Thèmes

Thème 1:	Thème 2:	Thème 3:
Traduction automatique	Qualité des traductions	Optimisation des processus
Optimiser la qualité et le déploiement de la TA au BtB	Aider les langagiers à assurer la qualité des textes gouvernementaux dans les deux langues officielles	Utiliser intelligemment les données existantes pour optimiser les processus du BtB

Thème 1: Traduction automatique

[OCL]
Outil de compréhension linguistique
- Portage (TA stat.)

[HF2]
TA incrémentale
- Essai au Parlementaire (automne 2019)
- Portage (TA stat.)
- TA neuronale

[HF3]
TA neuronale adaptative

Thème 2: Qualité des traductions

[HTQ1]
Nettoyage du Mégacorpus
- *WMT 2019 shared task on Parallel Corpus Filtering*
- Améliorations substantielles de qualité de la TA

[MTQ1]
Évaluation de la qualité de la TA
- Systèmes BtB expérimentaux
- modèles neuronaux profonds (BERT)
- *WMT 2019 shared task on MT Metrics*

[MTQ2]
Estimation de la qualité de la TA
- modèles neuronaux profonds bilingues (XLM)

[HTQ2]
Lisibilité des traductions

Thème 3: Optimisation des processus

[HF4]
Analyse des commandes
- Prédiction des ressources (traducteurs)
- Examen préliminaire des données
- Classification automatique par spécialités / sous-spécialités

[HF1]
Recherche intelligente dans le Mégacorpus

CNRC·NRC

MERCI

Michel Simard – Michel.Simard@nrc-cnrc.gc.ca

AI for Translation Quality

Project Update, with Revised Deliverables and Calendar

20 September 2019

Theme I: Machine Translation Quality (MTQ)

Package MTQ1 ("*MTEval*"): MT Quality Evaluation [NRC]

With the goal of better aligning with the needs and priorities of the Bureau, we propose to steer this project in the direction of supporting the Bureau with their ongoing evaluation of various MT offers (DeepL, Iconic, Reverso, Portage, etc.). The main goal of the project would then be to develop a protocol for comparative evaluation of MT systems, well adapted to the needs of the Bureau and their clients. That protocol would combine human and automatic evaluation methods, and would focus on producing quantifiable results. The YiSi automatic evaluation metric would still be part of the deliverables, but the project would also include human evaluation methods.

Deliverables

Name	Description	Date (end of month)
MTQ1.1:	Report on fine-resolution human and automatic comparative MTQ evaluation	10.2019
MTQ1.2:	MTEval system	12.2019

Package MTQ2 ("*MTQuEst*"): MT Quality Estimation [NRC]

Following encouraging results of YiSi-2 in WMT 2019 Quality Estimation shared task, we are currently working on developing BERT/XLM-based YiSi QE custom trained on BT data. The current plan is to deliver that technology to the Bureau before the end of FY2019-20. However, with little MT currently deployed and systematically used at the Bureau, we are faced with the challenge of obtaining usable feedback from Bureau translators on which to base customization or improvements. One place where we believe we are likely to see MT being

deployed and used in a systematic fashion in FY2020-21 is at Parliamentary Operations, as it is part of our project on *Incremental MT* (HF3). Therefore, our proposal is to integrate MTQuEst there.

Deliverables

Name	Description	Date (end of month)
MTQ2.1:	MTQuEst baseline system	03.2020
MTQ2.2:	Report on advanced MTQ estimation	03.2020
MTQ2.3:	MTQuEst on-site integration	09.2020
MTQ2.4:	Report on MTQuEst integration	12.2020

Theme II: Human Translation Quality (HTQ)

Package HTQ1 ("*Corpus Cleaning*"): Computer-assisted HTQ Control [RALI]

After some initial delays, the Bureau has now provided RALI and NRC with access to its Megacorpus content, *Bitexte* application (which allows access to documents in their original format) and the original "flagged" documents from the Megacorpus (documents known to contain alignment and/or translation errors). Using this data, RALI has completed the creation of a small subset of the Megacorpus, manually-annotated for alignment and translation errors, and is currently using that to develop and test high-precision automatic error detection. From there, the objective is to bootstrap semi-automatic creation of a much larger dataset. This dataset will then be used to train large-scale classifiers. In parallel, RALI and NRC are testing various unsupervised corpus cleaning approaches (methods that don't rely on annotated data), and building and evaluating baseline neural MT systems from the Megacorpus, as a way of measuring corpus cleaning effectiveness.

Deliverables

Name	Description	Date (end of month)
HTQ1.1:	Report on alignment errors in the Megacorpus	12.2019
HTQ1.2:	Report on Automatic HT error detection	10.2019
HTQ1.3:	Automatic HT error detection system	03.2020

Package HTQ2 ("*Readability*"): Translation Readability Assessment [NRC]

The deliverables in this project have been somewhat reorganized and simplified, with less reliance on explicit collaboration and feedback from the Bureau, and expanded work on exploring writing assistance tools for producing readable texts.

Deliverables

Name	Description	Date (end of month)
HTQ2.1:	Baseline document-level Translation Readability system	06.2020
HTQ2.2:	Report on the state-of-the-art of Cross-lingual readability assessment	09.2020
HTQ2.3:	Advanced document-level and segment-level Translation Readability system	03.2021
HTQ2.4:	Final report, with an outlook on automatic assistance for readable writing	03.2021

Theme III: Human Translator Feedback (HF)

Package HF1 ("*TM Matching*"): Advanced Translation Memory Matching [RALI]

This project was initially planned to begin in January 2020, but because the outcome of the *Corpus Cleaning* project (HTQ1) is a prerequisite, kick-off was delayed until April. There are also potential synergies between this and domain/subdomain prediction from text in the *Translation Assignment* project (HF4).

Deliverables

Name	Description	Date (end of month)
HF1.1:	Report on Advanced TM Matching	09.2020
HF1.2:	Advanced TM Matching system	09.2020

Package HF2 ("*Incremental MT*") Incremental Machine Translation Training [NRC]

This project was initially planned to take place within the METEO team (Weather warning translation), but due to technical and logistical issues, it has been relocated to Parliamentary Operations instead, focusing on Hansard (HoC & Senate debates). The goal is to integrate Portage (with incremental training capability) into the Parliament's *Prisme* environment, so that translators can start working with this technology as soon as the House resumes (possibly in early December 2019). NRC delivered a baseline Portage system to HoC Digital Services for the purpose of integration in mid-July 2019. HoC Digital Services are expected to provide NRC with data (HoC & Senate proceedings of the 1st session of 42nd parliament), from which NRC will build the operational systems. If all goes well, a first evaluation could be completed before the end of FY2019-20.

Deliverables

Name	Description	Date (end of month)
HF2.1:	Evaluation protocol for Incremental MT	11.2019
HF2.2:	Incremental MT Training system (@Parl.GC.CA)	11.2019
HF2.5:	Report on Incremental MT Training and Recommendations	03.2020

Package HF3 ("*Adaptive MT*"): Adaptive NMT Training [RALI]

This project was initially planned to begin in July 2020, but because the outcome of the Corpus Cleaning (HTQ1) and TM Matching (HF1) projects are prerequisites, kick-off was delayed until October 2020. However, training NMT systems has become an integral part of the Corpus Cleaning (HTQ1) project, as MT quality is one of the ways in which we evaluate corpus cleaning. This means that much of the groundwork for this project will already be in place before the project actually begins. There are also potential synergies between this and domain/subdomain prediction from text in the Translation Assignment project (HF4).

Deliverables

Name	Description	Date (end of month)
HF3.1:	NMT System	03.2021
HF3.2:	Report on NMT Integration at the Bureau	03.2021

Package HF4 ("*Assignment*"): Assisted Translation Assignment [NRC]

This project was initially scheduled to begin in April 2019. However, delays in obtaining the data for this project resulted in a delayed start in September 2019. The project has also been expanded to the more general question of analysing source texts to predict not only the most apropriate translator, but also the text's domain/subdomain and, from there, various other resources (TM, Term banks, MT system, etc.).

Deliverables

Name	Description	Date (end of month)
HF4.1:	Preliminary report on domain and resource prediction from text	11.2019
HF4.2:	Prototype Translation assignment assistant	01.2020
HF4.3:	Final report on Automating translation assignment	03.2020

			FY2019-20							FY2020-21											
Package	Name	Lead	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3
MTQ1	MTEval	NRC		report		system															
MTQ2	MTQuEst	NRC		start					system + report						integration			report			
HTQ1	Corpus Clean	RALI		report (HTQ)		report (align)			system												
HTQ2	Readability	NRC								start		base system			report (SOTA)						system + report
HF1	TM Match	RALI								start					system + report						
HF2	Incremental MT	NRC			system + report				report												
HF3	Adaptive MT	RALI														start					system + report
HF4	Translation Assignment	NRC			report		system		report												

Méthodes et outils d'évaluation des traductions automatiques au BtB

Rapport final de projet

Michel Simard
Serge Léger
Centre de recherche en technologies numériques
Conseil national de recherches Canada
prénom.nom@cnrc-nrc.gc.ca

30 janvier 2020

Introduction

Le présent rapport porte sur les travaux entrepris dans le cadre d'un projet sur les méthodes et les outils d'évaluation de la traduction automatique au Bureau de la traduction. Ce projet s'inscrit dans le contexte d'une entente de collaboration entre le Bureau de la Traduction (BtB) et le Conseil national de recherches Canada (CNRC) sur l'utilisation de méthodes d'intelligence artificielle (IA) pour soutenir la qualité des traductions au BtB: *Artificial Intelligence for Translation Quality* (AI4TQ).

Nous décrivons d'abord le contexte de modernisation des processus du BtB qui motive le présent projet, incluant l'utilisation croissante des techniques d'IA, dont la traduction automatique. Nous présentons ensuite un survol des méthodes d'évaluation automatiques de la qualité des traductions automatiques, notamment de la méthode YiSi développée récemment au CNRC. Nous proposons ensuite un protocole d'évaluation des traductions automatiques, couvrant les étapes de sélection de domaine et du jeu de test, de calcul des métriques, et d'interprétation des résultats. Nous présentons finalement ÉQuaTion, la trousse d'évaluation de la qualité des traductions automatiques développée par le CNRC pour le BtB autour de la technologie YiSi.

Contexte

Les récentes avancées en Intelligence Artificielle (IA) ont permis des progrès substantiels en Traduction Automatique (TA). Il est maintenant envisageable pour une organisation comme le BtB d'utiliser la TA de façon systématique dans ses processus. Différentes études semblent en effet indiquer que l'utilisation de la TA permettrait non seulement d'améliorer la productivité des traducteurs, mais également, dans certaines situations, leur permettre de produire des textes de meilleure qualité [Laübli et al. 2019]. Le BtB procède présentement à une modernisation de son flux de travail, et l'intégration graduelle de la TA fait partie intégrante de cet exercice.

Toutefois, on sait qu'une utilisation aveugle de l'IA peut avoir des conséquences catastrophiques. C'est pourquoi, dans une organisation où ces techniques sont déployées à grande échelle, le contrôle de la qualité devient un élément essentiel des opérations. Ainsi, à mesure que le BtB déploie des systèmes de TA au sein des flux de travail, il est impératif qu'il mette également en place des mécanismes pour mesurer en continu la qualité des traductions produites par chacun de ses systèmes de TA.

D'une façon intéressante, l'IA peut également jouer un rôle central dans ces mécanismes de contrôle de la qualité. Depuis une quinzaine d'année, un nombre croissant de chercheurs et de praticiens de la TA s'est intéressé à l'évaluation automatique de la qualité des traductions-machine. Leur travail a mené à la création de plusieurs métriques d'évaluation automatique de la TA, tel que BLEU [Papineni et al., 2002], TER [Snover et al. 2006], et Meteor [Lavie & Denkowski 2009]. Depuis quelques années,le CNRC a développé sa propre métrique d'évaluation de la TA, appelée YiSi[1] [Lo 2019]. Cette métrique utilise différentes ressources sémantiques, notamment des *plongements lexicaux* (en anglais: *word embeddings*) issus de réseaux de neurones, pour estimer dans quelle mesure la traduction-machine et le texte d'origine véhiculent les mêmes informations. Grâce à cette approche novatrice, la métrique du CNRC s'est rapidement démarquée des métriques concurrentes: depuis quelques années, YiSi arrive systématiquement en tête de la *Shared task on Metrics* de la conférence annuelle WMT, le plus important événement pour le communauté de recherche en TA [Ma et al. 2018, Ma et al. 2019]

Le principal objectif de ce projet était donc de soutenir la mise en place au BtB de mécanismes automatiques de contrôle de la qualité des traductions automatiques, en créant une version de YiSi spécialement adaptée aux besoins du Bureau, et en mettant au point les mécanismes permettant son intégration dans ses flux de travail.

Méthodes d'évaluation de la traduction automatique

La traduction automatique est l'un des plus anciens problèmes d'application envisagé pour les ordinateurs, et la recherche dans ce domaine remonte aux années 1950. Les premiers travaux de recherche sur l'évaluation de la TA sont presque aussi anciens (pour un survol du domaine, voir [Han 2016]). En effet, dès lors que les chercheurs ont développé des méthodes d'automatisation de la traduction, ils ont voulu mesurer leur succès et leurs progrès. Or, dans la plupart des situations auxquelles on avait appliqué l'informatique jusqu'alors, la solution était unique: calcul du chemin le plus court, trajectoire balistique optimale, décryptage d'un message codé, etc. Dans de telles situation, l'évaluation était triviale: ou bien l'algorithme produisait la bonne réponse, ou alors il commettait une erreur, qu'on pouvait quantifier facilement. La traduction s'est avérée un cas à part: il n'y a pas de solution unique. En fait, on peut même argumenter qu'il y a une infinité de façons différentes de traduire un texte. De ce fait, il est difficile de détecter automatiquement les erreurs de traduction, et encore plus d'en quantifier l'importance.

---

[1] « YiSi » est la romanisation du mot cantonnais 意思, qui signifie « sens, signification ».

On retrouve donc toute une littérature sur les méthodes d'évaluation de la TA. C'est dans ces travaux qu'on retrouve notamment l'idée de mesurer séparément la qualité d'une traduction machine selon deux axes principaux, soit celui de l'adéquation (*adequacy*: dans quelle mesure la traduction transmet le même message que le texte original) et celui de la fluidité (*fluency*: dans quelle mesure la traduction constitue un texte grammatical bien formé dans la langue cible) [White et al. 1994]. Jusqu'au début des années 2000 toutefois, l'évaluation de la TA était envisagée comme une tâche manuelle: la question était de définir la nature exacte du travail des évaluateurs humains. La recherche dans cette importante veine continue d'ailleurs à ce jour, et a récemment donné lieu aux cadres d'évaluation des traductions DQF [Görög 2014] et MQM [Lommel et al. 2014].

Avec l'émergence d'approche de TA reposant sur des méthodes d'apprentissage-machine, toutefois, est apparu un besoin de mesurer la qualité des traductions de façon automatique. En effet, certaines méthodes d'apprentissage-machine reposent explicitement sur une boucle de minimisation des erreurs: le système comporte un ensemble parfois colossal de paramètres numériques, et l'algorithme d'apprentissage explore graduellement l'espace de ces paramètres, à la recherche d'une combinaison de valeurs qui minimise l'erreur totale sur les données d'apprentissage. Cette recherche de l'espace de solution ne peut reposer sur une évaluation manuelle, il faut donc développer des méthodes d'évaluation automatiques. La Figure 1 ci-dessous illustre ce genre de processus.

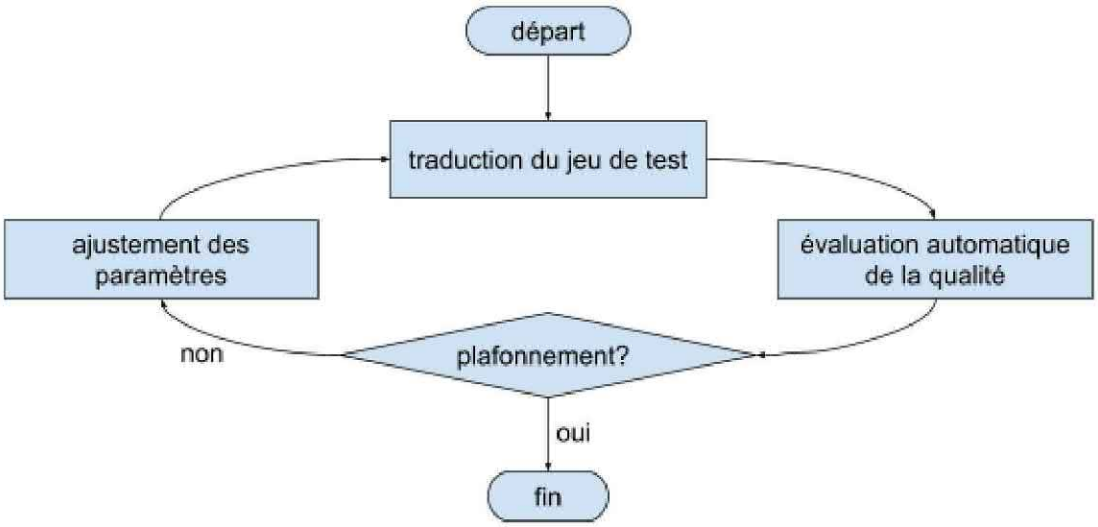


Figure 1: Processus d'optimisation d'un système de TA reposant sur l'évaluation automatique.

C'est de cette nécessité qu'est née la métrique BLEU [Papineni et al. 2002], de même que plusieurs métriques alternatives comme TER, hTER [Snover et al. 2006], Meteor [Lavie & Denkowski 2009], etc. Toutes ces métriques visent à évaluer globalement la qualité d'un système de TA. Pour ce faire, elles requièrent un échantillon de texte, représentatif du genre de texte pour lequel le système de TA est destiné, ainsi qu'une ou plusieurs traductions de bonne qualité de ce texte, typiquement des traductions produites par des traducteurs professionnels. On appelle *jeu de test* l'ensemble formé par les différentes versions de cet échantillon de texte.

La version source du jeu de test est soumise au système de TA dont on souhaite mesurer la qualité; les métriques d'évaluation procèdent alors en mesurant la similarité entre la traduction-machine et les traductions de référence. Plus cette similarité est grande, plus on jugera que la qualité de la traduction est élevée. C'est la manière de mesurer le degré de similarité qui différencie les métriques. BLEU, par exemple, n'est rien de plus qu'une mesure du chevauchement entre la traduction produite par la machine et la ou les traductions humaines de référence. Le chevauchement se mesure par le nombre de *n*-grammes partagés. Par *n*-gramme on entend une suite continue de *n* mots.

Par exemple, considérons la phrase anglaise « *the kitten laps the milk* » et sa traduction en français « *le petit chat lappe le lait* ». La traduction comprend les unigrammes, bigrammes, trigrammes et quadrigrammes suivants:

unigrammes	bigrammes	trigrammes	quadrigrammes
le (2 occurrences)	le petit	le petit chat	le petit chat lappe
petit	petit chat	petit chat lappe	petit chat lappe le
chat	chat lappe	chat lappe le	chat lappe le lait
lappe	lappe le	lappe le lait	
lait	le lait		

Si, pour la même phrase, un système de TA produit la phrase « *le chaton boit le lait* », on obtient les *n*-grammes suivants:

unigrammes	bigrammes	trigrammes	quadrigrammes
le (2 occurrences)	le chaton	le chaton boit	le chaton boit le
chaton	chaton boit	chaton boit le	chaton boit le lait
boit	boit le	boit le lait	
lait	le lait		

Pour obtenir le score BLEU, il suffit de calculer la proportion des *n*-grammes de la traduction machine qui apparaissent aussi dans une ou plusieurs traductions de référence. Par exemple, dans le tableau ci-dessus, ces *n*-grammes sont en rouge: on peut voir que, des 13 *n*-grammes uniques dans la traduction-machine, trois seulement se retrouvent également dans la traduction de référence, soit 23%.[2] Manifestement, BLEU est une approximation plutôt grossière de la similarité des traductions, et encore plus de la qualité des traductions-machine. En particulier, la métrique ignore complètement le fait que la traduction de référence ne constitue pas

[2] La formule complète de BLEU incorpore certains critères additionnels, notamment pour tenir compte des différences de longueur.

nécessairement la seule bonne traduction possible, et qu'un même concept peut être exprimé de plusieurs façons différentes dans la langue cible. Malgré tout, il a été observé expérimentalement que le classement effectué par BLEU sur la qualité relative des traductions du même texte par différents systèmes de TA est fortement corrélé avec le classement qui serait obtenu par un panel d'experts humains. C'est cette corrélation qui permet l'utilisation de BLEU dans un cadre d'optimisation d'un système de TA: si le système obtient un score BLEU plus élevé avec les paramètres A qu'avec les paramètres B, alors on peut considérer avec une certitude élevée que A est plus prometteur que B sur le plan de la qualité des traductions.

Avec l'amélioration croissante des systèmes de TA, plusieurs autre métriques ont vu le jour à la suite de BLEU, cherchant à raffiner la mesure de similarité, d'une part pour mieux tenir compte de la variabilité naturelle des langues naturelles, mais également pour permettre d'effectuer des jugements avec une résolution plus fine. La métrique YiSi du CNRC s'inscrit dans cette lignée [Lo 2019]. Comme on l'a vu, BLEU repose sur une notion de similarité pour le moins rudimentaire, de telle sorte qu'il jugera tout-à-fait différentes des traductions qui peuvent pourtant être équivalentes sur le plan sémantique. YiSi s'affranchit complètement de ces comparaisons de surface, en mesurant directement la similarité de sens entre les mots qui composent la traduction de référence et la traduction-machine. Cette comparaison s'appuie sur différentes ressources linguistiques, principalement des *plongements lexicaux*, mieux connus en anglais sous le nom de *word embeddings*.

Un plongements lexical est une représentation des mots de la langue dans un espace mathématique: il permet d'associer à chaque mot un vecteur multidimensionnel (typiquement quelques centaines de dimensions). Cet espace est structuré de façon telle que les mots dont le sens est apparenté sont représentés par des vecteurs qui «pointent» dans des directions similaires. Ainsi, on s'attend à ce que les représentations de «poule» et «cheval» soit moins distantes que celles de «poule» et «étable», mais plus distantes que celles de «poule» et «oie». Pour calculer la distance lexicale entre deux mots, on peut donc utiliser une généralisation multidimensionnelle du cosinus, cette fonction trigonométrique qu'on utilise pour comparer des angles: deux vecteurs qui pointent dans des directions similaires (parallèles) auront un cosinus près de 1, alors que des vecteurs orthogonaux auront un cosinus nul (voir la Figure 2).

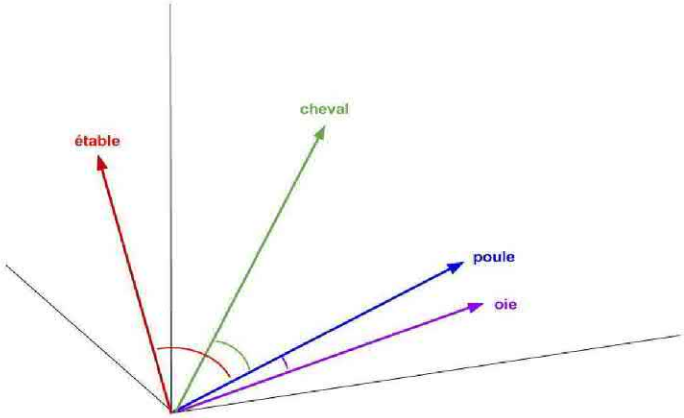


Figure 2: Représentation vectorielles de mots dans un plongement lexical.

Les plongements lexicaux utilisés par YiSi peuvent être obtenus de différentes façons, mais le plus souvent, on utilise une méthode d'apprentissage-machine à base de réseaux de neurones, telle que *word2vec* [Mikolov et al. 2013].

YiSi compare les traductions-machine produites pour chacune des phrases du jeu de test avec les traductions de référence correspondantes. À l'intérieur de chaque phrase, il combine les similarités lexicales pour calculer des similarités entre phrases. Le score produit par YiSi est une moyenne pondérée de ces similarités de phrases pour l'ensemble du jeu de test.

Comme pour toutes les métriques automatiques, les scores produits par YiSi ne doivent pas être interprétés de façon absolue: il n'existe pas d'échelle qui permette de dire, par exemple, qu'un score de 0.30 dénote un «mauvais» système ou qu'un score de 0.70 dénote un «bon» système. Tout dépend de la nature du jeu de test, de la paire de langue considérée, de la qualité des plongements lexicaux sur lesquels s'appuie le calcul, etc. En particulier, il est hasardeux de comparer entre eux des scores obtenus avec des jeux de test ou des plongements lexicaux (ou autres ressources linguistiques) différents.

Par contre, si les scores YiSi de deux systèmes A et B ont été obtenus dans les mêmes conditions (même jeux de test, mêmes ressources linguistiques), et si le score obtenu par le système A est plus élevé que celui obtenu par le système B, alors on peut raisonnablement penser que, pour des textes d'un nature similaire à celle du jeu test, le système A produira des traductions de meilleure qualité que le système B (et inversement si B obtient un meilleur score que A).

C'est donc dire que la métrique YiSi est bien adaptée pour comparer le potentiel de différents systèmes de TA pour traiter des textes d'une nature donnée. En particulier, les systèmes comparés peuvent être deux ou plusieurs instances d'un même système, avec des paramètres différents; en d'autres mots, lorsqu'un système de TA est soumis à un régime d'apprentissage continu, on peut utiliser YiSi pour suivre l'évolution du système au fil de son apprentissage.

Les taux de corrélation élevés de BLEU avec des évaluateurs humains ne s'observent qu'avec des échantillons de taille substantielle: typiquement on recommande d'utiliser des jeux de test de quelques milliers de phrases. Sur des phrases individuelles, BLEU est beaucoup moins fiable. Un des principaux avantages de YiSi, par rapport à des métriques comme BLEU, est sa capacité à mesurer des différences sur de petits échantillons : lors des récentes *shared tasks* sur l'évaluation de la TA, YiSi s'est surtout démarquée des autres métriques par sa fiabilité sur des phrases individuelles [Ma et al. 2019]. Il en résulte qu'il est possible d'utiliser des jeux de test substantiellement plus petits avec YiSi, de détecter des différences beaucoup plus fines entre les systèmes, et de s'en servir pour effectuer des analyses détaillées des forces et des faiblesses des différents systèmes sous la loupe.

Systèmes YiSi pour le BtB

Deux systèmes d'évaluation automatique de la TA ont été produits spécifiquement pour le BtB. Ces systèmes, un pour l'anglais et un pour le français, reposent sur la technologie YiSi, et s'appuient sur des plongements lexicaux *word2vec* [Mikolov et al. 2013]. Ces plongements ont été créés exclusivement à partir de données du Bureau: pour chacune des deux langues, on a utilisé un échantillon du Mégacorpus d'environ 50000 documents. Les détails de ces corpus apparaissent dans le tableau ci-dessous.

Tableau 1: Données d'entraînement des systèmes YiSi du BtB

Langue	Documents	Phrases	Mots
anglais	49842	3,60 M	59,6 M
français	50158	3,55 M	67,1 M

Les systèmes YiSi obtenus permettent une évaluation relativement précise de la qualité des traductions automatiques. À titre d'exemple, nous avons utilisé ces systèmes pour comparer la performance de deux systèmes de TA spécialisés pour la traduction des débats de la Chambre des communes du parlement canadien: un système statistique (Portage) et un système neuronal (Sockeye). Les deux systèmes ont été créés en utilisant les mêmes données d'entraînement (plusieurs millions de paires de phrases issues des Hansard), et sont évalués sur un extrait d'environ 3000 paires de phrases, distinct du corpus d'entraînement.

Le Tableau 2 rapporte les scores YiSi obtenus par les deux systèmes pour les traductions vers le français et vers l'anglais. Nous rapportons également les scores BLEU et WER (Word Error Rate, une variante de la métrique TER [Snover et al. 2005]), de même que le résultat d'une évaluation humaine des deux systèmes: 4 évaluateurs ont comparé à l'aveugle un sous-ensemble du jeu de test (400 paires de traductions), et déterminé pour chaque paire laquelle des deux était la meilleure. Les résultats de cette évaluation humaine sont rapportés en terme de «taux de préférence».

Tableau 2: Évaluation comparative de systèmes statistique (Portage) et neuronal (Sockeye).

	EN-FR		FR-EN	
	Portage	Sockeye	Portage	Sockeye
BLEU	37.9	**40.5**	**39.8**	38.7
1-WER	47.0	**48.5**	**50.2**	48.5
YiSi	70.2	**70.4**	**74.0**	71.6
Évaluation manuelle	45.0%	**55.0%**	**53.1%**	46.9%

Comme on peut le voir, dans cette expérience, les métriques sont unanimes: pour les traductions vers le français, le système neuronal produit de meilleures traductions que le système statistique, alors que pour les traductions vers l'anglais, c'est l'inverse.

Comme pour l'évaluation manuelle, pour chacune des métriques utilisées, nous pouvons calculer des taux de préférence par phrase, c'est-à-dire le pourcentage de phrases du jeu de test pour lesquelles une certaine métrique «préfère» la traduction produite par un certain système. Nous pouvons également comparer ces préférences avec celles des évaluateurs humains. Ce sont ces résultats que nous rapportons dans le Tableau 3. Dans les colonnes «Accord» de ce tableau, nous rapportons le pourcentage des phrases du test pour lesquelles la préférence de la métrique donnée est en accord avec la préférence de l'évaluateur humain. Par exemple, pour les systèmes anglais-français, la métrique BLEU affiche un accord de 55% : cela signifie que, pour 55% des phrases du jeu de test, BLEU donne un score plus élevé à la phrase préférée par les évaluateurs humains.

Tableau 3: Taux de préférence des différentes métriques quant aux traductions produites par des systèmes statistique (Portage) et neuronal (Sockeye).

	EN-FR			FR-EN		
	Portage	Sockeye	Accord	Portage	Sockeye	Accord
BLEU	39.2%	**50.5**%	55.0%	**52.0**%	43.3%	63.3%
1-WER	35.0%	**43.5**%	49.2%	**41.7**%	40.0%	54.7%
YiSi	42.0%	**53.5**%	**61.2%**	**55.7**%	44.3%	**67.3%**
Évaluation manuelle	45.0%	**55.0%**	**-**	**53.1%**	46.9%	-

Comme on peut le constater, YiSi affiche des taux d'accord avec les évaluateurs humains nettement supérieurs aux autres métriques. C'est ce qui en fait un outil particulièrement approprié pour analyser et comparer en détail les sorties de différents systèmes.

Protocole d'évaluation de la qualité de la TA

Pour que l'évaluation d'un ou de plusieurs systèmes au moyen d'une métrique d'évaluation telle que YiSi soit valide, elle doit respecter certaines conditions, notamment en ce qui a trait à la création du jeu de test. Nous examinons ici en détail les différentes étapes à suivre pour assurer la validité des résultats.

Identification du domaine

Les systèmes de TA à l'intention du grand public (Google Translate, DeepL, etc.) sont habituellement des systèmes « généralistes », c'est-à-dire qu'ils sont conçus pour traiter des textes de natures diverses. À l'opposé, les systèmes déployés en industrie sont souvent des systèmes « spécialisés », c'est-à-dire conçus de façon à produire des résultats optimaux sur des textes d'une nature spécifique, par exemple des descriptions de postes gouvernementaux, des contrats de service informatiques, des avis juridiques, etc. C'est ce qu'on appelle le *domaine* d'application d'un système de TA. Lorsqu'on utilise un système spécialisé pour traduire un texte du domaine, on s'attend généralement à ce qu'il produise des traductions de meilleure qualité que si on avait utilisé un système généraliste. En revanche, si on utilise le même système pour des textes d'une nature différente de ceux pour lequel il a été créé (des textes «hors-domaine»), ce système se comportera souvent moins bien que les systèmes généralistes.

Avant d'entamer l'évaluation d'un ou de plusieurs systèmes de TA, il faut se poser la question du domaine d'application envisagé pour ceux-ci. Si on prévoit utiliser un système pour traduire des rapports de marchés agricoles, ça n'est pas pertinent d'en tester la qualité sur des bulletins météo, et inversement.

Il est important aussi de bien cerner l'application visée. Si le système doit être utilisé par des traducteurs professionnels pour produire une première ébauche, on jugera peut-être qu'il est primordial que le système produise la terminologie juste pour le domaine. Si le système est destiné à un public large, comme outil de compréhension linguistique, on voudra peut-être plutôt prioriser une syntaxe fluide et la transmission de l'idée générale.

Identification du corpus

Une fois le domaine identifié, on voudra constituer ou localiser un corpus de texte à partir duquel seront extraits les échantillons de test. Ce corpus aura les caractéristiques suivantes:

- Le corpus est **bilingue**, constitué de textes rédigés en langue source et de traductions en langue-cible (la direction de traduction est importante, mais pas critique).

- Les traductions, qu'on appelle **traductions de référence** sont de bonne qualité: on veut des traductions produites par des professionnels, d'une qualité égale ou supérieure à ce que l'on souhaite pour l'application visée. Accessoirement, on peut disposer de plusieurs traductions de référence pour le même texte.

- Le corpus correspond au **domaine** identifié. Si le domaine est vaste et comporte des sous-domaines, on souhaite que le corpus soit représentatif de l'ensemble des sous-domaines, tant en nature qu'en quantité (distribution relative des sous-domaines dans le domaine).

Il peut arriver qu'un corpus répondant à ces trois critères n'existe pas, ou ne soit pas disponible. Par exemple, il est possible que les textes du domaine n'aient jamais fait l'objet de traduction par le passé. Dans ce cas, trois avenues sont à considérer:

Corpus apparenté: Une approche consiste à utiliser un corpus bilingue dont le domaine est apparenté au domaine visé. Il faudra néanmoins prendre en considération la distance entre les deux domaines dans l'interprétation des résultats. En particulier, si les deux domaines se distinguent par leurs terminologies ou par le niveau de langue, les conclusions de l'évaluation seront plus ou moins valables.
Un cas plus fréquent est celui où l'on dispose d'un corpus *englobant*, qui couvre un domaine plus large que le domaine visé, c'est-à-dire qui inclut des textes du domaine visé, entremêlés avec des textes d'autres domaines. Dans ce cas, si on dispose de textes en langue source ou en langue cible du domaine visé, on peut utiliser des méthodes de *sélection de données*, qui permettent d'extraire du corpus englobant les données qui s'apparentent le plus au domaine visé. Voir par exemple [Moore & Lewis 2010].

Création du corpus: En l'absence de corpus englobant il est possible qu'il soit nécessaire de créer un corpus. Il faudra alors se procurer des textes en langue source, et en confier la traduction à un ou des traducteurs. On voudra alors maximiser l'impact de l'opération et minimiser les coûts de production en ciblant des phrases qui représentent au mieux le domaine. (La question de la taille du corpus est abordée dans la section suivante.) D'une part, on doit prioriser des phrases qui présentent un vocabulaire et des tournures fréquemment observées. D'autre part, on veut aussi favoriser la diversité, en évitant des phrases trop similaires entre elles. Comme pour l'extraction d'un corpus apparenté, il existe des méthodes automatiques de sélection de données pour aider à l'identification des phrases.

Évaluation manuelle: Dans certains cas, plutôt que d'utiliser des métriques automatiques comme YiSi, il peut être préférable d'opter pour une évaluation manuelle des systèmes. Ça peut être le cas si l'évaluation ne concerne que deux ou trois systèmes de TA, si elle n'est pas appelée à être répétée périodiquement, ou si on prévoit qu'un corpus adéquat sera disponible à moyen terme, et donc que ça ne vaut pas la peine d'en créer un maintenant. Une approche simple et efficace consiste à comparer des traductions de phrases «à l'aveugle» sur un échantillon de phrases. Comme pour la création de corpus, on voudra maximiser la représentativité du jeu de test. Pour ce genre d'évaluation, les phrases très longues ou très courtes sont souvent difficiles à évaluer. C'est pourquoi on voudra limiter la taille des phrases, par exemple cibler des phrases contenant entre 5 et 20 mots.

Si l'évaluation concerne plusieurs systèmes de TA, il peut être utile d'envisager une méthodologie d'évaluation manuelle plus complexe, tel que la méthode *Appraise* [Federmann 2012].

Extraction du jeu de test

Une fois le domaine et le corpus identifiés, on peut procéder à l'extraction du jeu de test. Ce sont les phrases en langue source de ce jeu de données qui seront soumises à la TA, dont les traductions seront alors comparées aux traductions de référence. Dans une dynamique de

création de système de TA, le corpus peut contenir des dizaines ou même des centaines de millions de mots. De cette quantité, on extraiera un petit sous-ensemble, qui constituera le jeu de test. Cette extraction se fait avant l'entraînement du système de TA, et les données du jeu de test ne sont pas utilisées pour l'entraînement. Cette façon de faire permet de garantir que les données de test sont distinctes des données d'entraînement. (Nous discutons plus en détail de cette question ultérieurement, dans la section sur le Prétraitement du jeu de test)

Toutefois, si les systèmes de TA sont pré-existants et si le corpus ne sert qu'à l'évaluation, alors le corpus peut *être* le jeu de test. Dans tous les cas, le jeu de test doit être d'une taille suffisante pour obtenir des mesures fiables des métriques d'évaluation. En général, l'échantillon contiendra environ 1000 paires de phrases, soit entre 10000 et 20000 mots de texte dans chacune des deux langues.[3] Notons qu'il s'agit là d'une indication générale. Si le domaine visé est très spécialisé, avec un vocabulaire restreint, il est possible que 300 ou 500 phrases soient suffisantes. À l'opposé, pour un domaine plus large ou des systèmes généralistes, il pourrait être préférable de viser 1500 ou 2000 phrases.

L'extraction du jeu de test peut se faire de différentes façons. L'intention, c'est que l'échantillon constitue un exemple aussi représentatif que possible de la tâche réelle pour laquelle le système de TA est destiné. Étant donné un corpus de traduction relativement volumineux, la meilleure façon d'arriver à ce résultat est d'effectuer une sélection aléatoire de segments du corpus.

Cette sélection peut toutefois être assujettie à différentes contraintes. Par exemple:

- **Terminologie/phraséologie.** Même à l'intérieur d'un domaine spécifique, il peut exister des variantes qui dépendent, par exemple, du client de la traduction, et qui se manifestent par des différences sur le plan terminologique, phraséologique ou même typographique. Par exemple, des documents «bancaires» qui proviennent de deux banques différentes comporteront peut-être des différences subtiles sur le plan terminologique. Il faut s'assurer que l'échantillon de test correspond bien au profil des textes qu'on a l'intention de traduire avec le système.

- **Chronologie**: Lorsqu'on a affaire à un domaine qui évolue rapidement, comme la plupart des domaines scientifiques, on voudra le plus souvent privilégier pour le test des documents récents, qui sont plus susceptible d'être représentatifs des textes qu'on recevra dans l'avenir. Par contre, si la nature des textes du domaine varie de façon cyclique dans le temps (par exemple: des bulletins météorologiques), on voudra s'assurer que le jeu de test représente bien l'ensemble du cycle.

- **Hétérogénéité**: Certains domaines, notamment ceux qui sont liés à l'actualité, sont naturellement hétérogènes: le sujet qui était d'actualité le mois dernier ne l'est peut-être plus aujourd'hui, et celui qui intéressera demain ne ressemble peut-être pas beaucoup à ce qu'on a vu dans le passé. Dans ce cas, il peut être approprié de choisir un échantillon

[3] Tel que discuté plus haut, YiSi nécessite théoriquement moins de données que BLEU. Nous recommandons toutefois que le jeu de test soit d'une taille suffisante pour assurer une bonne représentation du domaine des textes. De plus, une plus grande quantité permet de détecter des différences plus fines entre les systèmes.

de test qui, tout en respectant le genre et le style du domaine, se démarque des textes d'entraînement par son sujet ou son vocabulaire spécialisé.

- **Contextualisation**: Si on a l'intention de soumettre le jeu de test à une évaluation manuelle, en tout ou en partie, on voudra possiblement constituer celui-ci à partir de documents complets ou de paragraphes ou de blocs de lignes contiguës, ce qui permettra aux évaluateurs de considérer les traductions en contexte[4].

Prétraitement du jeu de test

Déduplication. Dans une dynamique de création de systèmes de TA, il est primordial que le jeu de test n'ait pas servi à l'entraînement ou à la création du système de TA qu'on veut évaluer. En effet, la plupart des systèmes de TA reposent maintenant sur des méthodes d'apprentissage-machine. Ces systèmes ont tendance à mémoriser les textes rencontrés lors de leur phase d'apprentissage, et à en recracher la traduction telle quelle. En apprentissage-machine, la principale qualité recherchée est la capacité de *généralisation* d'un système, c'est-à-dire son aptitude à traiter correctement des situations jamais observées. La capacité d'un système de TA à traduire «par coeur» n'est généralement pas un bon indicateur de la qualité du système. C'est pourquoi on procède habituellement à une *déduplication* des données de test: on élimine du jeu de test toutes les phrases qui figurent également dans les données d'entraînement. En général, on éliminera aussi les répétitions internes, c'est-à-dire les duplicatas de phrases au sein même du jeu de test.[5]

Toutefois, si le domaine est naturellement répétitif, et si l'application visée ne prévoit pas l'utilisation conjointe d'une mémoire de traduction, il peut être pertinent de laisser les duplicatas dans le jeu de test. Encore une fois, tout est question d'obtenir un jeu de test qui soit aussi représentatif que possible des véritables conditions d'utilisation du système de TA.

Segmentation et Alignement. Les métriques d'évaluation automatique requièrent que les textes soit segmentés en phrases et alignées. Si le jeu de test est extrait d'une mémoire de traduction, les données sont directement disponibles sous cette forme. Si on utilise des documents en format original (documents Word, HTML, etc.), il faudra effectuer ces opérations. Un outil tel que *AlignFactory* (ou sa version en-ligne, *YouAlign*[6]) produit généralement de bons résultats. Toutefois, il est important de contrôler la qualité de la segmentation et de l'alignement. Si le jeu de test comporte des phrases mal segmentées, la qualité des traductions automatiques peut s'en trouver affectée. Si certains segments sont mal alignés, la qualité des traductions de ces phrases sera systématiquement sous-estimée. Il faut donc: soit corriger la segmentation et l'alignement des phrases affectées, soit les éliminer du jeu de test. Si on choisit cette dernière

---

[4] Dans ce cas, une approche alternative consiste à préserver séparément le contexte des phrases du jeu de test, de façon à permettre aux évaluateur de consulter celui-ci au besoin.

[5] En fait, lorsqu'on crée un système de TA, la procédure habituelle est de dédupliquer l'ensemble du corpus, puis de le partitionner en deux sous-ensembles disjoints, *test* et *entraînement*.
[6] https://youalign.com

option, il faut quand même s'assurer qu'on respecte toujours les contraintes de taille du jeu de test et de représentativité des sous-domaines (le cas échéant).[7]

Format et encodage. Il existe différentes façons de gérer les jeux de test. Une méthode relativement simple est d'utiliser un tableur comme Excel, en organisant les différentes versions d'un texte (source, traduction de référence, traductions machine) dans des colonnes distinctes, un segment par ligne. Cette façon de faire permet d'une part de préserver la segmentation du texte en phrase, ce qui garantit que tous les systèmes de traduction testés se basent sur le même segmentation, et d'autre part de préserver naturellement l'alignement entre les phrases source et toutes leurs traductions. Il est relativement aisé d'importer la sortie de AlignFactory dans ce format, et la plupart des systèmes de TA devraient être en mesure de traiter directement des fichiers Excel, ce qui facilitera la réinsertion des traductions-machine dans le fichier.

Un des aspects qui doit être traité avec soin est l'encodage des caractères. Lorsque des textes provenant de différentes sources sont combinés dans un même fichier, il n'est pas rare de voir apparaître des incohérences au niveau de l'encodage des caractères accentués. Ces incohérences doivent être résolues avant de soumettre les textes à l'évaluation. Par ailleurs, le système YiSi suppose en encodage des textes dans le format UTF-8. Il faudra donc s'assurer d'exporter le contenu du jeu de test dans ce format.

Calcul des scores et interprétation des résultats

L'échantillon de test peut maintenant être soumis au(x) système(s) de TA visé(s). On voudra s'assurer que le processus de traduction préserve la segmentation originale du texte. C'est ce qui suggère l'utilisation d'un format comme TMX ou XLIFF, ou l'utilisation d'un tableur comme Excel.

Les traductions machine peuvent alors être évaluées. YiSi prend normalement en entrée le texte source, la traduction de référence et la ou les traductions machine. Le système produit un score global pour chacune des systèmes. Il peut fournir des scores individuels pour les traductions de chaque phrase du jeu de test. Comme YiSi est une mesure de similarité (par opposition à une mesure des erreurs), les valeurs plus élevées dénotent des traductions de meilleures qualité.

Accessoirement, lorsqu'on compare deux systèmes de TA ou plus, YiSi peut calculer un *taux de préférence* pour chaque système. Ce taux, qui s'exprime comme un pourcentage pour chaque système, représente la proportion des phrases du jeu de test pour laquelle ce système obtient le score YiSi le plus élevé.

---

[7] Si le but de l'évaluation est de comparer plusieurs systèmes entre eux, les erreurs d'alignement ou de segmentation auront un impact limité, dans la mesure où elles ne sont pas trop nombreuses. En effet, comme tous les systèmes sont soumis au même test, ils rencontreront tous les mêmes problèmes, et on peut penser que leurs performances respectives seront affectées de façon à peu près égale.

ÉQuaTion: une trousse d'évaluation de la qualité des traductions automatiques

API

Les fonctionnalités des systèmes YiSi créés pour le Bureau sont rendues accessibles par le biais d'un serveur HTTP. Ce serveur, le logiciel YiSi et ses ressources linguistiques (plongements *word2vec*) sont livrés dans un conteneur Docker unique, qui peut être installé sur n'importe quelle plateforme compatible.

Les fonctionnalités sont accessibles via l'API REST décrite en annexe.

Interface utilisateur

Le serveur HTTP permet également d'accéder à un interface utilisateur (IU) simple pour YiSi. Nous décrivons ci-dessous les principales étapes et opérations de l'utilisation de celle-ci.

Soumettre un test

L'IU requiert que le jeu de test soit sous forme de fichiers en texte brut ("*plain text*"), une phrase par ligne. Tous les fichiers doivent comporter le même nombre de lignes : la $n^{\text{ième}}$ ligne de chaque fichier traduction doit contenir la traduction de la $n^{\text{ième}}$ ligne du fichier en langue source. La page d'accueil de l'interface (Figure 3) comporte des champs permettant de spécifier les différents fichiers constituant le jeu de test: texte en langue-source, traduction de référence (une seule traduction de référence est permise), et une ou deux traductions-machine.

Pour soumettre le test, il faut également spécifier un «analyseur», qui active des ressources différentes en fonction de la langue cible du test. Par exemple, dans la Figure 3, puisque le test concerne des traductions de l'anglais vers le français, on spécifie l'analyseur «YiSi anglais-français».

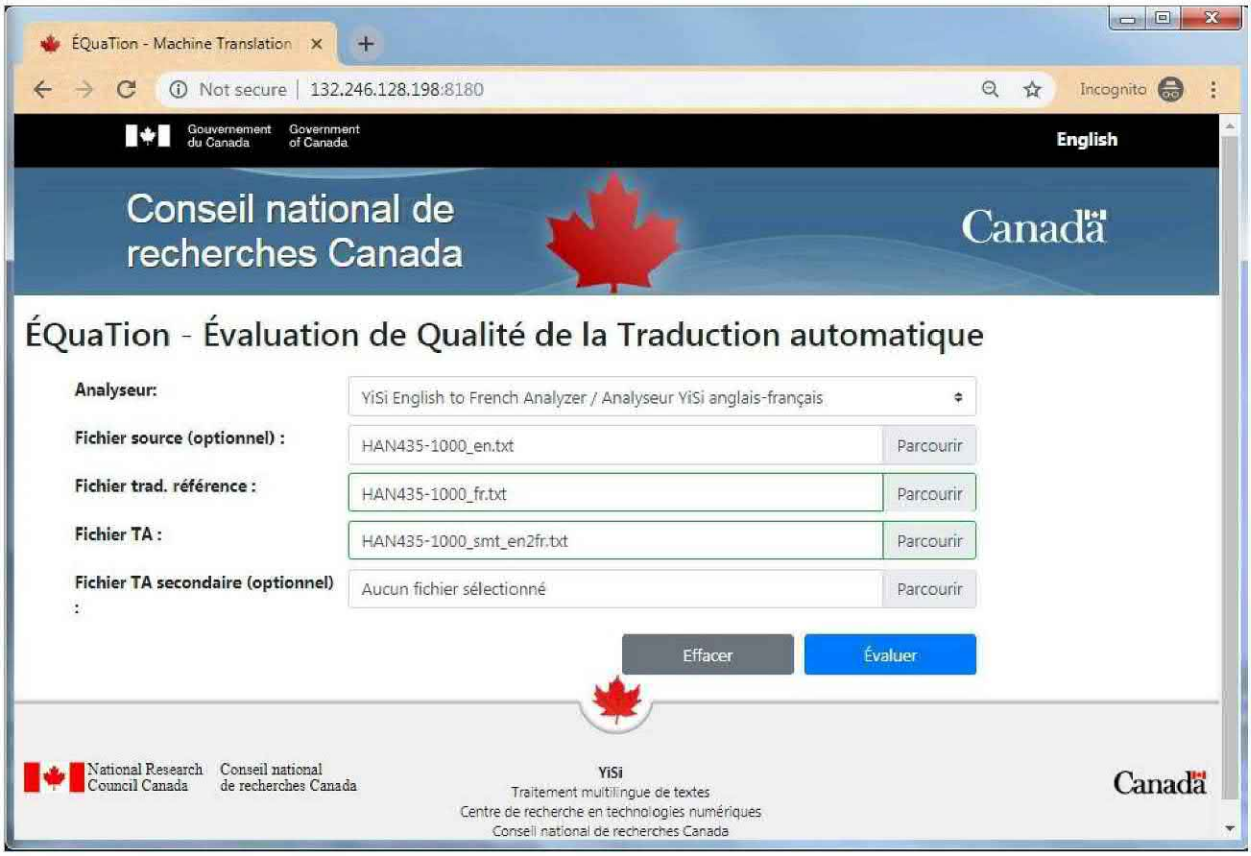


Figure 3: Page d'accueil de l'interface utilisateur

Interpréter le sommaire des résultats

En appuyant sur le bouton «Évaluer», le système lance le calcul des scores. Dépendant de la taille du jeu de test (nombre de phrases), ce calcul peut prendre de quelques secondes à quelques minutes. Dans tous les cas, il faut éviter de cliquer à répétition sur le bouton «Évaluer».

Une fois le calcul terminé, l'IU affiche un sommaire des résultats (Figure 4). Celui-ci comporte le nombre de segments du test, le nombre de mots dans chaque version, et le score YiSi moyen: en effet, YiSi calcule un score pour chaque segment du test, et le score global attribué au système est obtenu en calculant la moyenne des scores individuels pour chaque phrase du test, pondéré par la taille de chaque phrase (nombre de mots).

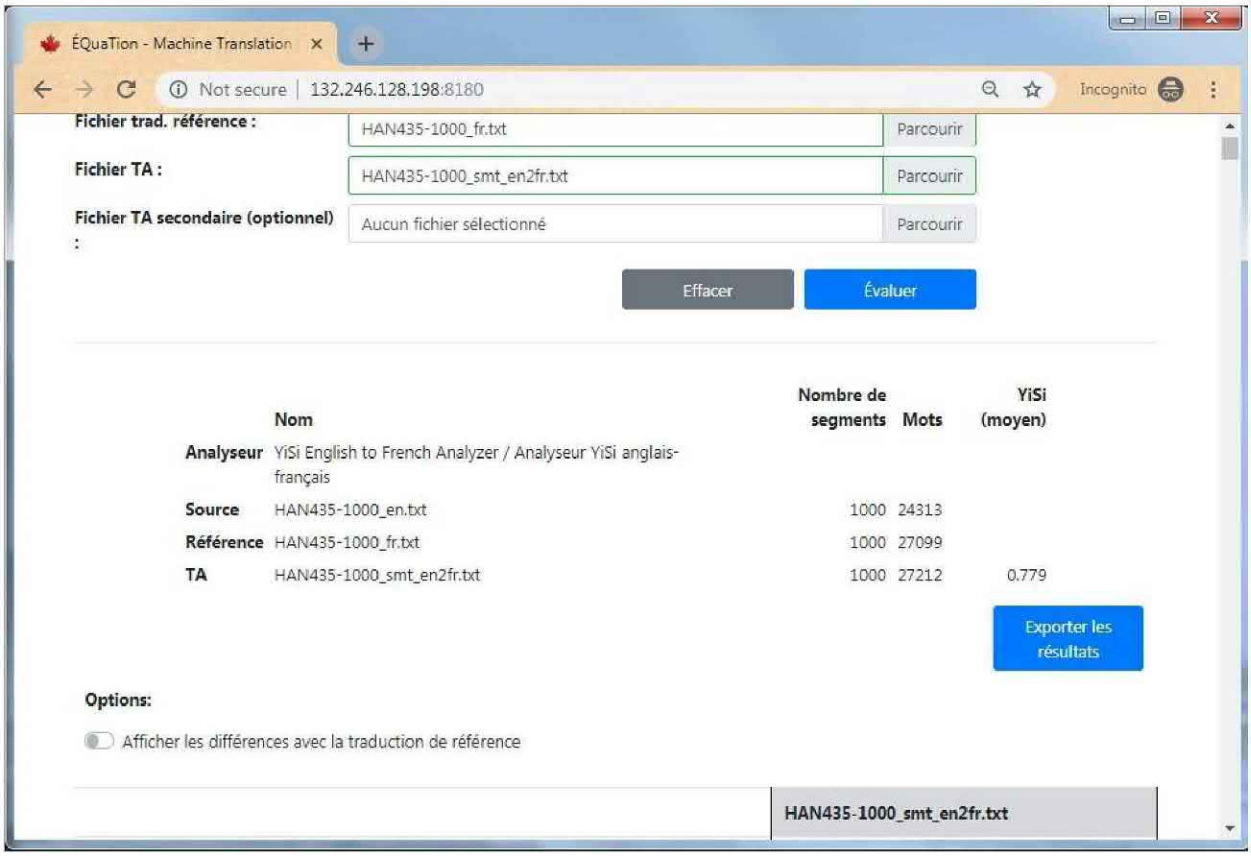
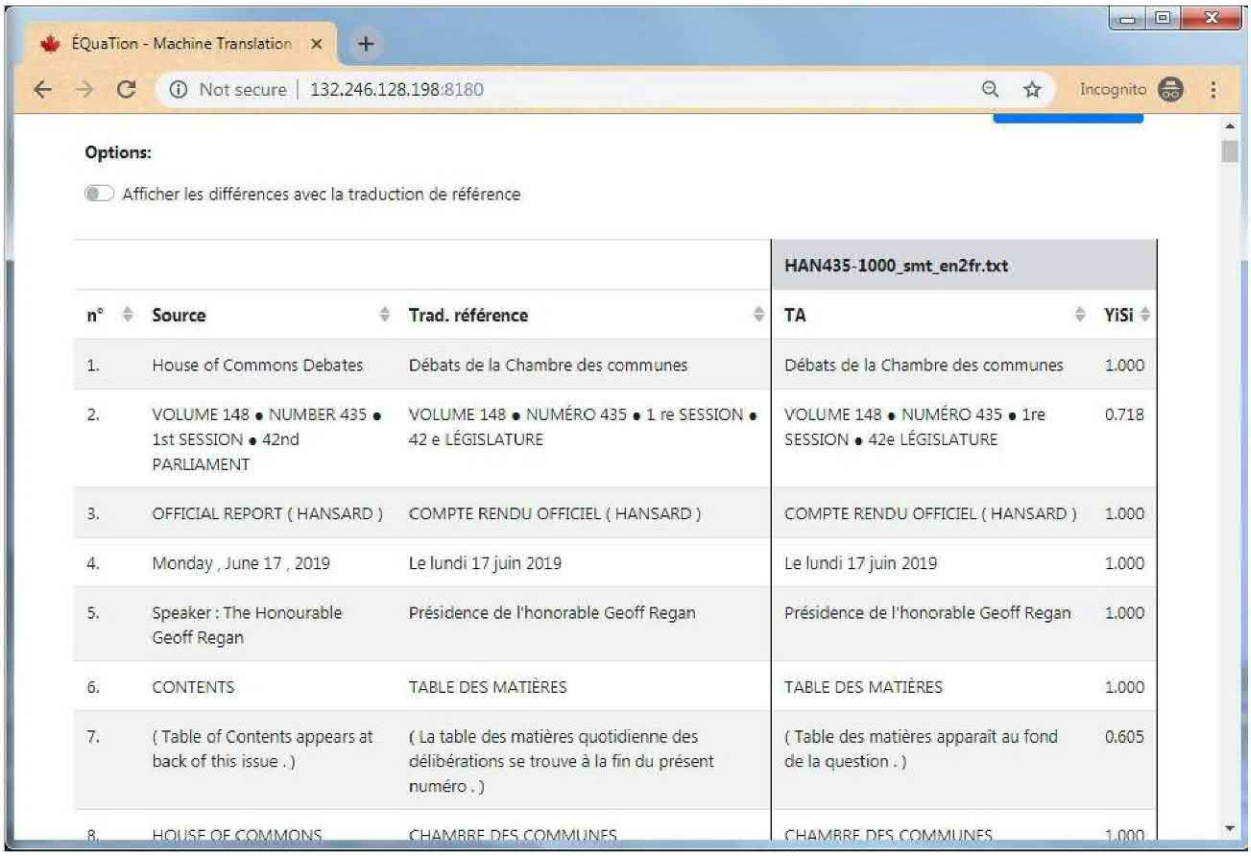


Figure 4: Sommaire des résultats

Analyser les résultats phrase par phrase

Pour mieux comprendre le comportement d'un système de traduction, il est parfois utile d'examiner les traductions qu'il produit phrase par phrase. À la suite du sommaire, l'IU affiche sous forme de tableau le score obtenu pour chaque phrase (Figure 5).



			HAN435-1000_smt_en2fr.txt	
n°	Source	Trad. référence	TA	YiSi
1.	House of Commons Debates	Débats de la Chambre des communes	Débats de la Chambre des communes	1.000
2.	VOLUME 148 • NUMBER 435 • 1st SESSION • 42nd PARLIAMENT	VOLUME 148 • NUMÉRO 435 • 1 re SESSION • 42 e LÉGISLATURE	VOLUME 148 • NUMÉRO 435 • 1re SESSION • 42e LÉGISLATURE	0.718
3.	OFFICIAL REPORT (HANSARD)	COMPTE RENDU OFFICIEL (HANSARD)	COMPTE RENDU OFFICIEL (HANSARD)	1.000
4.	Monday , June 17 , 2019	Le lundi 17 juin 2019	Le lundi 17 juin 2019	1.000
5.	Speaker : The Honourable Geoff Regan	Présidence de l'honorable Geoff Regan	Présidence de l'honorable Geoff Regan	1.000
6.	CONTENTS	TABLE DES MATIÈRES	TABLE DES MATIÈRES	1.000
7.	(Table of Contents appears at back of this issue .)	(La table des matières quotidienne des délibérations se trouve à la fin du présent numéro .)	(Table des matières apparaît au fond de la question .)	0.605
8.	HOUSE OF COMMONS	CHAMBRE DES COMMUNES	CHAMBRE DES COMMUNES	1.000

Figure 5: Résultats détaillés

Dans ce tableau, il est possible d'afficher en surbrillance les différences entre la traduction de référence et la traduction automatique, en sélectionnant l'option *«Afficher les différences avec la traduction de référence»* (Figure 6). Les mots de la référence omis par le système de TA apparaissent ~~rayés et en rouge~~, alors que les mots ajoutés par la TA sont soulignés et en jaune. Il est également possible de trier les phrases suivant différents critères, en cliquant sur les flèches à côté de la rubrique appropriée. Par exemple, pour voir d'abord les phrases pour lesquelles le système de TA a obtenu les plus mauvais scores, on peut trier le tableau en ordre croissant de score, comme dans la Figure 6.

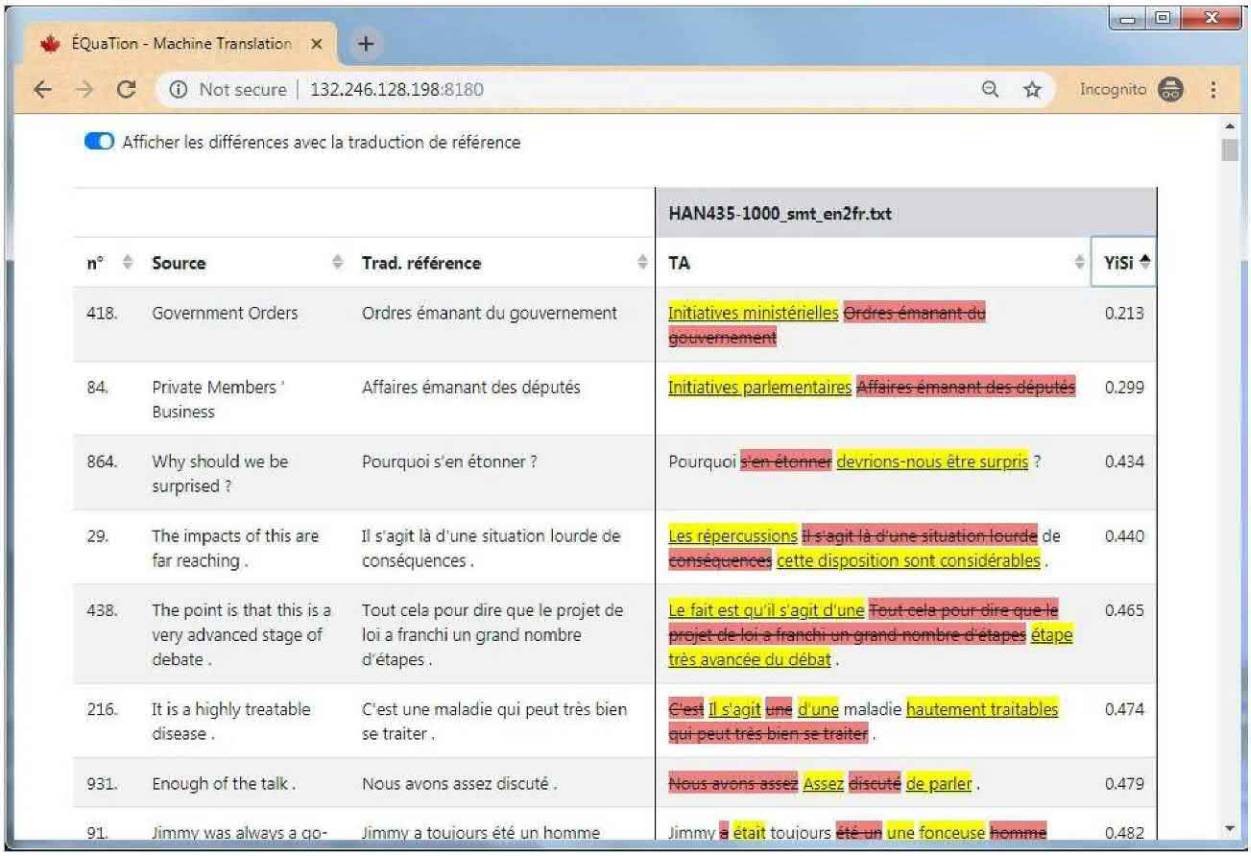


Figure 6: Affichage des différences et tri des résultats en fonction du score

Comparer deux systèmes

Il est possible d'utiliser l'IU pour comparer les traductions produites par deux systèmes de TA différents. Pour ce faire, en plus de spécifier les fichiers source, référence et TA, il faut spécifier un fichier de TA secondaire, comme dans la Figure 7. L'IU affiche alors non seulement des statistiques individuelles de chaque système, mais aussi un taux de «préférence». Il s'agit, pour chaque système, du pourcentage de phrases pour lequel ce système a obtenu le meilleur score. Par exemple, si on compare deux systèmes, A et B, sur un jeu de test de 1000 phrases, que le système A obtient un score plus élevé que B sur 550 de ces phrases, que le système B obtient un score plus élevé que A sur 300 phrases, et que les deux systèmes obtiennent des scores identiques sur les 150 phrases restantes, alors les taux de préférence seront de 55% pour A et 30% pour B.

En pratique, cette statistique est souvent plus éloquente et facile à interpréter que la comparaison directe entre les scores globaux (YiSi moyen). Dans l'exemple de la Figure 7, les deux systèmes obtiennent des scores YiSi moyens presque identiques (0.805 et 0.791) mais des taux de préférence substantiellement différents (58.46% et 37.68%). Une évaluation manuelle des deux systèmes révèle que ces taux de préférence reflètent de façon réaliste les préférences des évaluateurs.

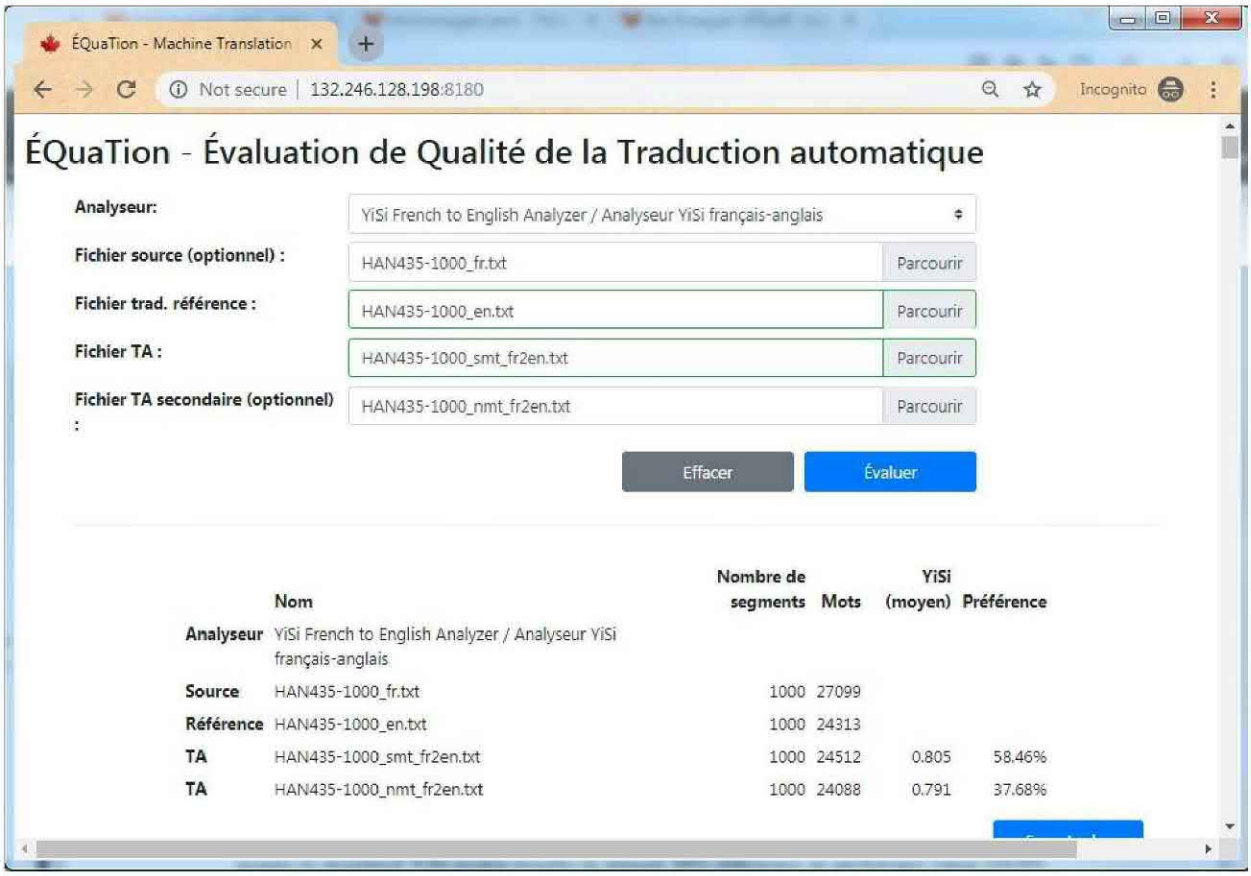
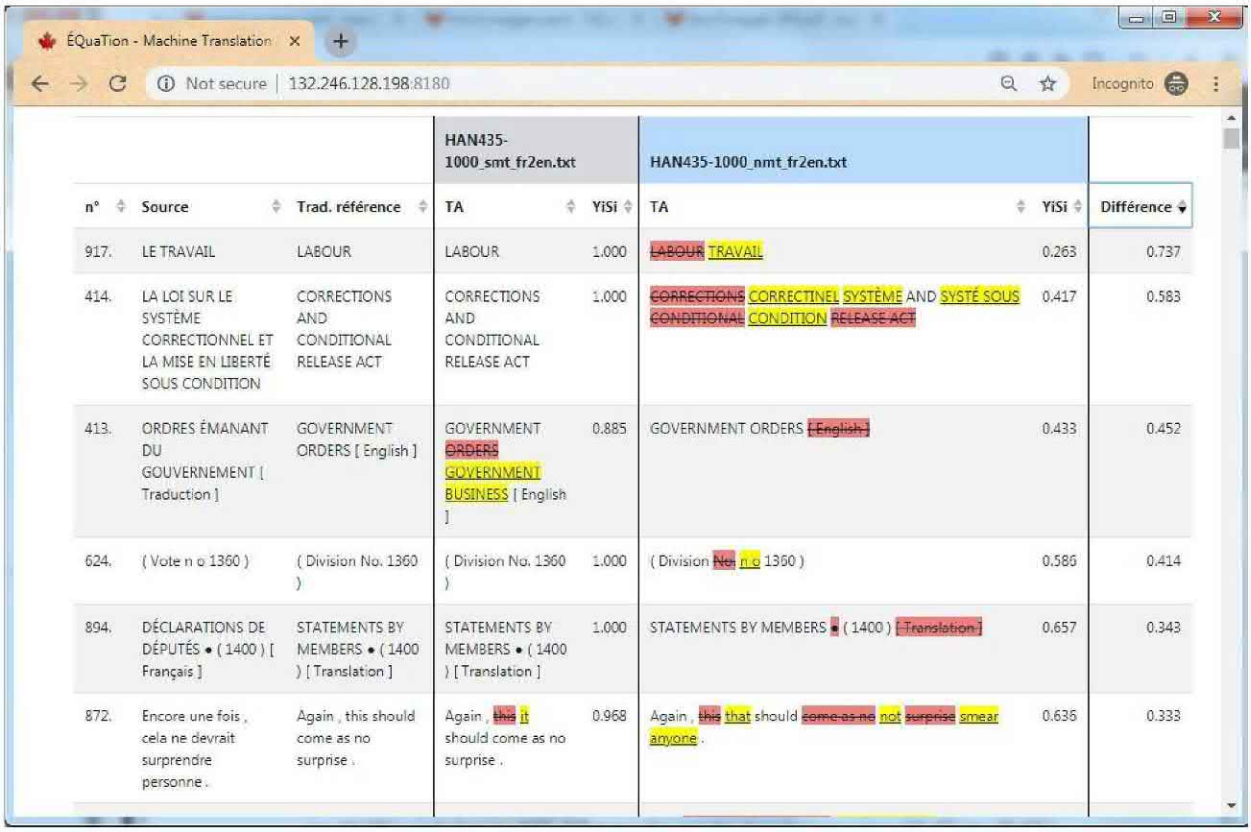


Figure 7: Comparaison de deux systèmes de TA

Comme pour l'évaluation d'un système unique, dans le cas d'une comparaison entre deux systèmes, l'IU affiche aussi des résultats détaillés (Figure 8). Dans ce cas, toutefois, il est également possible de restreindre l'affichage aux seules phrases où les deux systèmes obtiennent des scores différents, en sélectionnant l'option «*Afficher uniquement les segments dont les scores diffèrent*», ce qui élimine notamment les phrases pour lesquelles les deux systèmes produisent des traductions identiques.

De plus, l'IU affiche une colonne supplémentaire, intitulée «Différence»: il s'agit de la différence entre les scores obtenus par les deux systèmes. La présence de cette colonne permet de trier les résultats de façon à afficher en priorité les phrases pour lesquelles un des deux systèmes se comporte mieux que l'autre. Par exemple, dans la Figure 8, en affichant les phrases en ordre décroissant des différences, l'IU affiche d'abord les phrases pour lesquelles le premier système obtient des scores élevés alors que le second obtient des scores faibles.



ÉQuaTion - Machine Translation

Not secure | 132.246.128.198:8180

			HAN435-1000_smt_fr2en.txt		HAN435-1000_nmt_fr2en.txt		
n°	Source	Trad. référence	TA	YiSi	TA	YiSi	Différence
917.	LE TRAVAIL	LABOUR	LABOUR	1.000	~~LABOUR~~ TRAVAIL	0.263	0.737
414.	LA LOI SUR LE SYSTÈME CORRECTIONNEL ET LA MISE EN LIBERTÉ SOUS CONDITION	CORRECTIONS AND CONDITIONAL RELEASE ACT	CORRECTIONS AND CONDITIONAL RELEASE ACT	1.000	~~CORRECTIONS~~ CORRECTINEL SYSTÈME AND SYSTÉ SOUS ~~CONDITIONAL~~ CONDITION ~~RELEASE ACT~~	0.417	0.583
413.	ORDRES ÉMANANT DU GOUVERNEMENT [Traduction]	GOVERNMENT ORDERS [English]	GOVERNMENT ~~ORDERS~~ GOVERNMENT BUSINESS [English]	0.885	GOVERNMENT ORDERS ~~[English]~~	0.433	0.452
624.	(Vote n o 1360)	(Division No. 1360)	(Division No. 1360)	1.000	(Division ~~No.~~ n o 1360)	0.586	0.414
894.	DÉCLARATIONS DE DÉPUTÉS • (1400) [Français]	STATEMENTS BY MEMBERS • (1400) [Translation]	STATEMENTS BY MEMBERS • (1400) [Translation]	1.000	STATEMENTS BY MEMBERS • (1400) ~~[Translation]~~	0.657	0.343
872.	Encore une fois , cela ne devrait surprendre personne .	Again , this should come as no surprise .	Again , ~~this~~ it should come as no surprise .	0.968	Again , ~~this~~ that should ~~come as no~~ not ~~surprise~~ smear anyone .	0.636	0.333

Figure 8: Comparaison détaillée entre deux systèmes de TA

Exporter et manipuler les résultats

Si on souhaite effectuer une analyse plus poussée des résultats, ou simplement pour conserver le résultat d'une analyse, l'IU permet d'exporter les résultats d'une analyse dans un fichier en format CSV («*comma-separated values*»). Il suffit pour ce faire de cliquer sur le bouton «Exporter les résultats», se trouvant sous le sommaire des résultats (Figures 4 et 7). Il est alors possible d'insérer le contenu du fichier CSV dans un tableur Excel (par exemple).

Références

Federmann, C. (2012). Appraise: an open-source toolkit for manual evaluation of mt output. *The Prague Bulletin of Mathematical Linguistics*, *98*, 25-35.

Görög, A. (2014). Quantifying and benchmarking quality: the TAUS Dynamic Quality Framework. *Tradumàtica*, (12), 0443-454.

Han, L. (2016). Machine Translation Evaluation Resources and Methods: A Survey. *arXiv preprint arXiv:1605.04515*.

Läubli, S., Amrhein, C., Düggelin, P., Gonzalez, B., Zwahlen, A., & Volk, M. (2019). Post-editing Productivity with Neural Machine Translation: An Empirical Assessment of Speed and Quality in the Banking and Finance Domain. *arXiv preprint arXiv:1906.01685*.

Lavie, A., & Denkowski, M. J. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine translation*, *23*(2-3), 105-115.

Lo, C. K. (2019, August). YiSi-A unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)* (pp. 507-513).

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12), 0455-463.

Ma, Q., Wei, J., Bojar, O., & Graham, Y. (2019, August). Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)* (pp. 62-90).

Ma, Q., Bojar, O., & Graham, Y. (2018, October). Results of the WMT18 Metrics Shared Task: Both characters and embeddings achieve good performance. In *Proceedings of the third conference on machine translation: shared task papers* (pp. 671-688).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Moore, R. C., & Lewis, W. (2010, July). Intelligent selection of language model training data. In *Proceedings of the ACL 2010 conference short papers* (pp. 220-224). Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006, August). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas* (Vol. 200, No. 6).

White, J., O'Connell, T., & O'Mara, F. (1994, October). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas* (pp. 193-205).

Annexe:
API du système ÉQuaTion


```yaml
info:
 title: API for ÉQuaTion
 version: 1.0.0
 description: >-
   ÉQuaTion provides machine translation (MT) quality service using [YiSi]
(https://nrc.canada.ca/en/research-development/products-services/software-applications/yisi-so
ftware-semantic-machine-translation-evaluation-metric).

servers:
 - url: /api

components:
 schemas:
   Model:
     type: object
     description: Definition of a YiSi single model.
     properties:
       url:
         type: string
         format: uri
         description: The complete URL to access this YiSi model.
       name:
         type: string
         description: Model's description or name.

   MTEval:
     type: object
     description: The result of a single ÉQuaTion evaluation.
     properties:
       statistics:
         $ref: "#/components/schemas/Statistic"
       data:
         type: array
         description: A one-to-one mapping for each input segment.
         items:
           $ref: "#/components/schemas/MTEvalRow"

   MTEvalRow:
     type: object
     description: >-
       Reflects a single aligned segment across each input file, with ÉQuaTion results.
     properties:
       source:
         $ref: "#/components/schemas/Segment"
       reference:
         $ref: "#/components/schemas/Segment"
```

```yaml
    mt:
      type: array
      description: MT segments and ÉQuaTion results.
      items:
        $ref: "#/components/schemas/MTSegment"
      minItems: 1
      maxItems: 2

  Segment:
    type: object
    description: Metadata about a single Segment.
    properties:
      text:
        type: string
        description: The original segment.
      ntokens:
        type: integer
        description: The number of words contained in this segment.

  MTSegment:
    description: Metadata about a single Segment and ÉQuaTion results.s
    allOf:
      - $ref: "#/components/schemas/Segment"
      - type: object
        properties:
          diff:
            type: string
            format: "html"
            description: >-
              Contains the HTML-enriched segment. MT segments are compared with the reference
              segment and difference are highlighted. Uses `<ins />` to reflect text that was
              added, and `<del />` to reflect text that was removed.
          score:
            type: number
            format: double
            description: >-
              YiSi score of the reference segment and this MT segment.

  Statistic:
    type: object
    description: >-
      Metadata, statistics and global results for individual input files.
    properties:
      source:
        $ref: "#/components/schemas/FileStatistic"
      reference:
        $ref: "#/components/schemas/FileStatistic"
      mt:
        type: array
        description: Statistics about the machine translation input files.
```

```yaml
      items:
        $ref: "#/components/schemas/MTFileStatistic"
      minItems: 1
      maxItems: 2

FileStatistic:
  type: object
  description: Contains file statistics for source and reference files.
  required:
    - filename
    - n_segments
    - words
  properties:
    filename:
      type: string
      description: The file name as provided by the user.
    n_segments:
      type: integer
      description: The number of segments contained in the input file.
    words:
      type: integer
      description: Total number of words contained in the file.

MTFileStatistic:
  description: >-
    Contains file statistics and global results for the machine translated input files.
  allOf:
    - $ref: "#/components/schemas/FileStatistic"
    - type: object
      required:
        - duration
        - average
        - weighted_average
      properties:
        duration:
          type: number
          format: double
          description: The time in milliseconds required to perform YiSi evaluation.
        average:
          type: number
          format: double
          description: The non-weighted average of the YiSi score for each segments.
        weighted_average:
          type: number
          format: double
          description: >-
            The weighted average of the YiSi score for each segment, considering the
            number of words in each segment.
        yisi_preference:
          nullable: true
          type: number
```

```yaml
            format: double
            description: >-
              Available when two MT files are provided, this score represents how often YiSi
              preferred one MT over the other.

    RequestError:
      type: object
      description: A description of the user correctable error.
      properties:
        error:
          type: boolean
          description: Always true, represents an error.
        en:
          type: string
          description: The English error text, to be shown to the user.
        fr:
          type: string
          description: The French error text, to be shown to the user.

  paths:
   /models:
    get:
      description: Returns all known YiSi models.
      responses:
        "200":
          description: A list of models.
          content:
            application/json:
              schema:
                type: array
                items:
                  $ref: "#/components/schemas/Model"

   /models/{modelID}:
    post:
      description: Performs MT evaluation of the provided input files.
      parameters:
        - in: path
          name: modelID
          schema:
            type: string
          required: true
          description: Model ID to use for this MT evaluation.
      requestBody:
        required: true
        description: >-
          Each of these parameters are files containing segmented and aligned documents.
          ÉQuaTion will perform tokenization using Portage's tokenizer.
        content:
          multipart/form-data:
```

```yaml
      schema:
        type: object
        required:
          - reference
          - mt
        properties:
          source:
            type: string
            format: binary
            description: >-
              The text in the original source language.
          reference:
            type: string
            format: binary
            description: >-
              A reference text, usually the human translated version but can also be
              another machine translated version.
          mt:
            type: array
            description: >-
              At least one machine translated (MT) version of the text.
            items:
              type: string
              format: binary
            minItems: 1
            maxItems: 2

responses:
  "200":
    description: ÉQuaTion evaluation result.
    content:
      "application/json":
        schema:
          $ref: "#/components/schemas/MTEval"
  "400":
    description: Bad request.
    content:
      "application/json":
        schema:
          $ref: "#/components/schemas/RequestError"
  "404":
    description: A model with the specified ID was not found.
  "5XX":
    description: Unexpected error.
```

A2020-0026

Page(s) 247-263

fait/font l'objet d'un des prélèvement(s) dans leur intégralité

En vertu de l'article – ou des articles suivants :

20(1)(b)

de la *Loi sur l'accès à l'information*

Page(s) 247-263

Is/are withheld in their entirety

Pursuant to Section(s)

20(1)(b)

of the *Access to Information Act*

CONSEIL NATIONAL DE RECHERCHES CANADA

Gestion assistée des commandes au Bureau de la traduction

Rapport final sur la prédiction de domaine

Michel Simard Cyril Goutte Serge Léger

30 mars 2020

Résumé

Le présent rapport décrit les travaux entrepris dans le cadre d'un projet sur la gestion assistée des commandes de traduction (*Assisted Translation Assignment*) au Bureau de la traduction. Ce projet s'inscrit dans le contexte d'une entente de collaboration entre le Bureau de la Traduction et le Conseil national de recherches Canada, portant sur l'utilisation de méthodes d'intelligence artificielle pour soutenir la qualité des traductions au Bureau : "*Artificial Intelligence for Translation Quality*" – AI4TQ.

Le projet dont il est question ici porte d'une façon générale sur l'automatisation partielle ou entière de certaines des tâches des conseillers à la clientèle (CC) du Bureau, notamment l'identification du domaine de spécialité des textes soumis pour traduction. Nous avons utilisé des méthodes d'intelligence artificielle pour créer un système de prédiction du domaine de spécialité des textes. En exploitant les informations provenant d'une part de la mémoire de traduction du Bureau – le Mégacorpus – et d'autre part des bases de données des commandes, ce système arrive à distinguer les textes généralistes des textes spécialisés avec 78% d'exactitude, et à identifier le domaine de spécialisation des textes spécialisés avec moins de 15% d'erreurs. Une implémentation du système sous la forme de service Web permet une intégration simple dans les applications du Bureau.

Table des matières

1 Gestion des commandes au Bureau **5**

2 Données **6**
2.1 Mégacorpus 6
2.2 Base de données des commandes 7

3 Méthodes **13**
3.1 SVM 14
3.2 ProbCat 16
3.3 Classification biaxale 16
3.4 Segmentation automatique des documents 17

4 Évaluation **17**
4.1 Métriques d'évaluation 18
4.2 Résultats 19

5 Interface de démonstration **21**

6 Perspectives **25**
6.1 Prédiction des sous-domaines 25
6.2 Estimation de confiance 25
6.3 Entraînement continu 26
6.4 Découpage automatique 26

Introduction

Le présent rapport décrit les travaux entrepris dans le cadre d'un projet sur la gestion assistée des commandes de traduction (*Assisted Translation Assignment*) au Bureau de la traduction ("le Bureau" ci-après). Ce projet s'inscrit dans le contexte d'une entente de collaboration entre le Bureau et le Conseil national de recherches Canada (CNRC), portant sur l'utilisation de méthodes d'intelligence artificielle pour soutenir la qualité des traductions au Bureau : "*Artificial Intelligence for Translation Quality*" – AI4TQ.

Chaque année, les différents organismes du gouvernement fédéral soumettent au Bureau plus de 250 000 commandes. Plus de 60% de ces commandes concernent la traduction d'un ou de plusieurs documents vers l'une ou l'autre des langues officielles. Chaque commande est prise en charge par un "conseiller à la clientèle" (CC) du Bureau, qui l'accompagnera tout au long de son parcours dans les rouages de l'organisation. Cette gestion des commandes représente une fraction importante des coûts d'opération du Bureau. Le présent projet porte d'une façon générale sur l'automatisation partielle ou entière de certaines des tâches qui sont présentement effectuées de façon manuelle par les CC.

En particulier, nous nous intéressons, pour les tâches de traduction, à identifier les traducteurs ou fournisseurs externes qui sont les plus à même d'effectuer la tâche. En pratique, cette décision repose principalement sur le *domaine de spécialité* du texte, qui doit correspondre avec celui du traducteur ou fournisseur. L'objectif du présent projet est donc de mettre au point un système de prédiction automatique du domaine de spécialité d'un texte, adapté aux besoins des CC du Bureau. Nous brossons d'abord un portrait du processus de gestions des commandes au Bureau (Section 1), et des données sur lesquelles s'appuie notre travail (Section 2). Nous décrivons ensuite les méthodes que nous avons considérées (Section 3) et présentons les résultats que nous avons obtenus (Section 4) et le système que nous avons produit (Section 5. Finalement, nous discutons des perspectives d'avenir du projet (Section 6).

1 Gestion des commandes au Bureau

Chaque année, le Bureau reçoit plus de 250 000 commandes provenant de ses clients du gouvernement fédéral. Ces commandes concernent l'un ou l'autre des différents services offerts par le Bureau : relecture, révision, correction d'épreuve, etc. Mais la plus de 60% concernent la traduction d'un ou de plusieurs documents vers l'une ou l'autre des langues officielles. Chacune de ces commande est prise en charge par un des 80 "conseillers à la clientèle" (CC) du Bureau, qui suivra cette commande tout au long de son parcours dans les rouages de l'organisation.

La majorité des commandes de traduction arrivent via le *Système de commandes en direct*, que les clients du Bureau utilisent à la fois pour transmettre leurs textes et pour spécifier plusieurs des paramètres de la demande, tel que la nature du travail, la date butoir, etc. Mais la saisie doit être complétée par les CC. Cette tâche est relativement complexe et fastidieuse, nécessitant typiquement environ 20 "clics" par demande. Une importante proportion des commandes concernent de courts documents : 47% des documents français contiennent moins de 250 mots; pour l'anglais, c'est 54%. Pour ces documents, le travail des CC représente donc une part importante des coûts de traitement.

Chaque traducteur ou sous-traitant a son ou ses domaines de spécialités, qui sont consignées dans son profil. Le Bureau reconnaît 20 spécialités et 64 sous-spécialités. (Nous verrons ces domaines plus en détail dans la Section 2.2.) Comme c'est en partie sur la base de cette information que les CC déterminent à qui attribuer une tâche de traduction, l'identification du domaine de spécialité d'un document fait partie inhérente du leur travail. Les CC doivent donc connaître intimement les différents domaines de spécialité et être en mesure de les distinguer efficacement.

Une importante proportion des documents traduits par le Bureau appartient au domaine des "Textes administratifs et généraux" (*TAG*) : 55% des traductions en anglais, 56% des traductions en français. Il s'agit de documents de nature générale ou administrative qui portent sur les divers domaines d'activités du gouvernement du Canada. Comme la traduction de ces textes ne né-

cessite que des compétences générales, ils peuvent être confiés à n'importe lequel des traducteurs et des fournisseurs externes du Bureau.

Bien que la très grande majorité des documents traités par le Bureau ne concernent qu'un seul domaine de spécialité, il arrive occasionnellement qu'un document touche à plus d'un domaine : par exemple, un contrat pour l'achat de matériel informatique peut comporter à la fois une section "technique", décrivant la nature du matériel, et une section "légale", décrivant les obligations des parties. Pour identifier correctement les domaines de spécialité pertinents, les CC doivent donc parcourir l'ensemble du document. Dans le cas où le document couvre plusieurs domaines différents, le CC doit en outre spécifier un "découpage" du document (en pages ou en sections), permettant son partage entre plusieurs fournisseurs.

2 Données

Les données utilisées dans le présent projet proviennent de deux sources : d'une part, le *Mégacorpus*, qui constitue essentiellement la mémoire de traduction du Bureau; et d'autre part un jeu de données que nous appellerons "DC-23", qui est en fait extrait de différentes bases de données, et qui documente l'ensemble des commandes traitées par le Bureau entre 2015 et 2019. Nous décrivons plus en détail ces deux jeux de données ci-dessous.

2.1 Mégacorpus

Tel que mentionné plus haut, le Mégacorpus (MC) est la mémoire de traduction du Bureau. La version que nous avons obtenue du Bureau couvre une période allant de 2014 jusqu'à la fin de 2018 environ. Le corpus consiste en un très volumineux ensemble de fichiers TMX, un par document traduit. Ces documents sont regroupés par "client" : on dénombre 221 de ces clients : certains sont de réels clients du Bureau (Défense nationale, Pêches et océans, etc.), d'autres désignent en fait des regroupements de documents de même nature (SAP-SIGMA, "Staffing").

Sous chacun de ces clients, les documents sont regroupés par direction de traduction (anglais-français ou français-anglais). Pour chaque document, on retrouve un fichier TMX qui contient une suite d'unités de traduction (typiquement, des phrases) dans leurs versions anglaise et française. À chaque document est associé un identificateur numérique unique : c'est le numéro de commande. Le Tableau 2.1 résume les principales caractéristiques du corpus.

	anglais-français	français-anglais
Documents	1 474 365	332 277
Unités de traduction (phrases)	115,6M	23,7M
Mots en langue source	1602M	415M

Tableau 2.1 – Caractéristiques du Mégacorpus

Le Mégacorpus est au centre d'autres projets de AI4TQ, notamment un projet sur le "nettoyage de corpus". En effet, il s'avère que le corpus recèle une certaine quantité de "bruit", c'est-à-dire de contenu non-utilisable. Ce bruit peut prendre la forme de documents mal alignés, non-traduits, dont la traduction est d'une qualité douteuse, de documents comportant des problèmes d'encodage de caractère, etc. Pour les fins du présent projet, nous avons choisi dans un premier temps d'ignorer ces problèmes.

2.2 Base de données des commandes

Le jeu de données "DC-23" consiste en une collection d'enregistrements décrivant toutes les commandes traitées par le Bureau entre le premier avril 2014 et le 31 mars 2019, soit plus de 1,2 million de commandes. Chaque enregistrement comporte les champs suivants :

Numéro de commande Un identificateur unique associé à chaque commande. C'est ce champ qui nous permet de faire le lien avec les documents dans le Mégacorpus.

Code et nom du compte-client Ce champ identifie non seulement le

client, mais également le compte auquel la commande est facturée.

Nom du demandeur le nom de la personne chez le client qui a placé la commande

Spécialité le domaine de spécialité du texte, tel qu'identifié par le CC du Bureau

Sous-spécialité le domaine de sous-spécialité du texte, tel qu'identifié par le CC du Bureau

Tâche exécutée L'une des valeurs suivantes : Traduction, Traduction et service linguistique, Correction d'épreuves, Relecture, Relecture intégrée, Dactylo/éditique, Révision, Révision comparative – facturable, Révision d'une rédaction – facturable, Traducteur sur place, Terminologie/documentation, Services administratifs, Gestion de projet, Aide à la rédaction, Affermage.

Nom de la ressource Le nom de la ou des ressource(s) à qui la tâche est confiée : il peut s'agir du nom d'un ou plusieurs employé(s) du Bureau, ou du nom d'un fournisseur externe (pigiste ou agence), assorti d'une référence au contrat le liant au Bureau.

Date du début la date et l'heure où la commande a été reçue.

Date de livraison la date et l'heure où le livrable a été envoyé au client.

Prévu pour exécution le nombre d'heures prévues pour l'exécution de la tâche

Nombre de tâches exécutées dans la très grande majorité des cas (99,5%), ce champ prend la valeur 1, mais pour certains types de commandes, il peut être supérieur à 1.

Pour le présent exercice, nous nous concentrons uniquement sur les commandes de traduction, c'est-à-dire les commandes dont le champ "Tâche exécutée" prend la valeur "Traduction" (près de 790 000 enregistrements), "Traduction et service linguistique" (2763 enregistrements) ou "Traducteur sur place" (9496 enregistrements). De ce nombre, nous ne retenons toutefois que les commandes dont le texte apparaît dans notre version du Mégacorpus, et nous écartons toutes les commandes dont le champ "Spécialité" contient plus

d'un domaine (37 699 commandes). Au final, il reste environ 460 000 enregistrements utilisables pour nos fins.

Notre plan initial était de créer un système de prédiction des ressources : l'idée était d'exploiter les caractéristiques du texte associé à la commande pour suggérer directement une liste de traducteurs ou de fournisseurs externes susceptibles d'effectuer la traduction. En pratique, toutefois, il s'est avéré que ce plan comportait des inconvénients majeurs. Le plus important est certainement que la liste des traducteurs et fournisseurs disponibles est en constante mutation : chaque mois, des traducteurs partent à la retraite, d'autres sont embauchés, les contrats avec les fournisseurs externes viennent à terme et sont remplacés par d'autres contrats, etc. Cette situation rendait la gestion et la maintenance d'un tel système extrêmement complexe.

En revanche, il est rapidement apparu que pour les CC, les informations les plus importantes pour identifier les ressources appropriés étaient les domaine de spécialité et de sous-spécialité du document. En effet, les traducteurs et les fournisseurs externes ont tous un ou plusieurs domaines de spécialité, qui sont consignées à même leur profil, et c'est essentiellement en combinant cette information avec la disponibilité des ressources que le CC effectue l'affection des commandes. Donc, plutôt que de travailler directement avec les noms de ressources, nous avons opté pour la création d'un système de prédiction du domaine de spécialité. Nous avons combiné le contenu du Mégacorpus et du DC-23 pour créer un ensemble de données d'apprentissage pour cette tâche, qui prennent la forme de tuples :

$\langle$ *texte, spécialité, sous-spécialité* $\rangle$.

Après une série d'expériences avec ces données, il est apparu que les données DC-23 reposaient en fait sur deux systèmes distincts de dénomination des spécialités. En effet, le Bureau a redéfini l'ensemble des domaines de spécialité et de sous-spécialités en août 2018. À partir de cette date, les nouvelles spécialités ont graduellement remplacé les anciennes. Comme le DC-23 couvre la période allant d'avril 2014 à août 2019, les deux jeux de spécialités sont représentés. Il n'existe pas de correspondance biunivoque entre les anciennes et nouvelles spécialités. Malgré tout, avec l'aide des experts du Bu-

reau, nous avons pu créer une table de conversion permettant de projeter les anciennes spécialités vers les nouvelles, et ainsi ré-étiqueter toutes les données de façon uniforme. Le Tableau 2.2 contient la liste des spécialités et sous-spécialités présentement en vigueur au Bureau. Ce tableau donne le nombre de documents pour chaque spécialité et sous-spécialités. Les pourcentages dans la dernière colonne sont la proportion que représente la sous-spécialité au sein de la spécialité.

Toutes les spécialités ne sont pas subdivisées en sous-spécialités. Celles qui le sont (BIO, CRIM, DROIT, ENV, MEC, MED, MIL, SCN, SOC, TECH et TRA) comportent chacune une sous-spécialité "générale", par exemple : CRIM pour CRIM et SOC-Soc pour SOC. En pratique, on observe que les "vraies" sous-spécialités n'occupent qu'une petite fraction de l'ensemble des données. Par exemple, la spécialité Criminologie (CRIM) se subdivise en sous-spécialités concernant les services policiers (CRIM-Pol), les services frontaliers (CRIM-Front), la sécurité publique (CRIM-SecP) et le service correctionnel (CRIM-SCorr). Mais en pratique, la grande majorité (91%) des commandes de cette spécialité sont simplement classés sous CRIM. De ce fait, bien que nous disposions d'une quantité impressionnante de données, pour beaucoup de sous-spécialités, le nombre d'exemples n'est pas suffisant pour permettre l'utilisation de méthodes d'apprentissage-machine. Nous reviendrons sur cette question dans la Section 3.

Tableau 2.2 – Spécialités et sous-spécialités

Spécialité	**# doc.**	**% corpus**	**sous-spéc.**	**# doc.**	**% spéc.**
AUT	5923	1.3%	AUT	5923	100.0%
BI	2491	0.5%	BI	2491	100.0%
BIO	2220	0.5%	BIO	1953	88.0%
			BIO-Forêts	20	0.9%
			BIO-Pêche	77	3.5%
			BIO-ZooVeg	158	7.1%

Suite à la page suivante

Tableau 2.2 – *Suite de la page précédente*

Spécialité	# doc.	% corpus	sous-spéc.	# doc.	% spéc.
			BIO-techno	12	0.5%
CRIM	19122	4.1%	CRIM	17402	91.0%
			CRIM-Front	250	1.3%
			CRIM-Pol	492	2.6%
			CRIM-SCorr	733	3.8%
			CRIM-SecP	245	1.3%
DROIT	10103	2.2%	DROIT	9377	92.8%
			ConPéAutFa	49	0.5%
			DROIT-Comm	39	0.4%
			DROIT-Conv	15	0.1%
			DROIT-FisCi	70	0.7%
			DROIT-MarTr	20	0.2%
			DROIT-Trib	533	5.3%
EMP	20813	4.5%	EMP	20813	100.0%
ENV	11906	2.6%	ENV	10320	86.7%
			ENV-Agr	438	3.7%
			ENV-Eco	514	4.3%
			ENV-Fossil	34	0.3%
			ENV-Ind	460	3.9%
			ENV-Nuc	140	1.2%
FIN	4538	1.0%	FIN	4538	100.0%
IMM	32091	6.9%	IMM	32091	100.0%
INF	3990	0.9%	INF	3990	100.0%
MEC	2704	0.6%	MEC	2157	79.8%
			MEC-Air	120	4.4%
			MEC-Armes	42	1.6%
			MEC-Eau	136	5.0%

Suite à la page suivante

Tableau 2.2 – *Suite de la page précédente*

Spécialité	**# doc.**	**% corpus**	**sous-spéc.**	**# doc.**	**% spéc.**
			MEC-Machine	159	5.9%
			MEC-TerFer	90	3.3%
MED	5467	1.2%	MED	5350	97.9%
			MED-Dents	15	0.3%
			MED-SST	92	1.7%
			MED-Vet	10	0.2%
MIL	25014	5.4%	MIL	23952	95.8%
			MIL-Air	207	0.8%
			MIL-Mer	62	0.2%
			MIL-Terre	793	3.2%
MP	11547	2.5%	MP	11547	100.0%
PARL	11	0.0%	PARL	11	100.0%
SCN	2916	0.6%	SCN	2584	88.6%
			SCN-Atmosp	46	1.6%
			SCN-Chimie	102	3.5%
			SCN-Geo	69	2.4%
			SCN-Math	9	0.3%
			SCN-Ocean	19	0.7%
			SCN-Phys	13	0.4%
			SCN-Spatial	37	1.3%
			SCN-Teled	37	1.3%
SOC	29528	6.4%	SOC-Arts	557	1.9%
			SOC-Polit	785	2.7%
			SOC-Soc	28186	95.5%
TAG	260843	56.4%	TAG	260843	100.0%
TECH	6369	1.4%	TECH	5879	92.3%
			TECH-Bois	15	0.2%

Suite à la page suivante

Tableau 2.2 – *Suite de la page précédente*

Spécialité	# doc.	% corpus	sous-spéc.	# doc.	% spéc.
			TECH-Constr	185	2.9%
			TECH-Electr	100	1.6%
			TECH-Mat	21	0.3%
			TECH-Meteo	33	0.5%
			TECH-Mines	3	0.0%
			TECH-Radio	34	0.5%
			TECH-Textil	99	1.6%
TRA	4822	1.0%	TRA	4559	94.5%
			TRA-Air	120	2.5%
			TRA-Eau	58	1.2%
			TRA-Route	52	1.1%
			TRA-Train	33	0.7%
Total	**462418**	**100.0%**		**462418**	

3 Méthodes

La classification de textes en domaines est une tâche qui se prête naturellement aux méthodes d'apprentissage-machine. Chaque transaction de l'ensemble DC-23 produit naturellement un exemple d'apprentissage (x, y), où x est une caractérisation du document et y l'un des domaines de spécialité (ou sous-spécialité) [1]. Le grand nombre de transactions disponibles dans le DC-23 nous permet d'envisager l'utilisation de méthodes reposant sur des caractérisations très fines (c'est-à-dire : impliquant un très grand nombre d'attributs – *features* en anglais) produites de façon entièrement automatique. Les méthodes que nous avons considérées sont les *Support vector machines* (SVM) et

1. En pratique, nous produisons deux exemples d'apprentissage à partir de chaque enregistrement : un pour l'anglais et un pour le français. En effet, bien que la tâche réelle consiste à classifier des textes en langue source, jusqu'à preuve du contraire, il n'y a pas de raison de penser que les traductions constituent des données d'apprentissage inadéquates.

une approche probabiliste de type *Naive Bayes*, que nous appelons *ProbCat*. Dans les deux cas, les documents sont caractérisés simplement par les mots qu'ils contiennent : chaque document est représenté par un vecteur de la taille du vocabulaire du corpus, où chaque composante représente l'utilisation d'un mot du vocabulaire dans le document. Cette approche est appelée *sac de mots* (*bag of words* en anglais) car elle ne tient pas compte de l'ordre dans lequel apparaissent les mots du texte.

Au moment de caractériser un document, nous excluons les "mots vides", c'est-à-dire principalement les mots grammaticaux très communs qui se retrouvent dans beaucoup de documents et ne sont pas susceptibles de nous aider à distinguer les domaines de spécialités. Des listes de mots vides pour l'anglais et le français sont fournies au système lors de l'extraction des attributs de chaque document. Nous écartons aussi du vocabulaire indexé les mots très peu fréquents dans le corpus. En effet, un mot apparaissant dans seulement un ou deux document est peu susceptible d'apparaître dans de futurs documents à classifier, et donc d'être utile à la classification. Par ailleurs, la distribution naturelle des mots dans le langage fait que de très nombreux mots apparaissent très peu souvent, et très peu de mots constituent le gros du vocabulaire typiquement utilisé. En filtrant ainsi les mots apparaissant dans peu de documents, nous pouvons donc réduire de façon très significative la taille du vocabulaire et donc optimiser la taille des données, des modèles et le temps de calcul.

3.1 SVM

Les modèles SVM (*Support Vector Machine*) sont des modèles statistiques populaires et historiquement très efficaces sur les tâches de classification de textes. À partir de vecteur d'attributs caractérisant un certain nombre de documents x et les domaines de spécialité y, nous pouvons estimer un modèle $f(x)$ qui offre une prédiction d'un domaine pertinent pour tout nouveau document.

Les modèles SVM sont à la base de modèles de classification binaire, c'est à dire qu'ils permettent de classifier un document dans l'une de deux classes

mutuellement exclusives. Dans la situation qui nous intéresse, il s'agit de prédire le domaine pertinent parmi une liste contenant beaucoup plus de deux domaines. Dans ce cas, nous utilisons l'approche dite *un contre tous* : pour les 20 domaines du Bureau, nous construisons 20 classifieurs SVM, chacun prédisant la probabilité qu'un document soit dans l'une des 20 classes plutôt que dans l'une des 19 autres : le premier SVM prédit le domaine AUT contre le reste, le second le domaine BI contre le reste, etc.

Nos modèles de SVM sont construits en quatre étapes :

1. extraction et décompte des mots,
2. pondération des mots,
3. estimation de classifieurs statistiques SVM,
4. calibration des classifieurs.

L'étape d'extraction et décompte est décrite brièvement ci-dessous. Pour la pondération des mots, nous utilisons la méthode *ltc* dans la nomenclature du système de recherche d'information SMART [2], qui attribue un poids plus élevé aux mots plus discriminants. C'est une pondération de type *tf.idf* qui combine la fréquence des mots (*term frequency*), [3] la fréquence documentaire inverse (*inverse document frequency*), [4] et normalise le tout pour produire un score indépendant de la longueur du document.

Nous entraînons un classifieur SVM pour chacun des domaines de spécialité, avec l'approche *un contre tous*. Chaque classifieur produit alors un score qui indique si un document appartient ou non à "sa" spécialité, plutôt qu'à une autre. Il se peut cependant que deux classifieurs prédisent que le document appartient à "leur" domaine, et ces scores ne sont pas directement comparables. Nous passons donc par une étape de calibration, qui transforme le score produit par chaque classifieur en une "vraie" probabilité, c'est à dire un

2. https://en.wikipedia.org/wiki/SMART_Information_Retrieval_System

3. Dans la pondération *ltc* on utilise en fait le logarithme de cette fréquence pour limiter l'influence des mots très fréquents.

4. La fréquence documentaire est le nombre de documents dans lequel un mot apparaît ; en divisant par cette fréquence, on sous-pondère les mots qui apparaissent partout et sont donc peu discriminants.

nombre entre 0 et 1, tel qu'un document avec une probabilité calibrée de 0.8 pour le classifieur AUT a 80% de chance d'être effectivement dans le domaine "Autochtones". À l'utilisation, nous soumettons la caractérisation x d'un document à chacun des classifieurs : le domaine de spécialité prédit est alors celui du classifieur qui produit la probabilité calibrée la plus élevée. Les détails de cette approche pour une application similaire sont décrits dans [1].

3.2 ProbCat

Les classifieurs ProbCat reposent sur des modèles statistiques génératifs des co-occurrences de mots w et des documents d, sous l'hypothèse que chaque document appartient à une catégorie c (le domaine de spécialité) qui conditionne la probabilité d'observer ou non certains mots. Les paramètres du modèle sont essentiellement les probabilités $P(w|c)$ qu'un mot w apparaisse dans un document de la classe c. Ces probabilités sont estimées directement à partir des observations du nombre d'occurrences du mot w dans les documents de la catégorie c, d'une façon identique à celui de la méthode dite *Naive Bayes*. À l'utilisation, le calcul du domaine le plus probable pour un document donné se résout comme un problème d'optimisation convexe, au moyen d'une méthode itérative (*Expectation Maximization* – EM). Les détails de cette approche sont décrits par exemple dans [2].

3.3 Classification biaxale

Plus de la moitié des documents soumis pour traduction au Bureau tombent dans la catégorie des Textes administratifs et généraux – TAG. D'une certaine façon, cette spécialité constitue un cas à part : ce qui caractérise un document TAG, contrairement aux autres spécialités, ce n'est pas le sujet ou la terminologie, mais plutôt le degré de complexité. Un document TAG peut en essence traiter d'un sujet médical ou légal, mais dans la mesure où le texte est suffisamment simple dans sa structure et son traitement du sujet, il n'est pas nécessaire que le traducteur soit un expert du domaine.

Cette particularité du domaine suggère une catégorisation suivant deux

axes : un premier classifieur détermine si le document est TAG ou non, un deuxième classe le document dans l'une des autres spécialités. En pratique, il nous est apparu que cette façon de découper l'information était peut-être plus utile pour les CC. Par ailleurs, cette approche biaxale est possible aussi bien avec les classifieurs SVM que ProbCat et il est possible de mélanger les deux (utiliser un SVM pour un axe et ProbCat pour l'autre). Comme nous le verrons plus loin, cette façon de découper le problème comporte également des avantages du point de vue de la performance des classifieurs.

3.4 Segmentation automatique des documents

Un même document peut se rattacher à plus d'un domaine de spécialité : par exemple, un contrat pour l'achat de matériel informatique peut comporter à la fois une section "technique", décrivant la nature du matériel, et une section "légale", décrivant les obligations des parties. Pour identifier correctement les domaines d'un document, les CC doivent donc effectuer une lecture en diagonale de l'ensemble des documents. Dans le cas où le document couvre plusieurs domaines différents, le CC doit en outre spécifier un "découpage" du document (en pages ou en sections), permettant son partage entre plusieurs fournisseurs.

Notre ensemble d'apprentissage ne comporte pas de tels documents "découpés". Malgré tout, rien n'empêche d'appliquer nos classifieurs sur des unités plus petites qu'un document. Il est donc possible de segmenter un document en sous-parties, d'analyser séparément chacune de ces sous-parties, et d'utiliser le résultat pour proposer un découpage du document par spécialité.

4 Évaluation

L'estimation des modèles de classification statistique repose sur l'utilisation de données annotées, c'est à dire un ensemble de documents pour lesquels les domaines de spécialités sont connus. Comme les paramètres du classifieur sont ajustés pour produire des sorties les plus proches possibles de ces don-

nées annotées, les performance d'un modèle ont tendance à être optimistes sur les données utilisées pour son estimation. Afin d'obtenir une estimation réaliste et non biaisée de la performance réelle des modèles, il est donc nécessaire de tester sur des données qui n'ont pas été utilisées pour ajuster le modèle.

Pour cela, nous utilisons la méthode dite de *validation croisée* (*k-fold cross validation*). En partant de l'ensemble de tous les documents du DC-23 avec leurs domaine de spécialité, nous divisons les données en k tranches de taille égale. Nous utilisons alors la concaténation de $k-1$ tranches pour estimer les classifieurs et la tranche restante pour en évaluer la performance sur des données jamais vues. En répétant cette procédure k fois, chaque fois avec une tranche d'évaluation différente, nous obtenons des prédictions non biaisées pour la totalité des données disponibles, et nous pouvons alors utiliser différentes métriques d'évaluation sur ces prédictions.

4.1 Métriques d'évaluation

Pour les problèmes de classification multi-classe, la métrique la plus naturelle est l'*exactitude* (en anglais, *accuracy*). Dans notre cas, il s'agit du nombre de documents pour lesquels un classifieur donné prédit le bon domaine de spécialisation, divisé par le nombre total de documents. Par exemple, si on teste un certain classifieur sur un ensemble de 1000 documents, que le système prédit la bonne spécialisation pour 650 d'entre eux et qu'il se trompe de spécialisation pour les 350 autres, on rapportera un taux d'exactitude de 65%.

Nos classifieurs calibrés permettent aussi de produire, pour chaque document, une liste ordonnée de classes prédites, par probabilité décroissante. Lorsque la vraie classe n'est pas en tête de liste (donc avec la probabilité la plus élevée), on comprend qu'il est plus souhaitable qu'elle se trouve en seconde position qu'en dernière position, mais la métrique d'exactitude n'en tient pas compte : cet exemple sera simplement comptabilisé comme une erreur de classification.

Pour tenir compte plus précisément de l'ordre des prédictions, nous calcu-

lons la métrique du rang réciproque moyen (*mean reciprocal rank* ou **MRR**). Pour chaque document d, nous classons les domaines prédits par ordre décroissant de probabilité, et nous faisons la moyenne des inverses des rangs de la classe correcte.[5] Lorsque la classe correcte est en tête de liste, elle contribue 1 à la moyenne, 1/2 lorsqu'elle est en seconde position, 1/3 lorsqu'elle figure en troisième place, etc. Un MRR de 1 signifie donc que la classe correcte est toujours la plus probable, alors qu'un MRR de 0.5 suggère que "en moyenne" la classe correct est en deuxième position. Par construction le MRR est compris entre 0 et 1, et on désire un MRR le plus élevé possible.

4.2 Résultats

Nous avons expérimenté avec plusieurs configurations, notamment l'utilisation de classifieurs distincts pour l'anglais et le français, des caractérisations utilisant des "bigrammes" (paires de mots adjacents) plutôt que des mots, l'ajout de méta-information dans la caractérisation (taille du document, longueur moyenne des phrases, etc.). Nous ne rapportons pas les résultats de toutes ces variantes ici, seulement ceux des méthodes de base, et des variantes qui ont le mieux fonctionné.

Nous présentons d'abord les résultats obtenus dans un scénario de classification "simple" (par opposition à la classification biaxale). Les deux premières lignes du Tableau 4.1 ("base") donnent les résultats obtenus par les classifieurs *ProbCat* et *SVM* présentés à la section 3. D'une façon générale, les SVM l'emportent en termes d'exactitude, mais les ProbCats produisent un meilleur classement des spécialités (MRR plus élevé).

Comme nous l'avons expliqué plus haut, nos classifieurs se basent sur une caractérisation des documents en "sacs de mots" (*bag of words*). Toutefois, il nous est apparu que le code et le nom du compte-client (voir la section 2.2) pouvait également s'avérer utile pour prédire le domaine de spécialité. Les troisième et quatrième lignes du Tableau 4.1 ("+client") rapportent les performances de classifieurs qui ont également accès à cette information dans la

5. https://en.wikipedia.org/wiki/Mean_reciprocal_rank

Système	Exactitude	MRR
base :		
ProbCat	0.6145	0.7751
SVM	0.7231	0.6554
+client :		
ProbCat	0.6177	**0.8175**
SVM	**0.7462**	0.6979

Tableau 4.1 – Résultats de la classification simple (un seul axe).

caractérisation du document.

Comme on peut le voir, le simple ajout de cette information améliore substantiellement les résultats. Par exemple, pour les SVM, on gagne plus de 2 pour cent en exactitude (de 0.7231 à 0.7462) et 4 pourcent de MRR (de 0.6554 à 0.6979). Ceci suggère qu'il existe une dépendance entre les clients et les domaines de spécialités : un même document peut être affecté à une spécialité différente dépendant d'où il provient. Notons que nous avons expérimenté avec l'ajout d'autres méta-données dans la caractérisation des documents, tels que la longueur du document, la longueur moyenne des phrases, etc. Seul le nom du compte-client s'est avéré fructueux.

Le Tableau 4.2 présente les résultats des classifications biaxales : d'une part entre TAG et non-TAG, d'autre part entre les autres spécialités. Ici encore, on rapporte les résultats obtenus avec et sans l'information sur le client ayant soumis la demande ("base" et "+client").

Ici encore, l'ajout de l'information sur le client permet des gains substantiels, et ce sur les deux axes. Par ailleurs, on remarque que les résultats obtenus suivant les deux axes sont meilleurs que ceux obtenus en classification simple. Le fait de séparer les deux tâches semble simplifier substantiellement le problème pour les classifieurs.

Axe 1 (TAG vs. non-TAG)		
Système	Exactitude	MRR
base :		
ProbCat	0.7371	–
SVM	0.7666	–
+client :		
ProbCat	0.7563	–
SVM	**0.7815**	–

Axe 2 (spécialités autres que TAG)		
base :		
ProbCat	0.7387	0.8217
SVM	0.8051	0.8002
+client :		
ProbCat	0.8159	**0.8538**
SVM	**0.8583**	0.8348

Tableau 4.2 – Résultats de la classification biaxale.

5 Interface de démonstration

Nous avons mis au point un système de prédiction des spécialités, reposant sur des classifieurs SVM et une approche de classification biaxale tel que décrit plus haut. Nous décrivons ici l'interface de démonstration de ce système, que nous appelons *L'aide-conseiller à la clientèle*. Il s'agit d'une interface Web, dans laquelle l'utilisateur sélectionne le nom du client ayant soumis la demande dans la case "Ministère", et colle le contenu du document soumis dans la case "Texte" (Figure 5.1).

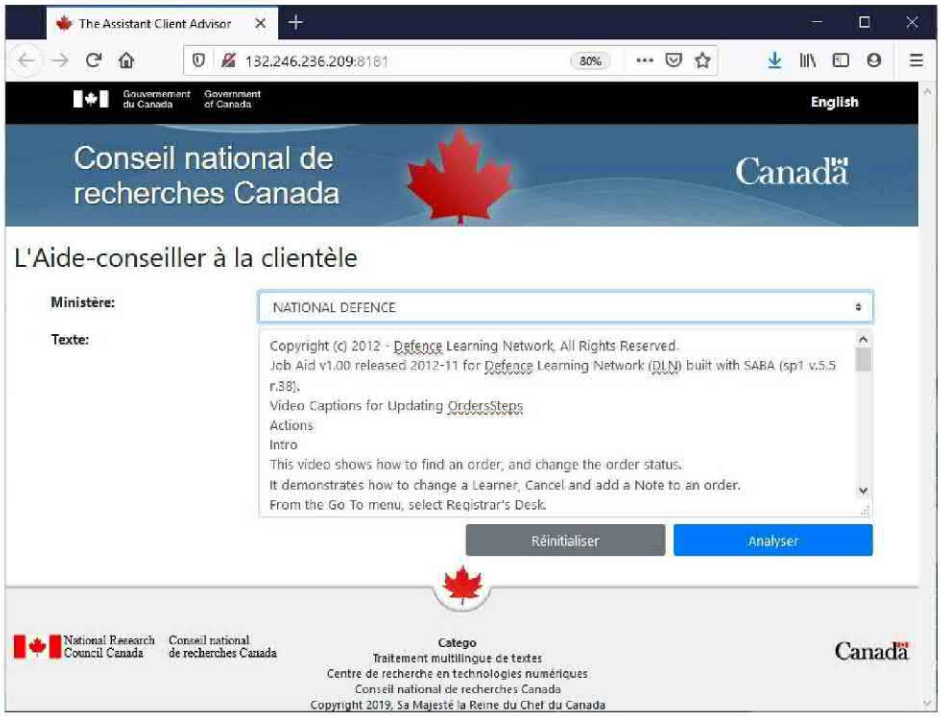


FIGURE 5.1 – Interface du *L'aide-conseiller à la clientèle*

Lorsque l'utilisateur appuie sur le bouton "Analyser", le système effectue une classification sur deux axes, et affiche sous "Domaines Prédits" d'abord le résultat de la prédiction TAG/non-TAG, puis celle du domaine de spécialité. Chaque décision est assortie de la probabilité produite par le classifieur correspondant. Pour l'axe des spécialités, le système affiche non seulement la spécialité la plus probable, mais aussi les principales spécialités concurrentes (Figure 5.2).

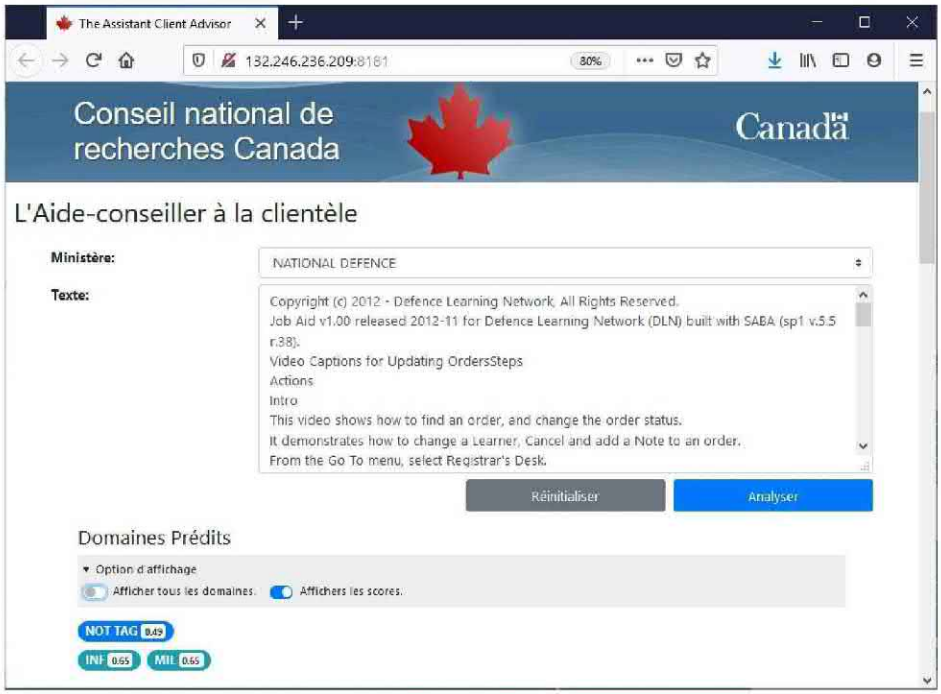


FIGURE 5.2 – Résultat d'une analyse

À la suite de cette information, le système affiche la "Distribution des domaines", c'est-à-dire un découpage du texte en différents domaines de spécialité. Les pastilles au haut de cette section indiquent la couleur employée pour désigner chaque fragment du découpage, et le pourcentage du texte représenté par chaque spécialité (Figure 5.3).

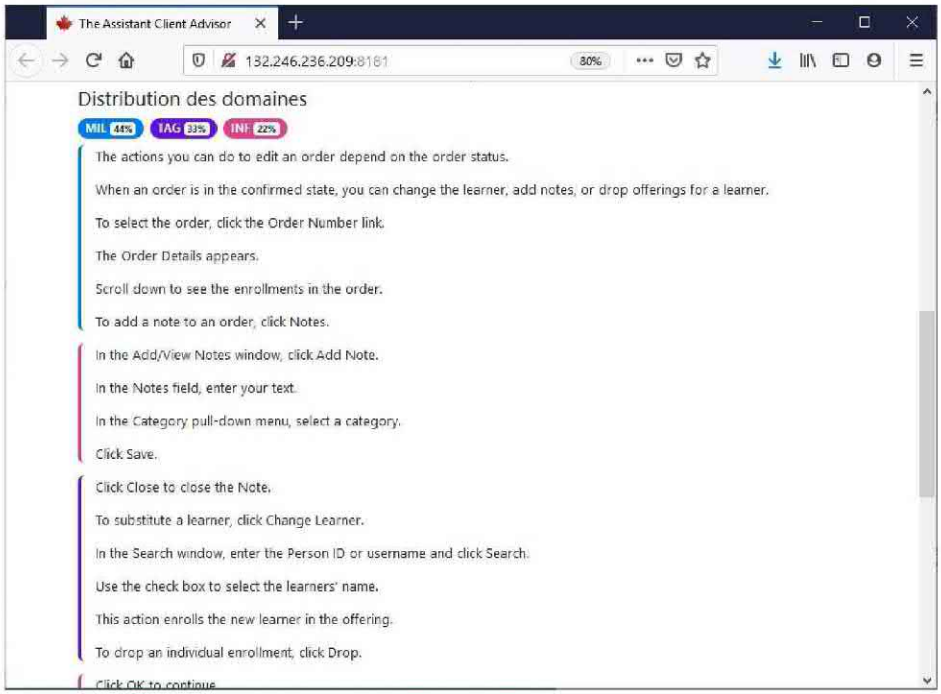


FIGURE 5.3 – Découpage par domaine de spécialisation.

Ce découpage repose sur une segmentation du documents en phrases ; les phrases sont ensuite regroupées pour constituer des blocs d'au moins 25 mots. Chaque bloc est alors analysé séparément. Les blocs consécutifs pour lesquels le système prédit la même spécialité sont fusionnés.

L'interface est disponible en français et en anglais, et peut traiter des textes dans l'une ou l'autre langue.

6 Perspectives

Nous avons décrit ci-dessus l'ensemble du travail effectué dans le cadre de ce projet. Nous abordons ci-dessous quelques-unes des thèmes qui pourraient éventuellement être explorés dans l'avenir.

6.1 Prédiction des sous-domaines

Dans les expériences rapportées plus haut, nous nous sommes concentrés sur la prédiction du domaine de spécialité. La classification en sous-spécialités ajoute une couche de complexité. Comme nous l'avons expliqué plus haut, il existe aussi certains problèmes de sous-représentation dans les sous-spécialités : pour certaines d'entre elles, nous n'avons pas assez d'exemples pour entraîner des classifieurs fiables.

Une solution possible à ce problème est de remanier les domaines, en regroupant certaines sous-spécialités peu représentées sous une étiquette commune. Dans un scénario d'automatisation partielle du travail des CC, les documents recevant une de ces étiquettes groupées devraient être revus par un expert.

La principales questions à résoudre ici sont celle du regroupement optimal des "petites" spécialités, et celle de la meilleure méthode pour effectuer ce genre de classification hiérarchique.

6.2 Estimation de confiance

En pratique, on observe que certains documents sont très faciles à classer, alors que d'autres présentent une forte ambiguïté. Les classifieurs que nous utilisons (SVM et ProbCat) produisent une valeur que nous avons décrite comme la probabilité que le document appartiennent à ce domaine de spécialité (Section 3). Alternativement, cette valeur peut être interprétée comme le niveau de confiance du classifieur, et utilisée pour séparer les cas faciles des cas difficiles : les cas faciles pourraient être traités de façon entièrement au-

tomatique, alors que les autres seraient redirigés vers un CC pour un examen plus approfondi.

Les questions à étudier ici sont celles de la manière d'obtenir des estimations de confiance des classifieurs, et de fixer des seuils permettant des garanties de fiabilité.

6.3 Entraînement continu

Dans un scénario où la prédiction de la spécialité des texte est employée comme aide aux CC (par opposition à une automatisation complète), il semble que la rétroaction des utilisateurs pourrait être utilisée pour améliorer la performance des classifieurs en continu. Différents méthodes ont été proposées pour l'apprentissage "en continu" des classifieurs tel que ceux que nous utilisons.

Pour exploiter efficacement la rétroaction des CC, il faudrait déterminer expérimentalement la ou les méthodes optimales. On voudrait également étudier comment gérer d'éventuels changements au systèmes de spécialité : comment faire lorsqu'une nouvelle spécialité apparaît, qu'une spécialité existante disparaît ou est subdivisée en sous-spécialités, ou que plusieurs spécialités existantes sont fusionnées ?

6.4 Découpage automatique

Tel que mentionné à la section 1, certains documents peuvent appartenir à plus d'un domaine : les CC doivent donc parcourir chaque document pour spécifier un "découpage" du document (en pages ou en sections), permettant son partage entre plusieurs fournisseurs.

Les méthodes que nous avons développées peuvent être appliquées à des documents entiers, mais également à des fragments plus petits. En classifiant séparément chaque page ou section d'un document volumineux, on peut facilement produire un découpage de celui-ci, et ainsi soumettre différentes parties à différents spécialistes. En pratique, plus l'échantillon de texte soumis est

petit, moins la prédiction produite par nos classifieurs sera fiable. Il est donc hasardeux de soumettre des fragments trop petits aux classifieurs.

Les questions à étudier ici sont celles de la fiabilité des prédictions en fonction de la taille des segments traités, la détection des points de changement, la granularité optimale des découpages et l'utilisation d'information structurelle pour le découpage automatique des textes.

Conclusions

Nous avons décrit les travaux effectués dans le cadre de ce projet portant sur la gestion assistée des commandes au Bureau. La tâche principale sur laquelle nous nous sommes penchés est la prédiction du domaine de spécialité d'un texte à partir de ses caractéristiques linguistiques. Nous avons proposé l'utilisation de méthodes d'apprentissage machine, et présenté nos résultats expérimentaux.

Globalement, les résultats obtenus nous semblent suffisamment bons pour envisager l'utilisation de ce genre de classifieurs dans un scénario d'automatisation partielle de la tâche des conseillers à la clientèle (CC) du Bureau. Nous préconisons une classification suivant deux axes, dans laquelle on détermine d'une part si le document est ou non un TAG (Texte administratif ou général), d'autre part à quelle domaine de spécialité il appartient. En effet, la distinction TAG / non-TAG nous semble orthogonale à la prédiction du domaine de spécialité, puisqu'elle dépend plutôt d'une mesure du degré de spécialisation que de la nature exacte de cette spécialisation. Nous croyons que les CC feraient meilleur usage d'une classification suivant ces deux axes. Par ailleurs, pour les documents présentant plus d'un domaine de spécialisation, nous croyons que nos méthodes peuvent être utilisées pour produire un découpage approximatif. Cette fonctionnalité devra toutefois faire l'objet d'essais plus poussés.

Nous n'avons pas abordé la question des sous-spécialités, parce que pour certaines de celles-ci, nous ne disposons pas de suffisamment d'exemples pour produire des classifieurs fiables. Une solution serait d'attribuer une étiquette unique aux documents tombant dans l'une de ces classes sous-

représentées, et de rediriger systématiquement ces documents vers un traitement manuel. Par ailleurs, il est possible de quantifier le niveau de confiance des classifieurs vis-à-vis d'une décision particulière ; cette information pourrait être utilisée pour départager les documents se prêtant à un traitement entièrement automatique de ceux qui requièrent une intervention manuelle.

Références

[1] G. Bernier-Colborne, C. Goutte, and S. Léger. Improving cuneiform language identification with bert. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 17–25, 2019.

[2] C. Goutte, S. Léger, and M. Carpuat. The nrc system for discriminating similar languages. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, pages 139–145, 2014.

Swagger Editor
Supported by SMARTBEAR

File ▾ Edit ▾ Insert ▾ Generate Server ▾ Generate Client ▾

The Assistant Client Advisor 1.0.0 OAS3

The Assistant Client Advisor provides domain prediction.

Servers

/api

default

GET /departments

Returns list of departments.

Parameters

Try it out

No parameters

Responses

Code	Description	Links

Code	Description	Links
200	A list of departments. Media type application/json Controls `Accept` header. **Example Value** | Schema `[ { "id": "314", "name": "ACOA" } ]`	*No links*

GET `/models`

Returns all known prediction models.

Parameters | **Try it out**

No parameters

Responses

Code	Description	Links

Code	Description	Links
200	A list of models. Media type application/json Controls Accept header. Example Value | Schema	*No links*

```
[
  {
    "api": "/api/models/0",
    "name": "2 Phase bilingual solution based on dept-bow1b"
  }
]
```

POST `/models/{modelID}`

Performs a domain prediction of the provided text.

Parameters Try it out

Name	Description
modelID * required `string` *(path)*	Model ID to use for this prediction. modelID - Model ID to use for this prediction.

Request body required application/json

Example Value | Schema

```
{
  "department": "134",
  "text": "This is the text to analyze"
}
```

Responses

Code	Description	Links
200	Prediction result. Media type application/json Controls `Accept` header. **Example Value** | Schema `[ { "scores": { "TAG": 0.75, "CRIM": 0.2 } } ]`	*No links*
400	Bad request.	*No links*
404	A model with the specified ID was not found.	*No links*
5XX	Unexpected error.	*No links*

Schemas

Model {

```
description:
                    Definition of a single prediction model.
api                 string($uri)
                    The complete URL to access this model.
name                string
                    Model's description or name.
}
```

Department ⌄ {

description:	Definition of a single department.
id	string A unique department identifier.
name	string Department's name.

}

PredictionRequest ⌄ {

description:	Prediction request.
text*	string The source text to perform a prediction.
department	string The department identifier.

}

Prediction ⌄ {

description:	The prediction results.
scores	> {...}

}