Système de question-réponse multilingue appliqué aux agents conversationnels

Wissam Siblini, Charlotte Pasqual
Axel Lavielle, Cyril Cauchois

Worldline, France
{wissam.siblini,charlotte.pasqual,cyril.cauchois}@worldline.com, lavielleaxel@gmail.com

Résumé. Les modèles de langages (e.g. BERT) permettent de résoudre avec brio des tâches de TALN complexes comme le question-réponse. Cependant, les jeux de données spécifiques à ces tâches sont principalement en anglais, ce qui rend difficilement compte des progrès dans les autres langues. Heureusement, les modèles commencent à être pré-entraînés dans des centaines de langues et ont une bonne capacité de transfert zero-shot d'une langue à l'autre. Dans cet article, nous montrons notamment que BERT multilingue, entraîné pour la tâche de question-réponse en anglais, est capable de généraliser au français et au japonais. Nous présentons alors une application pratique Kate, agent conversationnel dédié au support ressources humaines, qui répond aux questions posées par des utilisateurs dans plusieurs langues à partir de contenus de pages d'intranet.

1 Introduction

Depuis quelques années, on assiste à une révolution du domaine du Traitement Automatique du Langage Naturel (TALN) (Vaswani et al., 2017; Devlin et al., 2018), probablement motivée par des compétitions comme SQuAD (Rajpurkar et al., 2016) ou GLUE (Wang et al., 2018). Sur SQuAD, les modèles de langage tels que BERT (Devlin et al., 2018), XLNet (Yang et al., 2019) ; ont montré qu'ils pouvaient identifier, si elle existe, la réponse à une question dans une source donnée. Cette fonctionnalité est intéressante pour augmenter les agents conversationnels. En effet, ceux-ci sont souvent limités : ils identifient des intentions prédéfinies et fournissent des réponses scriptées, mais ne peuvent pas répondre à une requête inattendue. Pourtant les sociétés disposent d'un grand nombre de sources d'informations dans lesquels un modèle automatique de compréhension de texte serait susceptible d'identifier l'information recherchée. Malheureusement, la majorité des jeux de données (e.g. SQuAD) pour entraîner un tel modèle sont exclusivement en anglais. Comment résoudre cette tâche dans d'autres langages ? Recréer des ensembles de données étiquetées dans toutes les langues ciblées serait une solution peu flexible qui demanderait beaucoup de ressources. Une autre direction possible est le transfert zero-shot d'un modèle entraîné pour la tâche en anglais vers la langue cible (Loginova et al., 2018). Par exemple, les modèles de langage entraînés en multilingue semblent intégrer naturellement un alignement linguistique dans leur représentation des phrases, et obtiennent alors des performances étonnantes en transfert zero-shot. Dans cet article, nous démontrons cette

capacité pour le modèle BERT multilingue pour résoudre spécifiquement la tâche de question-réponse et la transférer de l'anglais vers d'autres langues (français et japonais) sans données étiquetées. Nous introduisons six nouveaux ensembles de données de question-réponse cross-lingues (question et paragraphe source dans une langue différente) afin de mieux comprendre et évaluer le transfert. Nous présentons enfin un cas d'application pratique : Kate, une assistante virtuelle pour le support des ressources humaines de l'entreprise.

2 Modèles de traitement automatique du langage naturel

Le TALN constitue un ensemble d'applications phares de l'apprentissage automatique. La classification de texte a été l'une des premières où les résultats ont été très prometteurs avec des stratégies comme la combinaison de TF-IDF et SVM (Manevitz et Yousef, 2001). Au départ, l'entraînement d'un modèle se limitait aux jeux de données spécifiquement construits pour la tâche ciblée. Par la suite, les chercheurs ont commencé à apprendre des représentations de mots sur des sources de textes externes non labelisées, pour intégrer et transférer des relations sémantiques. L'exemple le plus emblématique est *word2vec* (Mikolov et al., 2013), simple réseau de neurones à une couche cachée, entraîné avec de nombreuses astuces, pour prédire un mot dans une phrase à partir des mots qui l'entourent et inversement. Les poids du réseau permettent de dériver des représentations des mots, incorporant la sémantique du langage, qui améliorent significativement de nombreuses tâches de TALN notamment sur des ensembles de données de taille limitée. Cette stratégie a été largement adoptée et a évolué vers des représentations plus élaborées avec des réseaux de neurones récurrents (Le et Mikolov, 2014), voire des représentations contextualisées pour la désambiguïsation (Peters et al., 2018).

Les dernières propositions s'orientent vers l'uniformisation des progrès grâce à un format facilement utilisable par tous : les modèles de langage. Dans ces modèles profonds, la représentation de texte est intégrée au sein des premières couches, et les couches de sortie prévoient la résolution de plusieurs tâches (Devlin et al., 2018; Howard et Ruder, 2018). Ils prennent en entrée des textes bruts représentés comme des séquence de "tokens" (caractères ou séries de caractères) souvent généré avec Wordpiece Tokenization (Schuster et Nakajima, 2012). Les premières couches sont pré-entraînées sur des données textuelles non étiquetées pour produire les représentations contextualisées des tokens. Ensuite, le modèle est affiné sur des données spécifiques à la tâche cible en adaptant les couches de sortie.

Les modèles de langage récents s'appuient souvent sur l'architecture *Transformer* (Vaswani et al., 2017), un réseau de neurones avec plusieurs couches d'attention et de self-attention, pour pouvoir résoudre des tâches complexes de NLP en exploitant des interdépendances complexes sur des longues séquences de texte.

3 BERT et son extension multilingue

Dans la suite, nous nous concentrons sur BERT, modèle de langage récemment publié dans une version très multilingue. BERT apprend des représentations de tokens conditionnées conjointement par leur contexte gauche et leur contexte droit à l'aide de *Transformers* bidirectionnels. Il est conçu pour prendre en entrée une paire de phrases séparées par un token spécial. Il dispose d'un large éventail de sorties permettant de l'adapter pour plusieurs tâches

(classification, génération de texte, question-réponse) sans modification importante de son architecture.

Le pré-entraînement optimise les poids du modèle sur des documents non étiquetés issus de Wikipedia, dans les cent langues avec le plus grand nombre d'articles [1]. Un échantillonnage des articles et une pondération des tokens, basés sur la fréquence de la langue, permettent un certain équilibrage. Le pré-entraînement correspond à l'apprentissage de deux tâches. Dans la première, environ 15% des tokens, échantillonnés aléatoirement, sont masqués (en les remplaçant par un token spécial, par un token aléatoire, ou par le token précédent) et l'objectif est alors de les retrouver. Dans la deuxième, le but est de prédire la phase suivante/précédente, sur la base de la phrase courante. Le modèle n'est intentionnellement pas informé de la langue de l'échantillon, de sorte que les représentations des tokens apprises ne soient pas explicitement spécifiques à une langue.

Le pré-entraînement de BERT est très coûteux : quatre jours sur quatre à seize TPUs. Heureusement, les poids pré-entraînés ont été mis en ligne par les auteurs [2]. Au contraire, leur affinement sur des tâches spécifiques est relativement rapide. Par exemple, l'affinement sur SQuAD (plus de cent mille échantillons) prend au maximum deux heures sur un GPU classique (Tesla V100). BERT a été publié en deux versions (base et large). La différence réside dans la dimension du *Transformer* (dimension des couches cachées, nombre de blocs et nombre de têtes de self-attention). La version large est légèrement plus précise mais seule la version de base tient dans un GPU avec 12 Go de RAM.

4 Expériences en question-réponse multilingue

Nous basons notre analyse empirique sur le jeu de question-réponse SQuAD : étant donnée une paire question-paragraphe, le but est de déterminer l'emplacement de la réponse dans le paragraphe. Nous commençons par affiner BERT sur le jeu SQuAD anglais (les hyperparamètres et options utilisés sont ceux spécifiés dans le dépôt github officiel de BERT). Nous évaluons ensuite sa capacité à résoudre la même tâche dans d'autres langues (jeux tests SQuAD français et japonais), puis en cross-lingue.

Expérience 1 : transfert zero-shot vers le français et le japonais Les jeux test considérés ici sont des échantillons du jeu test SQuAD v1.1 (premier paragraphe de 48 pages de Wikipedia) traduits par des humains en français et en japonais [3].

Nous comparons les performances de BERT multilingue à une baseline (Asai et al., 2018) dont les résultats sont les meilleurs publiés à ce jour sur les jeux considérés. La baseline combine explicitement deux modèles : un modèle de compréhension de texte (RC) dans une langue pivot (anglais) et un modèle de traduction automatique (MT) attentif entre la langue pivot et la langue cible (japonais ou français). Le modèle MT traduit la paire paragraphe-question vers la langue pivot, puis le modèle RC extrait la réponse dans la langue pivot et, finalement, l'algorithme récupère la réponse dans la langue originale en utilisant les scores d'attention de MT. Le tableau 1 présente la correspondance exacte (EM) et le score F1 des réponses de BERT

1. https://meta.wikimedia.org/wiki/List_of_Wikipedias
2. https://github.com/google-research/bert
3. https://github.com/AkariAsai/extractive_rc_by_runtime_mt

multilingue, qui surpassent significativement ceux de la baseline pour le japonais et le français. De plus, par rapport à celle-ci, BERT a l'avantage supplémentaire d'être plus facilement transférable à encore d'autres langues.

TAB. 1: Comparaison du EM/F1 de BERT multilingue et de celui la baseline sur SQuAD français et japonais. F1 et EM sont les deux métriques officielles de la compétition SQuAD. EM mesure le pourcentage d'emplacements de réponses prédits qui correspondent exactement aux emplacements des vraies réponses. F1 est plus souple et mesure le chevauchement moyen entre les deux emplacements.

	Français		Japonais	
	F1	EM	F1	EM
Baseline	61.88	40.67	52.19	37.00
BERT multilingue	**76.65**	**61.77**	**61.83**	**59.94**

Expérience 2 : Question-réponse cross-lingue Pour effectuer des tests cross-lingues, nous construisons six jeux de données supplémentaires [4] à partir des jeux existants en mélangeant le paragraphe dans une langue avec la question dans une autre. La performance de BERT sur ces jeux est donnée dans le tableau 2. Puisque le modèle a été entraîné pour la tâche en anglais, la performance est la meilleure pour le jeu En-En. Les performances sur Fr-Fr et Jap-Jap sont également très bonnes comme indiqué dans la première expérience. Les résultats sur les jeux cross-lingues sont proches des résultats monolingues et dans certains cas commme En-Fr, la performance est même supérieure (par rapport à Fr-Fr). Nous observons qu'en général, la correspondance exacte et le score F1 ont des valeurs proches lorsque le paragraphe est en japonais alors qu'il y a un écart plus important pour les deux autres langues. C'est parce qu'en japonais, les tokens représentent des plus grandes parties de mots, et il y a donc moins de possibilité de chevauchement partiel entre les prédictions de BERT et la vérité terrain.

TAB. 2: Correspondance exacte et score F1 de BERT multilingue sur chacun des 9 ensembles de données monolingues ou cross-lingues. Chaque ligne (resp. colonne) correspond à une langue différente pour le paragraphe (resp. la question). Les chiffres en gras correspondent au meilleur EM pour chaque langue.

	Question	En		Fr		Jap	
		F1	EM	F1	EM	F1	EM
Paragraphe	En	90.57	**81.96**	78.55	**67.28**	66.22	52.91
	Fr	81.10	65.14	76.65	61.77	60.28	42.20
	Jap	58.95	57.49	47.19	45.26	61.83	**59.93**

4. https://github.com/wissam-sib/multilingualQA

Discussion Les performances impressionnantes de BERT en transfert zero-shot sont potentiellement obtenues grâce à son apprentissage de quatre concepts : (1) la structure sémantique de chaque langue, (2) l'alignement entre les espaces de représentation des différentes langues (au moins pour les plus fréquentes (Pires et al., 2019)), (3) la tâche de question-réponse en anglais et (4) les concepts intrinsèques du question-réponse s'appliquant à toutes les langues.

Les hypothèses (1) et (3) ont déjà été très largement validées par le passé. Nous nous penchons alors sur les hypothèses (2) et (4) qui sont intuitivement recevables. D'une part, de nombreux mots, incluant les noms propres, sont au moins partiellement les mêmes dans les Wikipedia de la plupart des langues et ont pu jouer le rôle de points de repère pour que le pré-apprentissage aligne les espaces de représentation des différentes langues. D'autre part, certaines caractéristiques du question-réponse sont indépendantes de la langue, comme l'identification de passages du paragraphe où une partie du vocabulaire correspond à la question (ou une paraphrase).

Les deux hypothèses sont également appuyées par les bons résultats empiriques. Dans les expériences cross-lingues, le fait que les couches d'attention reliant le paragraphe dans une langue et la question dans une autre langue aient permis malgré tout de focaliser sur l'emplacement de la réponse suggère des représentations alignées. Notons alors que, puisque la performance sur Jap-En est inférieure à celle sur Jap-Jap alors que la performance sur En-Fr est supérieure à celle sur Fr-Fr, l'alignement En-Fr semble être plus important que l'alignement En-Jap. Cela est sans doute dû à un nombre moins élevé de mots ayant des racines communes entre l'anglais et le japonais. De plus, le fait que les résultats sur Jap-Jap soient meilleurs que sur Jap-En suggère que la tâche En-En SQuAD apprise par BERT a été transférée à Jap-Jap par d'autres mécanismes que seul l'alignement de la langue, possiblement par les caractéristiques inhérentes de la tâche de question-réponse invariantes au langage.

5 KATE, le chatbot multilingue basé sur des contenus web

Les résultats empiriques étant très convaincants, nous avons intégré BERT dans notre assistant virtuel de support employé, KATE ("Knowledge ShAring ExperT for Employees"). Dans cette section, nous développons le cas d'usage, le processus de conversation et la manière dont BERT est intégré à l'architecture. Nous présentons également des exemples concrets des résultats impressionnants obtenus par BERT en matière de réponse aux questions.

Notre application est un assistant conversationnel qui répond à des interrogations de type ressources humaines. Nous avons montré, suite à une enquête interne, la difficulté de trouver une information dans des sources parfois mal structurées et insuffisamment indexées, conduisant souvent à un évitement ou un abandon. Nous avons donc décidé de créer un chatbot disposant d'une interface web permettant des interactions de messagerie texte standard, et connectée à un framework de conversation développé en interne. Nous avons ajouté deux boutons, "like" et "dislike", sur le dernier message bot (figure 1), pour obtenir un retour direct et explicite de l'utilisateur sur la pertinence de la dernière réponse donnée.

Dans le système de conversation de Kate (figure 2), la plate-forme de messagerie envoie le dernier texte entré par l'utilisateur à un classifieur d'intention, à savoir Google Dialogflow pour notre expérience, mais notre infrastructure peut également gérer d'autres outils pour cette tâche (Rasa, Snips, Luis.ai ou même BERT entraîné pour la classification). Le dialogue est géré par notre framework interne. Si le moteur de compréhension reconnaît l'intention de l'utilisa-

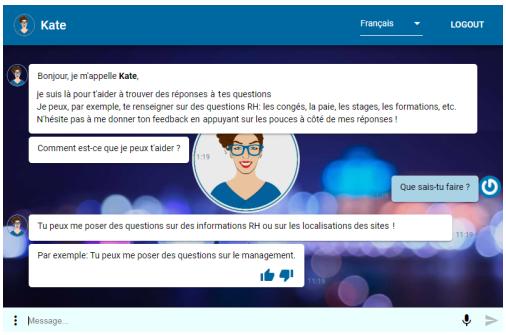

FIG. 1: Interface de Kate

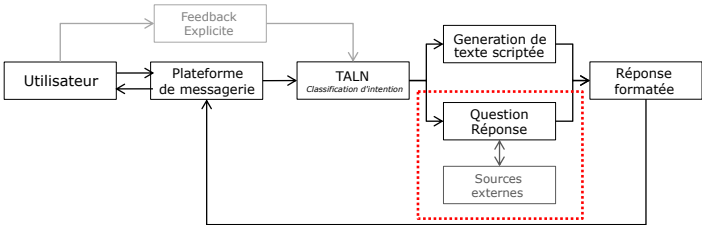

FIG. 2: Processus conversationnel de Kate

teur (e.g. "question sur les congés", "question sur la localisation d'un site"), nous générons une réponse scriptée, en sélectionnant au hasard dans une liste de réponses pré-écrites pour l'intention courante et complétées avec les entités adéquates. Lorsque le moteur de compréhension échoue, probablement parce que la demande est inattendue, ou qu'un retour négatif est envoyé par l'utilisateur, la fonctionnalité de question-réponse avec BERT est appelée.

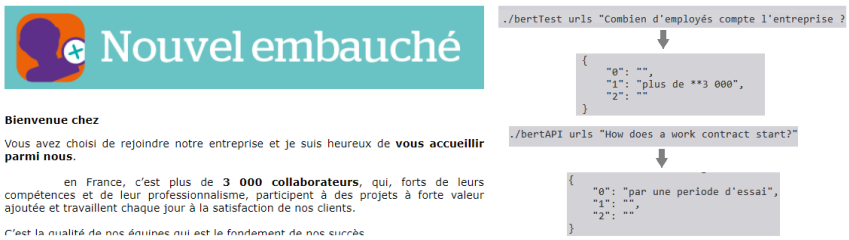

FIG. 3: Exemples de réponses de notre API (à droite). Ici, les sources d'informations utilisées par BERT sont trois pages web dont une est montrée à gauche.

Nous nous appuyons alors sur une base de connaissances externe, constituée d'une liste d'URLs correspondant aux références pouvant contenir les informations recherchées - dans notre cas, il s'agit de plusieurs pages Web de l'intranet d'entreprise comprenant des explications et des règles en matière de ressources humaines. Nous appliquons au code HTML de chaque page Web une fonction de filtrage qui supprime les balises et ne conserve que le

contenu textuel. Nous fournissons la question de l'utilisateur avec chacun des textes à une instance de BERT multilingue s'exécutant sur un serveur pour fournir la réponse. Finalement, le meilleur résultat est mis en forme et affiché par l'interface de messagerie. Dans la figure 3, nous donnons deux exemples de réponses de notre API de question-réponse, basée sur BERT multilingue, avec trois URLs comme sources d'information. La première question "Combien d'employés dans l'entreprise", posée en français, obtient une réponse "*plus de 3000*" située dans la deuxième page Web dans un passage qui n'a aucun mot en commun avec la question ("*en France, c'est plus de 3000 collaborateurs*"). La deuxième question en anglais aboutit quand même à la réponse sur la source en français (*"par une période d'essai"*).

6 Conclusion

Les résultats de BERT en matière de question-réponse dans plusieurs langues sont impressionnants, et confirment ses aptitudes pour les tâches complexes et le transfert zero-shot Siblini et al. (2019); Hsu et al. (2019). De façon pratique, cela permet d'augmenter les chatbots en les rendant capable gérer de nouvelles questions des utilisateurs à partir de contenus web externes. Les réponses sont très bonnes lorsqu'un texte contient une expression proche de la réponse attendue, même lorsque la question et le paragraphe n'ont aucun mot en commun.

L'inclusion d'un modèle de question-réponse dans un moteur de conversation présente d'énormes avantages en termes de maintenance. Les développeurs, linguistes et intégrateurs du chatbot n'ont plus à définir manuellement les intentions/corpus/réponses et se concentrent sur la mise à jour de la base de documents et du modèle linguistique de compréhension. En outre, l'agent conversationnel pourra traiter des questions que ses concepteurs n'auraient pas forcement anticipé. Les compétences cross-lingue de BERT apportent un avantage supplémentaire pour les entreprises internationales disposant de documents dans différentes langues.

Parmi les pistes d'amélioration, nous avons remarqué que BERT ne renvoie qu'une réponse partielle lorsque l'information est dispersée dans plusieurs parties de la source (sous-titres, listes à puces) nous cherchons donc comment agréger plusieurs réponses candidates. Ensuite, l'alignement linguistique de BERT est limité pour les langues rares (Pires et al., 2019) et pourrait bénéficier d'un entraînement plus fin avec des corpus explicitement alignés. Nous prévoyons enfin d'analyser des alternatives de BERT améliorant la rapidité de l'inférence et la consommation mémoire actuelle pour gérer des bases de connaissances plus vastes.

Références

Asai, A., A. Eriguchi, K. Hashimoto, et Y. Tsuruoka (2018). Multilingual extractive reading comprehension by runtime machine translation. *arXiv preprint arXiv :1809.03275*.

Devlin, J., M.-W. Chang, K. Lee, et K. Toutanova (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

Howard, J. et S. Ruder (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv :1801.06146*.

Hsu, T.-Y., C.-l. Liu, et H.-y. Lee (2019). Zero-shot reading comprehension by cross-lingual transfer learning with multi-lingual language representation model. *arXiv preprint arXiv :1909.09587*.

Le, Q. et T. Mikolov (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196.

Loginova, E., S. Varanasi, et G. Neumann (2018). Towards multilingual neural question answering. In *European Conference on Advances in Databases and Information Systems*, pp. 274–285. Springer.

Manevitz, L. M. et M. Yousef (2001). One-class svms for document classification. *Journal of machine Learning research 2*(Dec), 139–154.

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

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, et L. Zettlemoyer (2018). Deep contextualized word representations. *arXiv preprint arXiv :1802.05365*.

Pires, T., E. Schlinger, et D. Garrette (2019). How multilingual is multilingual bert ? In *ACL*.

Rajpurkar, P., J. Zhang, K. Lopyrev, et P. Liang (2016). Squad : 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv :1606.05250*.

Schuster, M. et K. Nakajima (2012). Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5149–5152.

Siblini, W., C. Pasqual, A. Lavielle, et C. Cauchois (2019). Multilingual question answering from formatted text applied to conversational agents. *arXiv preprint arXiv :1910.04659*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, et I. Polosukhin (2017). Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008.

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, et S. R. Bowman (2018). Glue : A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv :1804.07461*.

Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, et Q. V. Le (2019). Xlnet : Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv :1906.08237*.

Summary

Language models such as BERT are a great way to solve complex NLP tasks like Question-Answering. However, datasets are currently mostly in English which makes it difficult to acknowledge progress in other languages. Fortunately, BERT has recently been pre-trained in several hundred languages and show a good ability for zero-shot transfer from one language to another. In this paper, we show that multilingual BERT, trained to solve the question-answering task in English, is then able to generalize to French and Japanese. We also introduce our use-case: Kate, a human resources chatbot, that answers questions from users in multiple languages from intranet pages.