L’intelligence artificielle et la traduction générée automatiquement

Lingua e traduzione – lingua francese – corso avanzato (RISE-SA), a.a. 2023-2024

Prof.ssa Alida Maria Silletti

SOURCES: François Yvon. « Les deux voies de la traduction automatique ? ». *Hermès, La Revue- Cognition, communication, politique*, CNRS-Editions, 2019, 85, pp.62-68 ;

Nathalie Kübler, *La traduction automatique : traduction machine?*, https://core.ac.uk/download/pdf/47087967.pdf ;

Franck Barbin, « La traduction automatique neuronale, un nouveau tournant ? », *Palimpseste. Sciences, humanités, sociétés*, 2020, 4, pp. 51-53, https://shs.hal.science/halshs-03603588/document ;

https://www.lebigdata.fr/classement-meilleur-traducteur-automatique ; https://www.codeur.com/blog/outils-traduction-gratuite/

AUX ORIGINES DE LA TRADUCTION AUTOMATIQUE (TA)

- La traduction automatique (TA) est le domaine de recherche à la base du traitement automatique du langage (TAL)

1940-1960 :

- Les premiers ordinateurs permettent aux belligérants de déchiffrer leurs codes respectifs et on tente d'appliquer ces techniques de déchiffrage à la traduction automatique ;
- dans les premiers systèmes de TA, les connaissances en syntaxe et en analyse syntaxique sont encore très insuffisantes, les ordinateurs ont des capacités de stockage très limitées et sont peu puissants ;
- des systèmes de mise en correspondance de dictionnaires sont créés, ce qui génère des résultats de traduction voués à l'échec, incapables de traduire plus de quelques phrases
- 1954 : le premier système de TA, financé par les gouvernements russe et américain, est présenté au public ; il permet de traduire 49 phrases russes, sélectionnées au préalable, vers l'anglais se servant d'un dictionnaire de 250 mots et de six règles de grammaire ;
- 1957-1962 : des modèles syntaxiques de la langue sont créés, par lesquels une description syntaxique de la langue donne une structure aux phrases, indispensable pour les systèmes de traduction automatique

AUX ORIGINES DE LA TA

1966 : le rapport ALPAC

- le gouvernement des États-Unis demande à la commission ALPAC (Automatic Language Processing Advisory Committee) un rapport sur la TA : elle est plus lente, moins efficace et deux fois plus chère que la traduction faite par des humains et on recommande d'arrêter de financer la recherche dans ce domaine – la recherche en TA est temporairement arrêtée aux États-Unis

1967-1976 :

- d'autres pays continuent à développer la recherche en TA ;
- aux Etats-Unis la seule activité porte sur la traduction du russe en anglais de textes scientifiques et techniques ;
- 1970 : au Canada, le projet TAUM-Météo mène au développement ultérieur du langage de programmation PROLOG et s'utilise avec succès dans le domaine restreint de la traduction des prévisions météorologiques ;
- en France, on développe à Grenoble un système permettant de traduire du russe des textes mathématiques et physiques ;
- 1976 : le système Systran, produit par Peter Toma, aux États-Unis, est mis en place à la Commission Européenne- il continue à être actuellement utilisé

AUX ORIGINES DE LA TA

Années 1980 :

- à partir de la première version de Systran, qui consiste à traduire du russe vers l'anglais, de nouvelles versions auprès de la Commission européenne (français-anglais, et vice-versa ; anglais-italien, et vice-versa ; français-italien, et vice-versa…) et de l'OTAN sont créées et installées
- des systèmes concurrents de Systran sont développés : Logos et METAL
- les premiers systèmes multilingues sont créés : le projet européen EUROTRA est basé sur un système à transfert multilingue, le Distributed Language Translation system à Utrecht (construit sur un système à pivot, à savoir en passant par une interlangue, qui était l'espéranto)

AUX ORIGINES DE LA TA

Années 1990 :

- des mémoires de traduction, indispensables pour la traduction professionnelle, commencent à dominer le marché
- la montée en puissance des ordinateurs personnels permet aux systèmes commerciaux de TA de produire des systèmes individuels que l'on peut installer sur ordinateur
- le développement d'Internet permet de mettre en ligne des systèmes de TA : Systran est le premier à s'en servir en 1998 avec Altavista/Babelfish
- la recherche en traduction automatique progresse : des systèmes à transfert sont utilisés, où un module d'analyse de la langue source, un module de règles de transfert complexes entre la langue source et la langue cible et un module de génération de la langue cible interagissent à partir de règles syntaxiques et de très grands dictionnaires
- l'apport de la recherche en linguistique de corpus qui décrit la langue à partir des données concrètes du corpus permet de travailler sur le lexique et d'y ajouter des informations syntaxiques toujours plus nombreuses

AUX ORIGINES DE LA TA

- les outils de traitement statistique de corpus développés en traitement automatique du langage permettent l'émergence d'une nouvelle direction de recherche : la traduction automatique « basée sur les exemples », qui permet de dégager les traductions les plus fréquentes en effectuant des analyses statistiques sur des corpus traduits et alignés
- la traduction de l'oral représente un défi passionnant pour le domaine du traitement automatique du langage – la traduction automatique du langage parlé ou interprétariat automatique
- des projets combinant la reconnaissance de la parole, la traduction automatique et la synthèse de la parole sont créés : le projet allemand Verbmobil (1993-2000), pour réaliser un traducteur oral transportable pour aider les locuteurs germanophones et japonais à mener des négociations commerciales en anglais ; le projet TC-STAR2 dont le but est d'effectuer la reconnaissance vocale des discours au Parlement Européen, traduit le discours et le restitue oralement dans la langue cible

LES APPROCHES ACTUELLES ET LES ATOUTS DE LA TA

- Des systèmes computationnels sont créés, basés sur des architectures de calcul « neuronales », c'est-à-dire qui sont capables d'apprendre à mettre en correspondance une phrase écrite ou prononcée dans une langue « source » avec sa traduction dans une langue « cible » - cela nécessite de disposer d'exemples de telles correspondances, rassemblés dans un corpus parallèle
- Un corpus parallèle est une collecte de données écrites et traduites d'une langues source à une ou plusieurs langues cibles
- L'augmentation continue des capacités de calcul et de stockage permet le déploiement à très grande échelle de ces architectures de calcul et la production quasi instantanée de traductions
- La TA contemporaine se nourrit de grandes masses de données

LES APPROCHES ACTUELLES ET LES ATOUTS DE LA TA

- Si ces données existent déjà, des traductions professionnelles sont opérées pour le compte d'une entité solvable, qui produira des textes spécialisés à des fins de publication et soumis à de fortes exigences de qualité – ex. la direction générale de la traduction de l'Union Européenne, dont les traductions sont disponibles en grande quantité, mais un travail de révision en post-édition sera cependant requis pour satisfaire les niveaux de qualité requis
- Si les données ne sont pas disponibles, cela peut dépendre de diverses raisons : soit il n'existe aucune demande pour des traductions humaines ; soit cette demande n'est pas solvable ; soit les langues ne sont pas disponibles ; soit la notion même de traduction n'est pas bien définie, par exemple, si les textes à traduire sont des commentaires postés sur des réseaux sociaux – le résultat : des traductions approximatives ou médiocres … bien que la traduction atteigne son but de faire comprendre

TA : C'EST DE LA TRADUCTION ? (YVON 2019)

- Les analyses de corpus permettent de comparer des traductions humaines et des traductions automatiques
- Les textes traduits automatiquement sont en moyenne plus longs que les textes sources
- La diversité lexicales (le rapport entre types et occurrences) y est plus réduite
- D'une manière générale, une traduction automatique devrait présenter les mêmes caractéristiques qu'une traduction littérale : une moindre variété lexicale, une sur-représentation des vocables très fréquents, une tendance à reproduire l'ordre des mots des textes sources, auxquelles s'ajoutent des erreurs propres aux traductions automatiques, comme la méconnaissance de termes ou d'idiomes
- Ces traductions restent imparfaites et nécessitent toujours une intervention humaine

Ces remarques sur la TA sont-elles encore valables en 2023 ?

LES LOGICIELS DE TA

- Un logiciel de traduction automatique automatise le processus de traduction de texte d'une langue, source, à une autre langue, cible
- Si l'outil le plus basique repose strictement sur la substitution mot-à-mot, certaines technologies incluent une traduction basée sur des règles ou modélisée statistiquement pour des traductions plus précises
- Puisque les traducteurs automatiques peuvent traiter et traduire de grandes quantités de texte presque instantanément, ces outils se révèlent particulièrement utiles en pré-traduction – le texte pré-traduit passe ensuite entre les mains des traducteurs qui effectueront une relecture et une révision
- Chaque système de TA a son propre algorithme qui décrit la composition sémantique et syntaxique d'un langage
- Deux types de systèmes de traduction automatique : la traduction automatique statistique et la traduction automatique dite neuronale
- 2015 : passage de la traduction automatique statistique (TAS) à la traduction automatique neuronale (TAN)

LA TAS

- Elle puise ses ressources dans d'énormes caches de données contenant des traductions préalablement approuvées pour une paire de langues spécifique, au sein de corpus bilingues
- Elle est donc basée sur des dictionnaires et des modèles statiques construits à partir de corpus découpés en unités de traduction
- La machine utilise un modèle probabiliste pour analyser les corpus et établir des parallèles entre les structures de deux langues
- La « meilleure » traduction est ainsi le résultat d'une « prédiction »

LA TAN

- Elle fonctionne sur le modèle du cerveau humain au travers de l'intelligence artificielle – un réseau neuronal artificiel qui ressemble étroitement au cerveau humain à la façon d'y stocker les informations
- Un système de ce type organise les données linguistiques en groupes et couches complexes
- Résultat : la TAN permet de générer des traductions de meilleure qualité (plus précises et naturelles) qui se rapprochent de la traduction humaine
- La traduction est le résultat de bases de données bilingues existantes et également de son apprentissage automatisé, qui se nourrit de nouvelles informations – les moteurs de traduction automatique sont désormais basés sur l'auto-apprentissage
- L'idée est que ce système soit plus rapide, plus nuancé et plus précis dans la traduction de textes que TAS
- Le logiciel est capable d'apprendre en temps réel, il s'améliore au fur et à mesure que l'on insère des segments à traduire
- Plus on alimente cet apprentissage avec un volume considérable de données de qualité bilingues, plus on améliore la fiabilité des résultats obtenus

LA TA AUJOURD'HUI

- La TA gagne du terrain dans le secteur de la traduction, pour répondre aux exigences de coûts et de délais des clients - les économies engendrées par la post-édition l'emportent souvent sur la baisse de qualité du texte final
- Même si la traduction automatique exerce une pression croissante sur la traduction humaine, il revient à la traduction humaine de faire valoir son expertise linguistique et disciplinaire pour démontrer sa « valeur ajoutée » par rapport à de tels systèmes
- Post-édition : l'étape durant laquelle on relit le texte produit par la traduction automatique et on le corrige pour éliminer les erreurs sémantiques et linguistiques – des éléments tels que l'exactitude sémantique, la cohérence terminologique, l'harmonisation du style ou l'adaptation culturelle demandent une vérification et correction humaines

- Google Traduction est un traducteur gratuit de Google, à la fois disponible sur ordinateur et sur smartphone
- Il dispose de 109 langues
- La traduction (neuronale) est effectuée de façon unidirectionnelle uniquement
- Son interface est propre et simple, avec la possibilité de disposer également d'une fonctionnalité hors ligne
- Dans l'application mobile, il est possible de
  - traduire n'importe quel texte saisi dans l'application (mode Tapez) ;
  - utiliser son écran tactile pour écrire à la main n'importe quel texte et le traduire dans l'application (mode Écrivez) ;
  - traduire un texte sonore en utilisant le microphone du smartphone (mode Parlez) ;
  - reconnaître et traduire des photos d'un texte (mode Photo/Snap) ;
  - utiliser l'application pour traduire le texte en temps réel tout en enregistrant une vidéo (mode Voir)

DeepL

- Outil de traduction gratuit mis à disposition par la société allemande DeepL GmbH (également à la tête de *Linguee*)
- Les réseaux de neurones de DeepL s'entraînent sur des milliards de segments de traduction de haute qualité du moteur de recherche de *Linguee*, lui permettant de fournir des traductions toujours plus précises et naturelles, selon l'apprentissage profond (*deep learning*, à l'origine du nom de DeepL)
- Concrètement, il peut traduire 1 million de mots en moins d'une seconde
- 72 paires de langues possibles
- Selon le test Bilingue Evaluation Understudy (test BLEU), DeepL se révèle être 3 fois plus efficace que Google traducteur
- DeepL est considéré comme le service de traduction automatique le plus efficace du marché

- Un autre outil très connu et performant de la traduction gratuite en ligne
- Il a l'avantage d'accompagner ses traductions d'exemples d'utilisation, permettant de contextualiser la phrase traduite
- Reverso permet de traduire du français vers l'anglais, l'allemand, l'espagnol, l'italien, le néerlandais, le portugais, le russe, le chinois, l'arabe, l'hébreu et le japonais