fukui and Hideki Yamamoto, « Machine Translation System PENSÉE: System Design and Implementation », the 7th Machine Translation summit, September 1999, Singapore. URL: http://www.mt-archive.info/MTS-1999-Shimohata.pdf

[YVE, 07] Yves Lepage, Etienne Denoual,« ALEPH: an EBMT systembased on the preservation of proportional analogiesbetween sentences across languages ». URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.6502&rep=rep1&type=pdf

[SLI, 2008] YojiFukumochi, « A Way of Using a Small MT System in Industry », the 5th*Machine Translation Summit,* July 10-13, *1995,* Luxembourg. URL: http://www.mt-archive.info/MTS-1995-Fukumochi.pdf

[BIR, 2011] Doctor of Philosophy :Alexandra Birch ,Reordering Metrics for Statistical Machine Translation, School of Informatics- University of Edinburgh, 2011.

[SNO, 2006] M. Snover, B. Dorr, , R. Schwartz, L. Micciulla, J. Makhoul.: "A Study of Translation Edit Rate with Targeted Human Annotation". In Proceedings of AMTA, Boston, 2006.

[CHI, 2012] Chiheb Trabelsi , Traduction statistique vers une langue à morphologie riche : Combinaison d'algorithmes de segmentation morphologique et de modèles statistiques de traduction automatique, Mémoire de l'obtention

du grade de Maitrise ès Science (M. Sc.) en Informatique , Université de Montréal , Juillet, 2012