Inférence textuelle interprétable en langage naturelle

Marjorie Armando

Laboratoire d'Informatique et des Systèmes - LIS

Encadrant

Benoit Favre

Résumé

La reconnaissance de l'inférence textuelle (Recognizing Textual Entailment : RTE) est un domaine assez récent en traitement automatique du langage (TAL). Le but de la RTE est de savoir automatiquement si une phrase, appelée l'hypothèse, est déduite d'une autre phrase, appelée la prémisse. Pour résoudre cela, on utilise des systèmes issus de l'apprentissage automatique.
Le problème qui se pose avec l'utilisation de ces systèmes est que le modèle nous fournit uniquement les probabilités de chaque label. Ainsi, aucune information supplémentaire n'est donnée : comment savoir, de manière humainement compréhensible, pourquoi le modèle a associé un label particulier à une paire de phrases donnée ?
Pouvoir répondre à cette question permettrait de rendre un modèle utilisable dans des domaines où les décisions doivent être mûrement réfléchies telle que la médecine, car malgré l'expansion des réseaux de neurones, ceux-ci restent des boîtes noires et il est donc difficile de leur faire confiance sans avoir d'explications en retour.
Dans ce travail, nous proposons une nouvelle méthode respectant les règles d'une "bonne" explication pour rendre un modèle interprétable dans le cadre de la RTE avec l'utilisation du corpus SNLI [1]. Nous allons la comparer avec la méthode Local Interpretable Model-agnostic Explanations (LIME) [2] qui permet d'expliquer les prédictions de n'importe quel classifieur en apprenant localement un modèle interprétable dans le voisinage de l'entrée.

Mots-clés : inférence textuelle, apprentissage automatique,
traitement automatique des langues naturelles, interprétabilité, LSTM

1 Introduction

1.1 Contexte de l'étude

Le traitement automatique du langage naturel ou de la langue naturelle (TALN) ou des langues (TAL) est un domaine pluridisciplinaire, qui fait collaborer l'intelligence artificielle, l'informatique théorique, la logique, la linguistique ou encore les statistiques en vue de modéliser et de reproduire à l'aide de machines, la capacité humaine à produire et à comprendre les énoncés linguistiques dans des buts de communication.

Dans le domaine du TAL, nous retrouvons plusieurs niveaux d'analyses linguistiques pour représenter au mieux les langues naturelles par les machines. On peut par exemple citer l'analyse lexicale qui permet d'identifier quels sont et où sont les mots, ou encore l'analyse sémantique qui permet de comprendre le sens des mots.

La RTE permet d'apporter des méthodes pour l'analyse lexicale et sémantique. Ceci peut permettre le développement de diverses applications telles que la recherche d'information ou encore le système de question/réponse.

Le but de la RTE est de savoir automatiquement si, à partir de deux phrases, on peut en déduire la deuxième de la première. La première phrase est appelée la prémisse, et la seconde l'hypothèse. Trois étiquettes permettent d'illustrer la relation entre la prémisse et l'hypothèse :

- **Contradiction** : l'hypothèse contredit la prémisse.
  Exemple : P : "Le chat est entièrement blanc" ; H : "Le chat est entièrement noir"
- **Neutre** : l'hypothèse est possible dans le contexte de la prémisse.
  Exemple : P : "Le chat dort sur la banquette" ; H : "Le chat aime le chocolat"
- **Inférence** : l'hypothèse est déduite de la prémisse.
  Exemple : P : "Le chat aimerait manger la sourie" ; H : "Le chat a faim"

Pour résoudre cela, on utilise des systèmes issus de l'apprentissage automatique. Nous utilisons les réseaux de neurones récurrents (Recurrent Neural Network : RNN) dans ce projet.

1.2 Problématique

Le problème qui se pose avec l'utilisation de système issue de l'apprentissage automatique est que le modèle nous fournit uniquement les probabilités de chaque label. Comment savoir, de manière humainement compréhensible, pourquoi le modèle a associé un label particulier à une paire de phrases donnée ?

Pouvoir répondre à cette question permettrait de rendre un modèle utilisable dans des domaines où les décisions doivent être mûrement réfléchies, telle que la médecine. Si un modèle propose de donner un certain traitement à un patient, il faut que le modèle puisse donner de bonnes explications pour que le docteur

l'approuve, car les conséquences pourraient être catastrophique.

Ainsi, un modèle doit être digne de confiance. Cette confiance donnée par les humains à un système dépend des explications données. Nous verrons dans la partie "" quels sont les critères d'une "bonne" explication. De plus, si un système est jugé digne de confiance, il sera davantage utilisé : en effet, il a été observé, par exemple, que le fait de fournir des explications augmente l'acceptation des recommandations de films [2].

Pouvoir donner une interprétation du système peut donc permettre aux utilisateurs d'adopter un modèle. Mais cela peut également aider le développeur lors du choix d'un modèle parmi ceux qu'il a implémenté, car le taux de réussite n'est pas le seul critère à prendre en compte.

/* FAIRE LE PLAN DU RAPPORT */

2 Etat de l'art

2.1 Dans le domaine de la RTE

Avant la publication du corpus SNLI [1], la recherche sur l'apprentissage automatique dans ce domaine été limitée par le manque de corpus assez grand. Depuis, les plus grands taux de réussite ont atteint les 89,3% [3].

2.2 Dans l'interprétabilité

Dans la pratique, il y a souvent un compromis entre le taux de réussite et l'interprétabilité du modèle. Certains modèles tels que les arbres de décisions ou encore les modèle linéaires sont facilement interprétables, et sont donc parfois utilisés à la place de modèles complexes tels que les réseaux de neurones profonds, même si ceux-ci peuvent donner de meilleurs résultats. Cependant, au cours de ces derniers mois, de grandes avancées ont été effectuées pour tenter d'interpréter les modèles utilisés jusqu'alors comme des boîtes noires.

Nous étudions dans ce rapport la méthode LIME [2], mais nous pouvons également citer la méthode SHapley Additive exPlanations (SHAP) [**?**] : elle explique la prédiction de n'importe quel modèle en utilisant les valeurs de Shapley, introduites dans la théorie des jeux en 1953. Ces valeurs ont récemment été utilisées pour attribuer une mesure d'importance aux features [4]. SHAP appartient à la classe des "méthodes d'attribution de features additives" : elle attribue une valeur à chaque feature pour chaque prédiction (c'est-à-dire une attribution de feature). Plus la valeur est haute, plus l'attribution de la feature à la prédiction spécifique est grande : la somme de ces valeurs devrait donc être proche de la prédiction initiale du modèle.

SHAP a rassemblé et unifié six méthodes d'attribution de features additives, dont LIME, et est la seule méthode qui respecte les trois propriétés suivantes :

- Précision locale : les explications sont fidèles et véridiques au modèle.

- **Feature manquante** : les features retirées n'ont aucun impact attribué aux prédictions du modèle.

- **Cohérence** : les explications sont cohérentes avec l'intuition humaine. Techniquement, la cohérence indique que si un modèle change de sorte que la contribution d'une entrée augmente ou reste la même indépendamment des autres entrées, l'attribution de cette entrée ne devrait pas diminuer.

Le calcul de ces valeurs est cependant très coûteux : il faut entraîner le modèle sur tous les sous-ensembles de features S inclu F, où F est l'ensemble de toutes les features. Les valeurs Shapley attribuent une valeur d'importance à chaque feature, ce qui représente l'effet sur la prédiction du modèle d'inclure cette feature. Pour cela, un modèle est entraîné avec la feature présente, et un autre modèle est entraîné avec la feature retirée. Ensuite, les prédictions des deux modèles sont comparées sur l'entrée courante, c'est-à-dire qu'on calcule leur différence. Comme l'effet de la supression d'une feature dépend d'autres features du modèle, les différences précédentes sont calculées pour tous les sous-ensembles possibles de features. Les valeurs de Shapley sont une moyenne pondérée de toutes les différences possibles et sont utilisées comme attribution de feature.

Un algorithme efficace a été mis au point, cependant il fonctionne uniquement sur les modèles basés sur les arbres.

3 Réseaux de neurones récurrents

3.1 LSTM et Bi-LSTM

Dans ce travail, nous utilisons les réseaux de neurones récurrents, et plus particulièrement les LSTMs (long short-term memory : LSTM) et les Bi-LSTMSs (bidirectionnal long short-term memory : Bi-LSTM). /*Expliquer RNN et LSTM*/ Les états cachés sont calculés à chaque instant t comme suit :

$$entree : i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{1}$$

$$oublie : f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{2}$$

$$etat : u_t = \tanh(W_u x_t + U_u h_{t-1} + b_u) \tag{3}$$

$$sortie : o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot u_t \tag{5}$$

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

3.2 Systèmes utilisés

Nous avons implémenté 3 systèmes différents avec la librairie DyNet [5] :

Le premier système passe la prémisse et l'hypothèse au LSTM pour avoir une représentation pour chacune de ces deux phrases. Ensuite, il concatène les deux

représentations obtenues pour l'envoyer à une couche de décision en effectuant la formule suivante :

$$W * [LSTM(prémisse)\ ; LSTM(hypothèse)] + b \tag{7}$$

où W est la matrice de poids, LSTM(*prémisse*) et LSTM(*hypothèse*) sont respectivement la représentation de la prémisse et de l'hypothèse, b est le biais, et [;] dénote la concaténation.

Le deuxième système effectue le même mécanisme que le premier système pour avoir une représentation de la prémisse et de l'hypothèse. Ensuite, il compare les deux représentations puis il envoit cela à une couche de décision en effectuant la formule suivante :

$$W * LSTM(prémisse) * LSTM(hypothèse)^T + b \tag{8}$$

où T dénote la transposée.

Le troisième système est inspiré de la méthode KIM [6].
On représente les mots de la prémisse et de l'hypothèse en les passant dans un Bi-LSTM : il utilise un LSTM forward pour lire la phrase de gauche à droite, puis un LSTM backward pour lire la phrase dans l'autre sens. A chaque mot lu, un état caché est généré par les deux LSTMs. Ces deux états cachés sont alors concaténer pour obtenir une représentation du mot :
$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$, où $\overrightarrow{h_t}$ est l'état caché généré par le LSTM forward à l'instant t, $\overleftarrow{h_t}$ est celui généré par le LSTM backward à l'instant t, et h_t est la représentation du mot t.
On dénote par p^s (respectivement h^s) le vecteur de représentation des mots de la prémisse (respectivement de l'hypothèse).
Nous construisons ensuite une matrice d'alignement comme suit :

$$e_{ij} = (p_i^s)^T h_j^s \tag{9}$$

où p_i^s est la représentation du i$^{\text{ème}}$ mot de la prémisse, h_j^s est celle du j$^{\text{ième}}$ mot de l'hypothèse.
Avec cette matrice, nous pouvons alors construire les vecteurs de contexte suivants pour la prémisse et l'hypothèse :

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{N} exp(e_{ik})}, p_i^c = \sum_{j=1}^{N} \alpha_{ij} h_j^s \tag{10}$$

$$\beta_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{M} exp(e_{kj})}, h_j^c = \sum_{i=1}^{M} \beta_{ij} p_i^s \tag{11}$$

où M est la longueur de la prémisse, N est la longueur de l'hypothèse, $\alpha \in \mathbb{R}^{M*N}$ est un softmax sur l'hypothèse, $\beta \in \mathbb{R}^{M*N}$ est un softmax sur la prémisse, /*à finir, parler des 2 vecteurs de contexte*/

Leurs performances sont décrit dans la partie "résultats expérimentaux" en annexe.

4 Interprétabilité

Il n'y a malheureusement pas de consensus concernant la définition d' "interprétabilité". Miller définit cela comme étant le degré auquel un humain peut comprendre la cause d'une décision[7]. Un système a donc une meilleure interprétabilité qu'un autre si ses explications sont plus faciles à comprendre par un humain.

4.1 Qu'est-ce-qu'une explication ?

La définition donnée par Miller est assez simple : une explication est une réponse à une question commençant par "pourquoi". Une question commençant par "comment" peut être retournée en une question commençant par "pourquoi". Le terme "explication" désigne le processus social et cognitif d'expliquer, mais c'est également le produit de ces processus.

4.2 Qu'est-ce-qu'une "bonne" explication ?

La définition d'une bonne explication ne doit pas se baser sur l'intuition de l'auteur, mais plutôt sur des faits. Miller résume ce qu'est une bonne explication[4], basée sur ce que les humains attendent d'une explication. Grâce à cela, nous allons énoncer les types d'explications adéquats à notre projet. Cependant, il ne faut pas oublier que les humains ont tendance à rejeter toutes explications allant à l'encontre de leur croyance.

4.2.1 Explication contrastée

C'est une explication qui doit être comparée. Les utilisateurs se demandent généralement pourquoi cette prédiction a été faite et pas une autre, via la question "quelle aurait été la prédiction si cette entrée avait été changé par une autre ?".
Un docteur se demandant "pourquoi ce traitement ne marche pas sur ce patient ?" voudrait comparer les données de ce patient à un autre patient ayant des caractéristiques similaires mais pour qui le traitement marche.
La meilleure explication pour ce type d'explication est celle qui met en évidence les différences entre l'entrée traitée et l'entrée de comparaison.
L'entrée de comparaison peut être artificielle.

4.2.2 Explication sélective

C'est une explication qui doit être courte. Généralement, on peut expliquer un phénomène par plusieurs facteurs (c'est ce qu'on appelle l'effet de Rashomon). Il faut en donner peu, à savoir deux ou trois raisons, même si les explications peuvent être plus complexes que cela.

4.2.3 Explication sociale

Comme nous l'avons expliqué ci-dessus, une explication est un processus social, c'est-à-dire qu'il faut prendre en compte les connaissances de la personne à qui l'on veut donner une explication. Dans notre projet, nous partons du principe qu'une explication doit être comprise par tout le monde, que ce soit

par un expert du domaine de l'apprentissage automatique ou bien par quelqu'un qui n'en a jamais entendu parler.

4.2.4 Explication anormale

Les explications "anormales" sont beaucoup appréciées, c'est-à-dire que si une cause rare a influencé la prédiction, il faut la spécifier. Dans notre projet, une cause rare peut être un mot -ou un groupe de mot- de l'hypothèse qui ne peut pas être mis en relation avec un mot -ou un groupe de mot- de la prémisse, mais qui a influencé la prédiction.

Références

[1] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2015.

[2] Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. 'why should i trust you ?' : Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[3] Reza Ghaeini, Sadid A. Hasan, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Z. Fern, and Oladimeji Farri. Dr-bilstm : Dependent reading bidirectional lstm for natural language inference. In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL HLT)*, 2018.

[4] Christoph Molnar. *Interpretable Machine Learning*. Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, 2018.

[5] Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. Dynet : The dynamic neural network toolkit. *arXiv preprint arXiv :1701.03980*, 2017.

[6] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. Natural language inference with external knowledge. *arXiv preprint arXiv :1711.04289*, 2017.

[7] Tim Miller. Explanation in artificial intelligence : Insights from the social sciences. *arXiv preprint arXiv :1706.07269*, 2017.

A Résulats expérimentaux

Pour la représentation des mots, nous utilisons des words embeddings préentrainés de dimension 100 via GloVe.6B.100d. Pour les mots inconnus, c'est-à-dire les mots qui n'ont pas d'embedding dans GloVe, nous utilisons des embeddings initialisés au hasard. Tous les embeddings sont mis-à-jour par le réseau.

La taille des batches est de 16 et le dropout est à 0,3.
Concernant les RNNs utilisés, il y a qu'une seule couche et la dimension des états cachés est de 100.

Système	Taux de réussite Contradiction	Taux de réussite Inférence	Taux de réussite Neutre	Taux de réussite Dev	Tau
1	67,33%	73,90%	68,38%	69,89%	test
2	78,55%	87,98%	74,99%	80,57%	test
3	68,73%	74,92%	69,55%	71,09%	test