Proposition et développement d'un modèle de système ou d'agent cognitif intelligent

Mounir HAMANI, Djamila Boulahrouz, Karima ADEL-AISSANOU

Doctoriales de Recherche Opérationnelle, le 12 et 13 Décembre 2018

Introduction

La création d'un assistant virtuel intelligent nécessite la combinaison de plusieurs domaines de connaissances, dont le traitement automatique du langage naturel (TALN), la recherche d'information (RI) et les bases de données.

L'extraction de l'information a commencé par l'extraction et le classement des documents qui contenaient des mots-clés à partir d'une requête de l'utilisateur. Aujourd'hui, elle est plus affinée et vise à récupérer l'information exacte que l'utilisateur recherche, afin de lui permettre de gagner encore plus de temps. Le mécanisme le plus courant pour interagir avec un utilisateur s'appelle un système de réponse aux questions ou encore système de Questions/Réponses (QR). Ici, l'utilisateur pose une question, généralement en langage naturel, et le système, après l'avoir traitée et avoir parcouru sa base de connaissances, essaie de donner la réponse la plus précise s'il en a une. Les connaissances peuvent être structurées, semi-structurées ou non structurées (texte libre) selon le système et son application.

Avec l'application du TALN dans le domaine de la RI, les chercheurs ont commencé par essayer des techniques fondées sur des règles pour comprendre la sémantique des questions. Il s'agissait de règles grammaticales codées à la main qui n'étaient malheureusement pas assez robustes en raison des variations linguistiques. Après cela, avec la montée de l'apprentissage machine, l'intérêt s'est déplacé vers l'utilisation de l'inférence statistique pour apprendre automatiquement de telles règles à partir de grands corpus d'exemples du monde réel (par exemple des sources générales comme le WSJ ou WikiPedia, et des domaines spécifiques comme les publications médicales ou scientifiques).

Le succès récent des réseaux neuronaux (RN) et leurs résultats records dans des domaines comme la traduction automatique et la vision par ordinateur ont incité de nombreux chercheurs à explorer la possibilité de les adapter et de les appliquer à leurs domaines respectifs. Cette situation, conjuguée à la disponibilité de corpus de textes annotés de plus en plus volumineux, a donné naissance à la troisième vague de recherche sur la RI.

Depuis 1999, la Text REtrieval Conference (TREC) a organisé une piste d'évaluation annuelle pour les systèmes de réponse aux questions et fournit depuis lors des ensembles de données librement accessibles. Plus récemment, des ensembles de données plus récents ont été rassemblés et l'un des plus largement utilisés est l'ensemble de données Stanford SQuAD qui fournit des questions, des paragraphes (contexte) ainsi que des réponses extraites de ces paragraphes.

Dans le contexte des systèmes de questions-réponses, nous identifions de multiples types de questions, parmi lesquelles des questions dites factoides, des questions définitionnelles, des questions sur le pourquoi et le comment, des questions de liste et, plus récemment, des questions basées sur des scénarios.

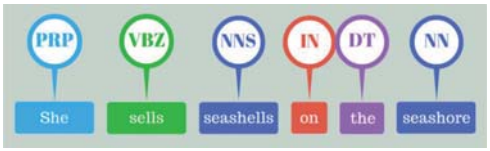
Dans notre recherche, nous nous concentrerons d'abord sur les questions de type factoide, la raison en étant que l'intérêt suscité par ce domaine a favorisé la création d'ensembles de données standards riches et de méthodes d'évaluation relativement peu controversées qui nous permettront de comparer nos résultats.

Les types de RN les plus utilisés dans le TALN sont les réseaux neuronaux récurrents (RNN) et les réseaux neuronaux convolutionnels (CNN). Dans la RI, l'approche dominante est le RNN, cependant, les RNN simples souffrent de ce qu'on appelle le problème de disparition du gradient où plus la séquence, ici de texte, est longue, plus il oublie son début. Pour y remédier, diverses solutions ont été proposées, telles que la mémoire à long terme et à court terme (LSTM) et le Gated Recurrent Unit (GRU) qui sont des mécanismes qui tentent de mieux mémoriser l'information à travers de longues séquences en utilisant des mécanismes de portes qui décident ce qui doit être mémorisé et ce qui doit être oublié.

Approche

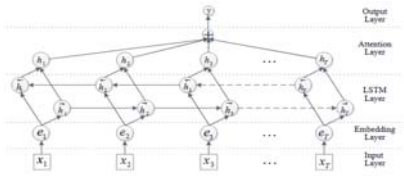
Notre approche repose sur deux principes. Le premier étant de ne rejeter ou ignorer aucune des méthodes précédemment essayées dans le domaine. En effet, chacune d'entre elles a ses points forts et ses points faibles. L'idée étant de les combiner de manière intelligente et adaptée afin de tirer profit de leur points forts et faire en sorte d'atténuer les points faibles des unes grâce aux autres. Le deuxième principe est celui de la séparation des préoccupations. Dans un cerveau par exemple, chaque partie possède sa spécialisation principale. En appliquant ce principe, il nous sera plus aisé de comprendre et d'améliorer les fonctions de chaque partie.

Dans ce but, nous avons décidé de combiner l'étiquetage morpho-syntaxique aussi appelé étiquetage grammatical avec un réseau neuronal dédié au système de questions-réponses. L'idée étant de profiter à la fois des structures grammaticales afin d'offrir un tremplin de compréhension du texte pour le réseau neuronal, et de profiter des propriétés d'apprentissage de ce dernier.



Etiquetage Grammatical
Source nlpforhackers.io

Dans un premier temps nous sommes en train de développer un modèle basé sur un réseau LSTM bidirectionnel [2]. Afin d'améliorer les performances de ce réseau, nous devons choisir un méchanisme d'attention performant ou en développer un nous même.



Réseau LSTM Bidirectionnel avec Attention
Source nlpforhackers.io

Méthodologie

Un aspect important des systèmes de QR est la recherche ou l'extraction des réponses. Le succès de cette opération peut être mesuré à l'aide d'une combinaison de critères multiples, dont la précision, le rappel et le classement réciproque moyen, qui peuvent tous être comparés ouvertement à des résultats concurrents dans l'ensemble du domaine. Et dans le cas de l'ensemble de données SQuAD, nous utilisons deux métriques appelées EM et F1, la première signifiant Exact Match qui veut dire correspondance exacte entre la réponse trouvée et celle de référence considérée comme vraie, il s'agit là d'une métrique binaire qui prend pour valeur 1 s'il y'a une correspondance exacte et 0 dans le cas contraire. La deuxième métrique quant à elle mesure le chevauchement moyen entre les prédictions du modèle et les intervalles des réponses de référence. Formellement, elle s'exprime comme suit:

F1=2* (Précision * Rappel) / (Précision + Rappel)

où la précision est définie comme le rapport entre le nombre de mots correctement prédits dans la plage de réponse et le nombre total de mots prédits. Le rappel, quant à lui, est le rapport entre le nombre de mots de la plage de réponse correctement prédits et le nombre total de mots de la plage de réponse. [3]

Résultats

Comme vous pouvez le voir ci-dessous, l'état de l'art se rapproche de plus en plus du score de la performance humaine sur l'ensemble de données SQuAD 2.0. sachant que sur la version 1.1., plusieurs modèles issus principalement des laboratoires de Google et Microsoft ont *dépassé* le score de la performance humaine.

Il est également important de noter que la plus part des modèles appliqués à cette branche du TALN sont basés sur les RNN, cependant, plus récemment certains modèles basés sur les réseaux de neurones convolutionnels ont réussis à rivaliser avec leurs performances sur des problèmes de traduction alors qu'ils étaient jusque là la panache du domaine de la vision par ordinateur. [4]

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Nov 23, 2018	PAML+BERT (ensemble) PINGAN GammaLab	83.435	85.992
2 Nov 04, 2018	AoA + DA + BERT (ensemble) Joint Laboratory of HIT and iFLYTEK Research	82.374	85.310
3 Nov 07, 2018	PAML+BERT (single model) PINGAN GammaLab	81.347	84.560
4 Nov 04, 2018	AoA + DA + BERT (single model) Joint Laboratory of HIT and iFLYTEK Research	81.178	84.251
5 Nov 04, 2018	Candi-Net+BERT (single model) 42Maru NLP Team	80.106	82.862
5 Nov 18, 2018	BERT (single model) Google AI Language	80.005	83.061
6 Nov 18, 2018	Lénet + BERT (single model) Layer 6 AI	79.181	82.259
7 Nov 08, 2018	SLQA+BERT (single model) Alibaba DAMO NLP	77.003	80.209

Classement SQuAD

Dans de prochaines publications, nous auront des résultats concrets à communiquer autour de nos travaux ainsi qu'une démonstration interactive accessible par internet.

Conclusions

Dans ce travail nous avons conçu un agent virtuel intelligent alliant plusieurs disciplines de savoir, parmi lesquelles nous pouvons citer le traitement du langage naturel, la recherche d'information ainsi que les bases de données.

Notre projet, dans son état actuel, consiste en un système de Question/Réponse. L'une des raisons de ce choix est l'existence d'ensembles de données de qualité permettant de comparer nos résultats avec l'état de l'art, comme par exemple l'ensemble appelé SQuAD.

Alliant les techniques à base de Réseaux de Neurones avec des éléments de la linguistique nous réussissons à former un assistant offrant un accès plus direct et plus précis à l'information.

La prochaine étape consiste à élargir l'ensemble des formats de données supportés par notre système et à lui offrir d'autres possibilités et capacités d'interaction.

Références

1. **Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250, 2016.**
2. **Schuster, M. and Paliwal, K.K., 1997. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), pp.2673-2681.**
3. **Park, D.H. and Lakshman, V., Question Answering on the SQuAD Dataset.**
4. **Gehring, J., Auli, M., Grangier, D., Yarats, D. and Dauphin, Y.N., 2017. Convolutional sequence to sequence learning. arXiv preprint arXiv:1705.03122.**