Comparaison des solutions de NLU sur un corpus français pour un chatbot de support COVID-19

Marion Schaeffer[1] et Christophe Bouvard[1]

[1] Wikit, Lyon, France

marion@wikit.ai et christophe@wikit.ai

Résumé

Les chatbots sont de plus en plus déployés au sein des organisations afin de répondre à des requêtes utilisateur·rice·s en temps réel. Ils utilisent un moteur de compréhension du langage (Natural Language Understanding) en charge d'assimiler les phrases utilisateur·rice·s et ainsi transformer l'information implicite en information explicite interprétable par la machine. De nombreuses offres de ce type existent sur le marché, et il peut être compliqué de choisir une technologie plutôt qu'une autre.

Nous proposons une comparaison de moteurs de NLU disponibles pour la langue française, suivant des critères de performance et de confiance dans un objectif de faciliter l'industrialisation. Nous utilisons deux jeux de données en français sur le cas d'usage du support à destination des employé·e·s : plus de 1000 phrases annotées concernant la COVID-19 que nous rendons disponibles en libre accès et un ensemble de phrases annotées concernant les ressources humaines.

Mots-clés

Comparaison, agent conversationnel, chatbot, traitement automatique du langage naturel (TALN), compréhension du langage naturel, classification d'intentions, reconnaissance d'entités, français, support aux employé·e·s, COVID-19.

Abstract

Chatbots have been increasingly deployed into organizations in order to provide real-time answers to user requests. They leverage a Natural Language Understanding engine to understand sentences and transform implicit information into explicit information for the machine. There are many similar offerings on the market, and choosing one technology over another can be complicated.

We propose a comparison of available NLU engines for the French language, based on criteria like performance and confidence with the aim of facilitating industrialization. We use two French datasets on the employee support use case : more than 1000 annotated sentences about COVID-19 freely available and a set of annotated sentences about human resources.

Keywords

Benchmarking, chatbot, natural language processing (NLP), natural language understanding (NLU), intent classification, entity detection, French, employee support, COVID-19.

1 Introduction

Au cours des dix dernières années, les chatbots sont devenus de plus en plus présents sur les sites web que nous consultons, ainsi que sur les applications que nous utilisons [17]. Ces programmes informatiques ont pour but d'interagir avec un·e utilisateur·rice lors d'une conversation qui se veut naturelle, c'est-à-dire telle que celle que l'on pourrait avoir avec un humain. Au cours d'un échange textuel avec des phrases exprimées en langage naturel, le chatbot doit apporter une réponse à une demande de l'utilisateur·rice [6]. L'assistance et le support utilisateur sont donc des applications privilégiées des chatbots au sein de diverses organisations telles que des entreprises ou des administrations publiques [8].

Pour comprendre la demande de l'utilisateur·rice, les chatbots utilisent un moteur de compréhension du langage naturel (NLU). L'algorithme de compréhension est un composant central du chatbot qui classifie des intentions et reconnaît des entités. Les intentions sont des catégories de sujets que le chatbot peut traiter [21] et les entités sont des mots-clés identifiés pour un traitement spécifique [10]. Les informations structurées issues du moteur NLU sont alors exploitables par le gestionnaire de dialogue du chatbot pour animer l'interaction avec l'humain. De nombreuses solutions de NLU sont disponibles sur le marché, par exemple : Dialogflow de Google (https://cloud.google.com/dialogflow), Wit de Facebook (https://wit.ai), LUIS de Microsoft (https://www.luis.ai), Amazon Lex d'AWS (https://aws.amazon.com/fr/lex/), Watson Assistant d'IBM (https://www.ibm.com/fr-fr/products/watson-assistant), SiriKit d'Apple (https://developer.apple.com/documentation/sirikit), Rasa NLU de Rasa (https://rasa.com/), Snips NLU de Snips (https://github.com/snipsco/snips-nlu), etc. Certains de ces outils sont largement configurables pour fonctionner avec des paramètres personnalisés. D'autres outils se veulent simples d'utilisation et proposent des plateformes clés en main non modifiables [4].

Lors du développement d'un chatbot, le choix de la technologie pour assurer la compréhension du langage n'est pas

une tâche aisée. Il existe différentes études comparatives et les résultats varient en fonction du *dataset*, c'est-à-dire du domaine d'application et des données en elles-mêmes (type de phrases, vocabulaire, périmètre, ...) [3]. De plus, la majorité des comparaisons sont faites en anglais [1, 5]. Pour travailler dans d'autres langues comme le français, les ressources évaluant ce type de solutions sont rares et difficiles à trouver. C'est pour cela que nous présentons une analyse comparative des moteurs de compréhension sur un jeu de données en français. Aussi, nous souhaitons évaluer les performances de ces solutions sur différents critères : la pertinence des résultats et les scores de confiance associés qui sont des critères essentiels pour l'industrialisation du chatbot. Nous avons choisi de réaliser cette comparaison sur un cas d'application concret issu d'une expérience client : un chatbot de support pour les employé·e·s pendant la pandémie de COVID-19. En effet, durant la pandémie, une grande partie de la population s'est retrouvée isolée, engendrant ainsi un sentiment de confusion et de questionnement chez de nombreux·euses salarié·e·s. L'objectif de ce jeu de données est donc de construire un chatbot pour informer les employé·e·s d'une organisation sur les dispositions spécifiques liées à la COVID-19 [2]. Pour compléter l'analyse, nous avons également effectué la comparaison sur un autre corpus traitant un cas d'usage assez proche : le support pour les employé·e·s sur les questions courantes concernant les ressources humaines.

L'article est structuré comme suit : la Section 2 présente l'état de l'art sur les moteurs de compréhension du langage et l'utilisation des chatbots durant la pandémie de COVID-19. Ensuite, la Section 3 aborde les étapes nécessaires à la réalisation de notre étude comparative. Les jeux de données utilisés sur le cas d'usage du support aux employé·e·s sur la thématique des ressources humaines et lors de la pandémie de COVID-19 sont décrits, ainsi que les solutions de NLU choisies pour l'étude. Puis, la Section 4 présente les métriques utilisées et les résultats obtenus par les moteurs de compréhension du langage suivant différents critères. Enfin, la Section 5 conclut notre étude en résumant notre contribution et en citant des perspectives d'évolution pour de futurs travaux.

2 État de l'art

2.1 Comparaison de moteurs de compréhension du langage

Des publications récentes ont comparé les services de NLU dans différents domaines et suivant diverses motivations. Dans [4], les outils des principaux fournisseurs industriels de plateformes NLU sont comparés. Les critères de comparaison sont variés : facilité d'utilisation, langues supportées, entités et intentions pré-construites, intention par défaut, intégration en ligne ou encore le coût financier d'utilisation.

De multiples jeux de données sont utilisés afin d'évaluer les technologies de NLU. En effet, les résultats varient d'un domaine d'application à l'autre. Dans [3], LUIS de Microsoft est le plus performant sur tous les jeux de données. Juste derrière se trouve Rasa, puis IBM Watson et enfin Dialogflow. Un autre article ajoute la technologie Snips au benchmark [5]. Elle se classe à la deuxième place, juste derrière LUIS et devant Rasa avec le même ensemble de données. D'autre part, la technologie d'IBM est la plus performante sur des corpus différents : [1, 20]. Les jeux de données varient de par leur taille (nombre d'intentions et nombre de phrases d'exemples par intention), mais aussi de par leur domaine d'application (vocabulaire et formulations de phrase). Les discussions s'attardent rarement sur l'explication des variations de performances d'un corpus à l'autre, ou sur la justification des scores de confiance.

Le marché étant en constante évolution, la liste des solutions de NLU change d'une étude comparative à l'autre. Nous avons choisi de présenter six des principaux acteurs du marché permettant de travailler en français, ainsi qu'une méthode qui est un point de comparaison (*baseline*) pour la tâche de classification : les Machines à Vecteurs de Support (SVM).

2.1.1 IBM Watson

Watson Assistant est l'agent virtuel intelligent d'IBM. Il s'agit d'une plateforme conversationnelle qui permet la compréhension du langage (NLU) mais aussi la gestion du dialogue (*Dialog Manager*). L'outil supporte plusieurs langues et permet une gestion en ligne ou via une API. Aussi, la partie NLU comprend les intentions et interprète les entités en évaluant la question de l'utilisateur·rice à partir de la base de connaissances disponible. Un score de confiance est ainsi attribué aux prédictions faites.

2.1.2 Rasa

Rasa propose un service de NLU open source. Il permet aux développeur·euse·s de configurer, de déployer et d'exécuter le moteur NLU sur des serveurs en local ou déployés en production. De multiples configurations sont disponibles, avec des paramètres par défaut mais aussi la possibilité d'intégrer des outils tels que spaCy ou BERT par exemple. La solution gère les intentions, les entités et les dialogues avec des scores associés aux résultats. Elle offre une adaptabilité, un contrôle des données et donc des avantages importants pour une solution déployée directement au sein de l'entreprise.

Les configurations incluant des modèles de langue pré-entraînés de type BERT sont particulièrement intéressantes pour bénéficier de la fonctionnalité multilingue [7, 19].

2.1.3 Classification par SVM avec la représentation du langage pré-entraînée CamemBERT

Les machines à vecteurs de support sont des modèles d'apprentissage supervisé utilisés pour des problèmes de discrimination ou de régression. Comme présenté dans [21], les SVM cherchent à définir une frontière de décision entre deux classes en utilisant une fonction noyau. Cette frontière doit avoir une marge maximale dans un espace latent.

Pour classifier les requêtes utilisateur·rice·s, il faut d'abord transformer le texte en vecteur. La méthode de représentation du langage pré-entraînée CamemBERT peut être utilisée à cet escient. CamemBERT [16] est un modèle de langage français basé sur l'architecture RoBERTa [14] et la

partie française du corpus OSCAR [18]. Il permet d'obtenir des vecteurs représentatifs et contextualisés au sens sémantique.

2.1.4 Snips

L'écosystème Snips permet de construire des assistants vocaux à partir d'une console web. Un moteur de compréhension du langage parlé (*Spoken Language Understanding*) est composé d'un moteur de reconnaissance automatique de la parole (*Automatic Speech Recognition*) et d'un moteur de compréhension du langage naturel (NLU). Le modèle peut être entraîné en anglais, français et allemand [5]. Pour la partie classification d'intentions et reconnaissance d'entités, des scores de confiance sont également attribués aux prédictions.

2.1.5 Wit

La plateforme Cloud de NLU wit.ai est détenue et maintenue par Facebook. L'utilisation de la plateforme est gratuite et plusieurs langues y sont disponibles. La solution se concentre uniquement sur l'extraction du sens de l'énoncé d'un·e utilisateur·rice avec la classification d'intentions et la détection d'entités et ne gère aucun type de conversation ou d'intégration via des plateformes de messagerie. L'outil fournit des scores de confiance.

2.1.6 LUIS

Le service Language Understanding, communément appelé LUIS, est l'une des briques de l'offre Microsoft pour construire des systèmes avec Intelligence Artificielle conversationnelle. LUIS utilise des algorithmes d'apprentissage pour analyser les requêtes des utilisateur·rice·s. Les intentions et les entités sont prédites avec des scores de confiance. Plusieurs langues sont disponibles, ainsi qu'un accès via son portail personnalisé, des API et des bibliothèques de développement (dont SDK appelé Bot Framework) pour compléter LUIS lors de l'implémentation de chatbot.

2.1.7 Dialogflow

La plateforme de création d'agents conversationnels Dialogflow de Google permet d'analyser des entrées textuelles ou audio pour en extraire du sens. La compréhension du langage naturel de Dialogflow utilise des modèles fondés sur BERT. De nombreuses langues sont disponibles pour analyser les intentions et les entités avec des scores de confiance ainsi que pour la gestion du dialogue.

2.2 Les langages moins représentés que l'anglais

La plupart des résultats sur les comparaisons de solutions de NLU sont obtenus à partir de corpus écrits en anglais. L'article [22] est un travail pionnier qui aborde ce problème pour la langue italienne. À notre connaissance, il est très difficile d'obtenir des informations similaires pour la langue française. Les difficultés sont multiples : seuls les prestataires disposant d'outils français peuvent être comparés (c'est actuellement le cas pour une majorité). De plus, il est nécessaire de créer un jeu de données spécifique à la tâche et à la langue car il n'existe pas de *datasets* français standardisés comme c'est le cas en anglais, par exemple avec le corpus de dialogue Ubuntu [15, 3].

2.3 L'utilisation des chatbots durant la pandémie de COVID-19

Depuis le début de la pandémie, les recherches sur la COVID-19 se multiplient, notamment en informatique. Les chatbots sont un moyen de centraliser les informations et de les rendre disponibles en continu. C'est pourquoi ils ont été largement utilisés pendant la crise sanitaire, dans différentes langues, sur divers canaux de communication et pour différentes applications [9]. À titre d'illustration, nous pouvons mentionner :

- le suivi des patient·e·s [12], pour collecter simplement des informations sur l'évolution des symptômes prolongés,
- le dépistage des employé·e·s du système de santé [11], pour éviter la propagation nosocomiale de l'infection,
- la recherche d'information [13, 2], pour fournir des réponses vérifiées aux demandes des utilisateur·rice·s et ainsi lutter contre la désinformation.

Les moteurs de compréhension du langage ont donc été largement étudiés sur divers cas d'application mais principalement en langue anglaise. De plus, les chatbots ont montré un réel intérêt durant la crise sanitaire engendrée par la pandémie de COVID-19. La suite de l'article détaille donc notre apport pour la comparaison de solutions de compréhension du langage, en français, sur le cas d'usage de la COVID-19 au sein des entreprises.

3 Méthodologie

3.1 Corpus COVID-19

La comparaison des services de compréhension du langage est basée sur le corpus d'un chatbot en production dédié aux questions concernant la COVID-19 au sein d'une collectivité territoriale française. Les utilisateur·rice·s finaux·ales de cet agent conversationnel sont les employé·e·s de cette organisation, qui ont accès à des services numériques dans le cadre de leur travail. Cet outil d'aide à l'information a été mis en œuvre pour répondre aux thématiques rencontrées lors des premiers confinements.

L'objectif de ce chatbot est de soutenir les employé·e·s lorsqu'ils·elles rencontrent des problèmes liés à la crise sanitaire et de les aider dans leurs interrogations quotidiennes concernant la COVID-19 (tests PCR, gel désinfectant pour les mains, masques, vaccins, cas contact, travail à distance, etc.).

Le jeu de données d'entraînement et de test en français sur le cas d'usage de la COVID-19 utilisé pour évaluer les performances des moteurs de compréhension du langage est disponible à l'adresse suivante : https://github.com/wikit-ai/nlu-french-benchmark.

Les phrases d'entraînement ont été majoritairement exportées du chatbot en production. De ce fait, elles ont été rédigées par les clients et correspondent donc exactement à

leurs besoins. Le corpus a ensuite été complété par des experts afin d'égaliser le nombre de phrases par intention et de garantir les meilleures performances possibles pour chacune des solutions de NLU. Les experts en question travaillent dans le domaine du chabot depuis plusieurs années et leur quotidien consiste à optimiser l'utilisation du chatbot du point de vue fonctionnel et de la compréhension du langage. Au total, 330 phrases ont été annotées pour 22 intentions différentes.

Le corpus de test a été conçu uniquement par les experts de façon manuelle. Pour chaque intention, ces derniers ont rédigé trois phrases différentes par intention en intégrant des entités afin que les différentes classes et mots-clés soient testés dans les mêmes proportions. L'expérience des annotateur·rice·s leur permet de s'approcher au mieux de la façon dont les utilisateur·rice·s discutent et surtout du type de requêtes qui sont généralement utilisées. Les doublons au sein du corpus de test et entre le corpus d'entraînement et de test ont ensuite été retirés. Au final, 913 phrases sont annotées.

3.1.1 Le cas d'usage de la COVID-19

La crise sanitaire de la COVID-19 a bouleversé la vie quotidienne des travailleur·euse·s. Le télétravail a été présenté comme une réorganisation de la vie professionnelle, avec de nouveaux outils et de nouvelles règles. Des processus spécifiques ont été mis en place avec des changements rapides et fréquents en réponse à la propagation du virus. Ainsi, ils·elles ont beaucoup de questions à poser au quotidien [2, 13].

Par exemple, voici quelques énoncés d'utilisateur·rice·s issus du jeu de données d'entraînement du chatbot en production précédemment présenté :

- *J'ai été en contact avec une personne qui a été testée positive au COVID-19.*
- *Quels sont les gestes barrière ?*
- *Je suis très angoissée par cette pandémie.*
- *Peut-on travailler à distance ?*
- *Je dois acheter du gel désinfectant pour les mains.*
- *Quels types de masques peuvent être utilisés au bureau ?*
- *Quelles sont les règles en vigueur à la cantine de l'entreprise ?*

Les réponses à ces entrées utilisateur·trice sont des données informatives pour guider les employé·e·s dans les processus et instructions qui évoluent rapidement.

3.1.2 Les intentions

La classification d'intentions consiste à interpréter la requête de l'utilisateur·rice en fonction des connaissances du système. En effet, il existe une liste exhaustive d'intentions que le moteur NLU est capable de reconnaître. Par conséquent, le problème est de savoir quelle intention existante est la plus proche sémantiquement de la phrase de l'utilisateur·rice [21].

Nous avons conservé 22 intentions différentes qui couvrent un large éventail de questions que les employé·e·s d'une organisation peuvent se poser pendant la pandémie de COVID-19.

Les intentions sont listées dans le tableau 1. Chacune a été entraînée avec 15 phrases d'exemples variées.

Index	Intentions du corpus COVID-19
1	Comment les absences sont-elles gérées ?
2	Comment se passent les repas au restaurant d'entreprise ?
3	Découvrir le champ d'action du bot
4	Que faire en situation de cas contact ou avéré dans l'équipe ?
5	Explique-moi les gestes barrières
6	Quelles sont les dispositions pour que je puisse garder mes enfants ?
7	J'ai une question sur la cellule psychologique
8	J'ai une question sur la prise de repas en salle commune
9	J'ai une question sur le travail à distance
10	J'ai une question sur les campagnes de dépistage
11	J'ai une question sur les campagnes de vaccination
12	J'ai une question sur les cas contacts
13	J'ai une question sur les formations
14	J'ai une question sur les masques
15	J'ai une question sur les symptômes
16	J'ai été testé·e positif·ve à la COVID-19
17	Les réunions en présentiel sont-elles autorisées ?
18	Obtenir du gel hydroalcoolique
19	Obtenir une attestation de déplacement pendant le couvre-feu
20	Est-ce qu'on peut prendre le café entre collègues ?
21	Quel est le dispositif pour les agents vulnérables ?
22	Retrouver tous les modes opératoires

TABLE 1 – Intentions du jeu de données sur la COVID-19

3.1.3 Les entités

La reconnaissance d'entités nommées (*NER*) est également une tâche très importante pour les chatbots. Elle consiste à identifier les portions de texte qui désignent des entités nommées telles que des personnes, des lieux, des noms d'organisations, ... Pour notre cas d'usage, nous étendons cette détection à des entités au sens large afin de pouvoir reconnaître des noms de logiciels ou des mots-clés liés au cas d'usage.

Notre comparaison intègre donc les méthodes d'extraction d'entités sur des entités personnalisées listées dans le tableau 2 afin d'améliorer la classification des intentions. Ces entités peuvent être spécifiées de manière exhaustive avec une liste de synonymes associés ou de termes connexes. En effet, les auteurs de [1] ont démontré que les intentions contenant des mots exclusifs et des entités distinctes étaient plus faciles à identifier par tous les moteurs NLU. Cela peut s'expliquer par le fait que les systèmes de NLU utilisent les types d'entités extraits comme entrée pour la classification

des intentions, mais aussi et surtout parce que certains mots sont associés à des intentions spécifiques.

Entités	Synonymes et termes connexes
absence	congé, RTT, vacances
agent	employé, collègue
attestation	attestation de déplacement, attestation dérogatoire
café	chocolat, chocolat chaud, choco, thé, eau chaude
cellule psychologique	cellule psy, soutien psychologique, soutien psy
compte épargne temps	CET
conjoint	conjointe, mari, femme, époux, épouse
dépistage	PCR, test, test antigénique
enfant	bébé, nourrisson, fils, fille
établissement	collège, crèche, école, école élémentaire, école maternelle, maternelle, lycée
gel	gel hydro, gel hydroalcoolique, hydroalcoolique, spray désinfectant, désinfectant
gestes barrières	règles, règles sanitaires
HDD	département, Hôtel du département
masque	FFP2, masque chirurgical, masque FFP2
mode opératoire	dispositif, dispositif sanitaire, mesure sanitaire, mode opé, protocole, protocole sanitaire
psychologue	psy
rendez-vous	rdv, rendez vous
restaurant administratif	restaurant, resto, resto admin, resto administratif
réunion	réunions, réu, meeting
trad	télétravail, télé travail, télé-travail, travail à distance, travail à domicile
véhicule	voiture

TABLE 2 – Entités du jeu de données sur la covid-19

3.2 Corpus RH

Afin de conforter les tendances dégagées grâce au corpus traitant le cas d'usage de la COVID-19, nous comparons les performances des différentes solutions de NLU sur un autre corpus. Ce corpus regroupe des intentions et des entités relatives aux ressources humaines dans le domaine de l'entreprise.

Le dataset a été conçu de façon incrémentale par les experts de la société Wikit et leurs clients afin de correspondre au mieux à leurs besoins. Seules 15 intentions ont été sélectionnées pour ce test, et chaque intention est entraînée avec 12 phrases d'exemple. La partie test rassemble 8 phrases différentes par intention.

Le corpus couvrent les congés, les périodes d'essai, les formations en entreprise, le comité social et économique de l'entreprise, la mutuelle, les interlocuteurs des ressources humaines, les questions relatives aux salaires, les risques de burnout ou encore le télétravail. Les interrogations autour de ces sujets sont converties en intentions, et les mots-clés sont utilisés comme entités.

3.3 Les plateformes de compréhension du langage

Les algorithmes de NLU visent à extraire des informations utiles et structurées à partir de données non structurées, c'est-à-dire d'une entrée en langage naturel. Nous sélectionnons six services largement utilisés par les chercheurs et les entreprises proposant une analyse en français afin de comparer leurs performances. Ces NLU apparaissent également dans d'autres études à des fins de comparaison mais avec des domaines d'application différents [1, 4, 5, 22, 20].

3.3.1 Interrogation via API

Pour les plateformes disponibles dans le Cloud comme Watson Assistant, LUIS, Dialogflow et Wit, les modèles de compréhension du langage sont entraînés et interrogés grâce à des API. L'entraînement est réalisé en associant chaque phrase d'exemple à l'intention correspondante et en annotant chaque mot-clé avec son entité. Suivant les plateformes, l'entraînement est automatique (Watson Assistant) ou doit être déclenché (Dialogflow, LUIS, Wit). Finalement, la prédiction sur le jeu de test se fait phrase par phrase avec le classement des intentions associées à leurs scores ainsi que les entités trouvées et leur position.

3.3.2 Création de modèles en local

En ce qui concerne Rasa, Snips et la classification par SVM, des modèles sont créés en local et sont testés directement en local également. Tout comme pour les plateformes disponibles dans le Cloud, les phrases d'exemples sont associées à leur intention et il en va de même pour les entités. Le modèle est entraîné (non automatique) puis requêté afin de prédire les intentions et entités d'une nouvelle phrase. Les intentions sont classées et associées à un score. Les entités sont identifiées avec leur position dans la phrase.

4 Résultats

Dans cette section, nous présentons la comparaison des NLU en termes de classification d'intentions et de détection d'entités. Nous entraînons chacun des services de NLU avec les corpus en français, puis nous testons la solution avec les phrases de test annotées.

4.1 Classification d'intentions

Pour évaluer la performance des moteurs de compréhension du langage sur la tâche de classification d'intentions, nous ne considérons que l'intention classée en première position, c'est-à-dire l'intention ayant le score de confiance le plus élevé. En effet, lors d'une conversation réelle avec le chatbot, celui-ci ne fournit qu'une seule réponse en fonction de la meilleure intention trouvée. C'est donc cette intention candidate qui doit être évaluée.

4.1.1 Métriques

Différentes métriques peuvent être utilisées pour évaluer les performances des NLU. Elles fournissent des informations complémentaires sur les performances du système. Afin de comprendre ces métriques, nous définissons pour une classe A fixée :

- les *vrais positifs (TP)* comme étant les valeurs appartenant à la classe A et effectivement prédites comme telles
- les *vrais négatifs (TN)* comme étant les valeurs n'appartenant pas à la classe A et effectivement prédites comme telles
- les *faux positifs (FP)* comme étant les valeurs qui n'appartiennent pas à la classe A mais qui sont prédites comme y appartenant
- les *faux négatifs (FN)* comme étant les valeurs appartenant à la classe A mais qui ne sont pas prédites comme y appartenant.

Exactitude : L'exactitude (ou l'*accuracy*) est la métrique la plus largement utilisée. Elle divise le nombre d'observations correctement prédites par le nombre total d'observations, comme le montre l'équation 1. Cependant, cette mesure peut être insuffisante si l'ensemble de données n'est pas symétrique.

$$Exactitude = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

Précision : Dans le domaine de la recherche d'information, la précision est la fraction des documents retrouvés qui sont pertinents pour la requête. Il s'agit d'une évaluation quantitative de la performance. Ainsi, comme nous pouvons le voir dans l'équation 2, la précision est le rapport entre les observations d'une classe A correctement prédites et le total des observations prédites comme étant de classe A.

$$Précision = \frac{TP}{TP + FP} \tag{2}$$

Rappel : Toujours dans le domaine de la recherche d'information, le rappel (ou *recall*) est la fraction des documents pertinents qui sont retrouvés avec succès. Il s'agit d'une évaluation qualitative de la performance. Ainsi, comme nous pouvons le voir dans l'équation 3, le rappel est le rapport entre les observations d'une classe A correctement prédites et l'ensemble réel des observations de la classe A.

$$Rappel = \frac{TP}{TP + FN} \tag{3}$$

Score F1 : Le score F1 est la moyenne harmonique de la précision et du rappel. Cette métrique reflète au mieux la qualité d'un modèle, en cas de distribution inégale des classes par exemple, car elle prend en compte les faux positifs et les faux négatifs comme le présente l'équation 4.

$$\begin{aligned} \textit{Score F1} &= 2 \times \frac{Rappel \times Précision}{Rappel + Précision} \\ &= \frac{2 \times TP}{2 \times TP + FP + FN} \end{aligned} \tag{4}$$

4.1.2 Classement

Nous comparons les moteurs de compréhension du langage avec les scores F1 obtenus sur le jeu de test COVID-19 et RH. Comme présenté dans le tableau 3 pour le corpus COVID-19, Wit a le meilleur score F1 d'une valeur de $0,806$. La performance de Watson est proche avec un score de $0,794$. Juste après, le modèle de classification par SVM associé à la méthode d'*embedding* CamemBERT et Rasa obtiennent un score très proches de $0,782$ et $0,780$. LUIS et Snips se placent respectivement en cinquième et sixième position avec des scores de $0,756$ et $0,746$. Enfin, Dialogflow se positionne à la fin du classement avec un score de $0,668$.

NLU	Exactitude	Précision	Rappel	Score F1
Watson	**0,800**	0,812	**0,800**	0,794
Rasa	0,783	0,794	0,783	0,780
SVM	0,784	0,795	0,785	0,782
Snips	0,731	0,815	0,729	0,746
Wit	0,752	**0,891**	0,753	**0,806**
LUIS	0,766	0,785	0,766	0,756
Dialogflow	0,617	0,812	0,616	0,668

TABLE 3 – Performance de la classification d'intentions sur le corpus COVID-19

En ce qui concerne le corpus RH, les résultats sont présentés dans le tableau 4. Watson se place largement en tête avec un score de $0,923$. Juste derrière, Rasa obtient un score de $0,902$. Wit, LUIS et Snips se placent respectivement en troisième, quatrième et cinquième position avec des scores de $0,891$, $0,880$ et $0,876$. Dialogflow et le SVM se partage le bas du classement avec des scores de $0,844$ et $0,808$.

NLU	Exactitude	Précision	Rappel	Score F1
Watson	0,9	**0,959**	0,9	**0,923**
Rasa	**0,908**	0,912	**0,908**	0,902
SVM	0,817	0,841	0,817	0,808
Snips	0,875	0,899	0,875	0,876
Wit	0,85	0,961	0,85	0,891
LUIS	0,883	0,899	0,883	0,880
Dialogflow	0,808	0,929	0,808	0,844

TABLE 4 – Performance de la classification d'intentions sur le corpus RH

4.1.3 Score de confiance

Les résultats de la classification d'intentions étant assez similaires d'une technologie à une autre, il est intéressant de comparer les scores de confiance attribués aux prédictions pour chacun des modèles et des corpus.
En effet, un modèle efficace doit avoir des scores de confiance élevés sur des prédictions correctes, avec un écart-type petit, alors que le score de confiance doit être le plus bas possible pour les mauvaises prédictions, avec un écart-type plus élevé.

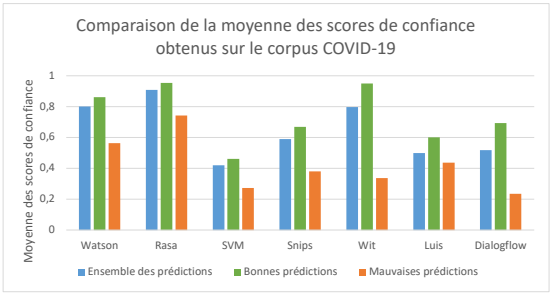

FIGURE 1 – Moyenne des scores de confiance obtenus sur le corpus COVID-19

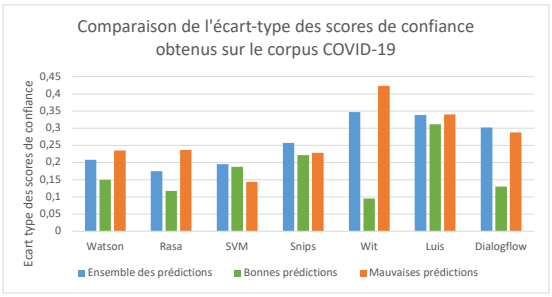

FIGURE 2 – Écart-type des scores de confiance obtenus sur le corpus COVID-19

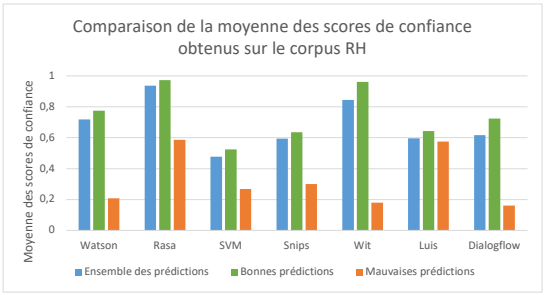

FIGURE 3 – Moyenne des scores de confiance obtenus sur le corpus RH

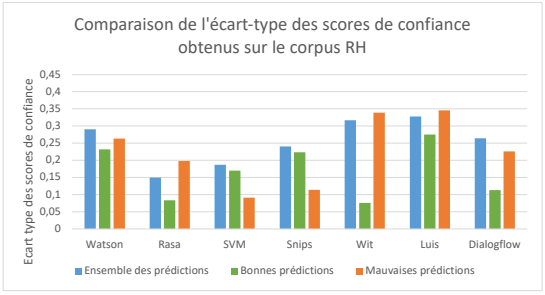

FIGURE 4 – Écart-type des scores de confiance obtenus sur le corpus RH

Comme on peut le voir sur les figures 1 et 2, le comportement attendu est bien présent sur les services Watson Assistant, Rasa, Wit et Dialogflow. Il est toutefois notable que sur Snips, Dialogflow, LUIS et de façon accentuée sur le SVM, les scores obtenus sont plus bas que les scores des autres services. De même, les écart-types des bonnes et des mauvaises prédictions du SVM, de Snips et de LUIS sont très proches en comparaison avec les autres solutions.

Pour le corpus sur le cas d'usage des ressources humaines, les résultats sont présentés sur les figures 3 et 4. Les observations sont très similaires à celles faites sur le corpus COVID-19. On remarque que les scores de confiance de Rasa sont très élevés, tout comme pour Wit. Watson Assistant et Dialogflow ont également des scores assez élevés. Les scores de confiance les plus faibles sont attribués par le SVM, Snips et LUIS.

Concernant les écart-types, le comportement attendu (écart-type plus élevé pour les mauvaises prédictions que pour les prédictions correctes) est bien présent sur Rasa, Wit, LUIS et Dialogflow. Pour Watson, les écart-types des prédictions correctes et mauvaises sont très proches. Finalement, les écart-types des bonnes prédictions sont supérieurs aux écart-types des mauvaises prédictions pour le SVM et Snips.

4.2 Détections d'entités

Dans un second temps, nous comparons les moteurs de compréhension du langage avec les performances obtenues sur la détection d'entités. Pour cela, l'exactitude est mesurée, c'est à dire la proportion d'entités correctement identifiées parmi l'ensemble des entités présentes. Seules les entités présentes dans les phrases du corpus de test sont évaluées.

Comme présenté sur la figure 5 pour le corpus COVID-19, Dialogflow a les meilleures performances avec un score de $0,876$. Watson est juste derrière avec un score de $0,849$. La suite du classement est occupée par Wit avec un score de $0,747$. LUIS, Rasa et finalement Snips se positionnent au bas du classement avec des scores inférieurs.

Le SVM n'a pas été testé sur la tâche de reconnaissance d'entités. Pour implémenter notre propre algorithme de détection d'entités, il est possible, par exemple, d'utiliser les champs aléatoires conditionnels (*conditional random fields*, CRF).

En ce qui concerne le corpus RH, les résultats sont présentés en figure 6. La meilleure performance est réalisée par Dialogflow avec un score de $0,957$. LUIS se positionne juste derrière avec un score de $0,932$. Les autres solutions se suivent avec des scores assez proches, dans l'ordre il y a Rasa, Watson, Snips et un peu plus loin Wit.

5 Discussion et conclusion

Dans la section 4, nous examinons les performances des différents services de NLU sur un corpus traitant le cas d'usage de la COVID-19. Nous croisons également nos résultats avec des tests réalisés grâce au dataset du cas d'usage des ressources humaines.

5.1 Classification d'intentions et détection d'entités

Globalement les performances des services de NLU sont meilleures sur le corpus RH que sur le corpus COVID-19

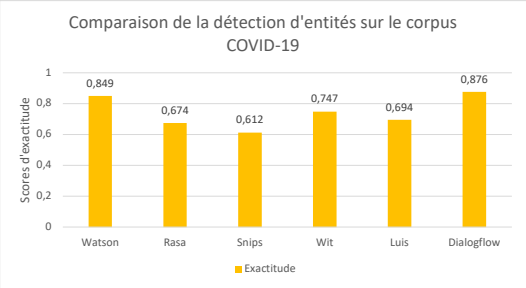

FIGURE 5 – Résultats de la détection d'entités pour le corpus COVID-19

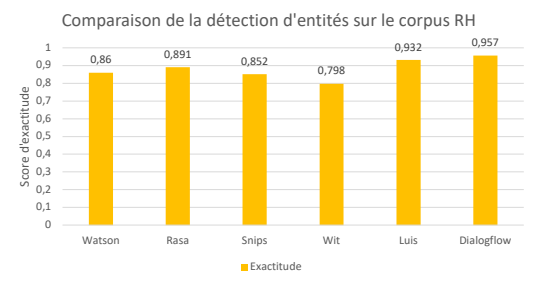

FIGURE 6 – Résultats de la détection d'entités pour le corpus RH

simplement car la quantité de données de test du corpus COVID-19 (plus de 900 phrases) est supérieure à celle du corpus RH (120 phrases).
En ce qui concerne la classification d'intentions, les moteurs de compréhension ont des scores F1 très proches. Watson semble cependant prendre l'avantage, suivi de très près part Wit et Rasa. Sur les deux corpus, Dialogflow, LUIS et Snips se positionnent en fin de classement pour cette tâche.
Lors de cette comparaison, nous utilisons l'annotation d'entités lorsque les plateformes disposent de la fonctionnalité directement (c'est-à-dire toutes les plateformes sauf la méthode de classification par SVM). D'autres études telles que [1] ont montré que la présence d'entités bénéficie à la classification des intentions. Finalement, le SVM reste concurrent aux autres méthodes de classification malgré son désavantage.
En ce qui concerne la détection d'entités, Dialogflow se positionne nettement devant ses compétiteurs.
De plus, Rasa et le classifieur SVM ont un avantage supplémentaire car il est possible de les configurer et de mieux ajuster leurs paramètres pour améliorer les résultats. Par exemple, il serait intéressant d'affiner un modèle de représentation du langage pour mieux s'adapter au vocabulaire spécifique du cas d'usage du support aux employé·e·s durant la pandémie de COVID-19 ou sur de domaine des ressources humaines ; et ainsi mesurer le bénéfice de ces méthodes par rapport à d'autres services plus statiques.

5.2 Synthèse des caractéristiques

Le tableau 5 rassemblent quelques unes des caractéristiques intéressantes à comparer en vue de l'industrialisation d'une solution de compréhension du langage.

5.2.1 Fonctionnalités

Différentes fonctionnalités des solutions ont été comparées dans la section 2. Nous les reprenons ici afin de conclure notre étude.

API. Le fait de disposer d'une API est un atout majeur pour une solution : en effet, elle permet d'accélérer l'industrialisation du moteur de compréhension du langage. Toutes les solutions en proposent une, plus ou moins complète, qui permet à minima de créer un modèle et de l'interroger lors de la prédiction. Les API poussées permettent également de gérer l'entraînement et la mise à jour des modèles.
Le SVM ne dispose pas d'une API initialement car celle ci doit être implémentée. C'est un inconvénient : cela prend du temps, mais c'est aussi un atout : les spécifications peuvent être personnalisées.

Multilingue. Le multilingue est une fonctionnalité très intéressante, elle permet de gérer plusieurs langues dans un seul moteur de compréhension du langage. Les seules techniques qui proposent simplement cette fonctionnalité sont Rasa et le SVM, grâce à des modèles de représentation du langage multilingue comme Bert par exemple.
Pour les autres solutions, elles proposent différentes langues mais ces langues ne peuvent pas être gérées simultanément. En d'autres termes, il est nécessaire d'avoir un moteur de compréhension du langage par langue souhaitée.

Configurabilité du NLU. En ce qui concerne la configurabilité, nous nous intéressons ici à la possibilité d'ajuster les composants du moteur de compréhension du langage. C'est le cas pour Rasa, qui permet de choisir la représentation du langage souhaitée ainsi que les technique de traitement de la langue à utiliser. Il en va de même pour le SVM, avec des paramètres ajustables et une méthode de représentation du langage au choix. Il est également possible d'ajuster les configurations de Snips concernant la partie NLU.
Les autres solutions sont très peu configurables et proposent, au mieux, des options comme le *fuzzy matching*, l'utilisation de méthodes d'apprentissage pour la détection d'entités, etc... Ce critère est finalement très lié à celui de l'open source discuté par la suite.

Hébergement proposé. L'hébergement est une partie importante de l'industrialisation, car elle impacte la mise en production du modèle de compréhension du langage. La majorité des solutions proposent l'hébergement du modèle conjointement au logiciel en tant que service (SaaS) : Watson Assistant, Wit, LUIS, Dialogflow.
En ce qui concerne le SVM, Rasa et Snips, l'hébergement est géré par l'équipe technique en charge du projet de chatbot.

Open source. Un logiciel ou une librairie est dit open source si le code source est libre d'accès, réutilisable et modifiable. Dans cette étude, c'est bien le cas pour Rasa et Snips, dont le code source est téléchargeable et adaptable. Il en est de même pour le SVM dont différentes implémentations sont disponibles ou peut être implémenté directement. Watson Assistant, Wit, LUIS et Dialogflow communiquent

Caractéristiques	Watson	Rasa	SVM	Snips	Wit	LUIS	Dialogflow
API	✓	✓	×	✓	✓	✓	✓
Multilingue	×	✓	✓	×	×	×	×
Configurabilité du NLU	×	✓	✓	✓	×	×	×
Hébergement proposé (SaaS)	✓	×	×	×	✓	✓	✓
Open source	×	✓	✓	✓	×	×	×
Tarification proposée	✓	×	×	×	×	✓	✓
Entraînement automatique	✓	×	×	×	×	×	✓

TABLE 5 – Synthèse des caractéristiques des moteurs de compréhension du langage

très peu sur l'implémentation. Seules les grandes lignes des technologies utilisées sont abordées mais il est compliqué d'obtenir ce genre de renseignements en tant qu'utilisateur·trice.

Tarification proposée. L'aspect financier de la mise en production d'un service de NLU pour le déploiement d'un chatbot est un critère de sélection important pour des entreprises, ou bien même les universitaires lorsqu'ils souhaitent utiliser ce type de technologies. Différentes tarifications existent :

- Rasa (version open source), Snips et un modèle de classification SVM sont des technologies gratuites à utiliser. Cependant, il est nécessaire de déployer une instance de production et donc de payer un hébergement. Le tarif dépend alors de l'hébergeur mais aussi des ressources nécessaires pour faire fonctionner l'instance, ainsi que de l'utilisation faite.
- LUIS, Dialogflow et Watson Assistant proposent le moteur de compréhension du langage hébergé dans des versions d'essais gratuites et avec différents plans de tarification selon le besoin. Le coût dépend alors du nombre de requêtes émises ou des options souscrites.
- Wit.ai est totalement gratuit, et ce même pour une utilisation commerciale. Cela comprend l'entraînement et l'hébergement du modèle.

Pour continuer cette étude dans de futurs travaux, nous souhaitons définir un cadre représentatif de l'utilisation de ce type de chatbot et ainsi comparer le coût financier de son déploiement sur une période et un volume de requête donné.

Entraînement automatique. Le temps d'entraînement (ou d'apprentissage) des moteurs de compréhension du langage est un argument important dans le choix de la technologie utilisée pour développer son chatbot. Par exemple, IBM Watson Assistant et Dialogflow proposent un entraînement qui se fait automatiquement et très rapidement alors que d'autres services comme LUIS ou encore Rasa, nécessitent un déclenchement de l'entraînement. Lorsque l'entraînement est terminé, un remplacement du modèle courant par un nouveau modèle est obligatoire sur Rasa, Snips ou le SVM, la prédiction est donc indisponible durant le temps de chargement du nouveau modèle. Ces contraintes sont intéressantes à étudier car leur impact est important pour l'industrialisation d'une solution de ce type et seront donc ajoutées à l'étude dans de futurs travaux.

De la même façon, le temps de prédiction des intentions et des entités est important pour l'industrialisation du chatbot. En effet, un des principaux atouts de ce dernier est le fait de converser en temps réel. Lors de notre utilisation, toutes les solutions ont permis cela. Cependant, étudier la robustesse des plateformes lorsque de nombreuses demandes sont formulées pour des questions d'industrialisation serait une démarche intéressante et complémentaire à notre étude.

5.3 Conclusion

Les chatbots sont de plus en plus populaires et, par conséquent, les services de NLU sont largement utilisés. Ces derniers constituent une pièce centrale du chatbot, car ils permettent d'interpréter les demandes utilisateur et donc de fournir les réponses les mieux adaptées. Le choix de la technologie pour un NLU est une tâche complexe. Les services ont des caractéristiques différentes et de nombreuses études comparent leurs performances. Cependant les résultats sont souvent différents selon le jeu de données et il y a très peu de ressources disponibles pour comparer les moteurs en langue française.

Avec la pandémie, les chatbots ont été largement utilisés pour répondre aux questions des utilisateur·rice·s sur les changements induits par la situation sanitaire. Dans ce contexte, nous avons créé, annoté et mis à disposition un jeu de données d'entraînement et de test sur le cas d'usage de la COVID-19 en entreprise. Nous avons entraîné et testé différents services de traitement du langage sur la tâche de classification d'intentions et de détection d'entités afin d'identifier l'outil le plus efficace pour le français.

Nous avons constaté que Watson, Wit et Rasa obtiennent les meilleures performances sur la classification d'intentions alors que Dialogflow est le plus performant pour la détection d'entités sur nos jeux de données en français. Cependant, les différents services ont des résultats assez similaires lorsqu'ils sont comparés dans leur globalité. Il est donc intéressant de s'intéresser aux avantages de chacun des moteurs en fonction de l'usage qu'il est prévu d'en faire. Dans un travail futur, nous comparerons les services de NLU sur des aspects complémentaires tels que le temps d'entraînement, le coût financier ou encore la robustesse lors du déploiement, afin d'avoir un classement objectif basé sur autant de critères que possible, pour faciliter l'industrialisation des chatbots.

Références

[1] Ahmad Abdellatif, Khaled Badran, Diego Costa, and Emad Shihab. A comparison of natural language understanding platforms for chatbots in software engineering. *IEEE Transactions on Software Engineering*, PP :1–1, 05 2021.

[2] Eslam Amer, Ahmed Hazem, Omar Farouk, Albert Louca, Youssef Mohamed, and Michel Ashraf. A proposed chatbot framework for covid-19. pages 263–268, 05 2021.

[3] Daniel Braun, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185, Saarbrücken, Germany, August 2017. Association for Computational Linguistics.

[4] Massimo Canonico and Luigi De Russis. A comparison and critique of natural language understanding tools. 2018.

[5] Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. Snips voice platform : an embedded spoken language understanding system for private-by-design voice interfaces. *ArXiv*, abs/1805.10190, 2018.

[6] Menal Dahiya. A tool of conversation : Chatbot. *International Journal of Computer Sciences and Engineering*, 5 :158–161, 05 2017.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[8] Dario Fiore, Matthias Baldauf, and Christian Thiel. "forgot your password again ?" : acceptance and user experience of a chatbot for in-company it support. *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*, 2019.

[9] Sviatlana Höhn and Kerstin Bongard-Blanchy. Heuristic evaluation of covid-19 chatbots. pages 131–144, 02 2021.

[10] Anran Jiao. An intelligent chatbot system based on entity extraction using rasa nlu and neural network. *Journal of Physics : Conference Series*, 1487 :012014, 03 2020.

[11] Timothy Judson, Anobel Odisho, Jerry Young, Olivia Bigazzi, David Steuer, Ralph Gonzales, and Aaron Neinstein. Case report : Implementation of a digital chatbot to screen health system employees during the covid-19 pandemic. *Journal of the American Medical Informatics Association : JAMIA*, 27, 06 2020.

[12] Hannah Lei, Weiqi Lu, Alan Ji, Emmett Bertram, Paul Gao, Xiaoqian Jiang, and Arko Barman. COVID-19 smart chatbot prototype for patient monitoring. *CoRR*, abs/2103.06816, 2021.

[13] Yunyao Li, Tyrone Grandison, Patricia Silveyra, Ali Douraghy, Xinyu Guan, Thomas Kieselbach, Chengkai Li, and Haiqi Zhang. Jennifer for covid-19 : An nlp-powered chatbot built for the people and by the people to combat misinformation. In *NLPCOVID19*, 2020.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta : A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[15] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The ubuntu dialogue corpus : A large dataset for research in unstructured multi-turn dialogue systems. *CoRR*, abs/1506.08909, 2015.

[16] Louis Martin, Benjamin Müller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert : a tasty french language model. *CoRR*, abs/1911.03894, 2019.

[17] Quim Motger, Xavier Franch, and Jordi Marco. Conversational agents in software engineering : Survey, taxonomy and challenges. *CoRR*, abs/2106.10901, 2021.

[18] Pedro Ortiz Suarez, Benoît Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. 07 2019.

[19] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert ? *CoRR*, abs/1906.01502, 2019.

[20] Haode Qi, Lin Pan, Atin Sood, Abhishek Shah, Ladislav Kunc, and Saloni Potdar. Benchmarking intent detection for task-oriented dialog systems. *CoRR*, abs/2012.03929, 2020.

[21] Jetze Schuurmans and Flavius Frasincar. Intent classification for dialogue utterances. *IEEE Intelligent Systems*, PP :1–1, 11 2019.

[22] Matteo Zubani, Luca Sigalini, Ivan Serina, and Alfonso Gerevini. Evaluating different natural language understanding services in a real business case for the italian language. In *KES*, 2020.