

Interprétation de prédictions de modèles pour l'inférence textuelle en perturbant significativement les entrées

Marjorie Armando

Laboratoire d'Informatique et des Systèmes - LIS

Encadrant

Benoit Favre

Résumé

La reconnaissance de l'inférence textuelle (Recognizing Textual Entailment : RTE) est au coeur de tous les aspects de la compréhension de texte 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. En utilisant des réseaux de neurones complexes, nous pouvons obtenir de bons taux de réussite pour la RTE. Cependant, ces réseaux ne sont pas interprétables, ainsi nous n'avons pas la possibilité de savoir si le modèle s'est basé sur de bonnes informations pour sa décision. Dans ce travail, nous proposons la méthode Best Adversarial eXemple for Interpretability (BAXI), qui respecte 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 à 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. Pour cela, nous avons extrait un échantillon du corpus de test de SNLI, et demandé à six annotateurs de donner les mots expliquant l'étiquette associée à la paire prémisse/hypothèse. Grâce à ce nouveau corpus, nous mesurons le taux de la qualité de l'explication fournit par la méthode BAXI et la méthode LIME, sur plusieurs systèmes que nous avons implémenté.

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

Table des matières

1	Introduction	4
1.1	Contexte et motivations	4
1.2	Tâches	4
2	Hypothèse scientifique	5
3	Etat de l'art	5
3.1	Dans le domaine de la RTE	5
3.2	Dans l'interprétabilité pour le TAL	6
3.3	Dans l'interprétabilité de la RTE	7
4	Définition d'interprétabilité	7
4.1	Qu'est-ce-qu'une explication?	7
4.2	Qu'est-ce-qu'une "bonne" explication?	7
5	Systèmes implémentés	9
5.0.1	Systèmes basés sur l'encodage des phrases	9
5.0.2	Système avec mécanisme d'attention	9
6	Description de l'approche BAXI	10
7	Cadre expérimental	11
7.1	Corpus SNLI et représentation de mots	11
7.2	Mots de remplacement	12
7.3	Corpus d'explications	13
7.4	Métrique	13
7.5	Paramètres	14
8	Résultats	14
9	Discussion	15
10	Conclusion	15
11	Remerciements	15
	Références	16

Liste des tableaux

1	Echantillon du corpus de développement de SNLI	12
2	Taux des étiquettes présentes dans le corpus de l'échantillon de test.	12
3	Résultats des tests de la RTE pour le fichier de validation.	14
4	Résultats des tests de la RTE pour le fichier de test. . . .	14
5	Résultats des tests de la RTE pour le corpus de l'échantillon de test.	14
6	Résultats des mesures de BAXI et LIME	15

Table des figures

1	Exemple présentant l'intuition de LIME.	7
2	Concept des intelligences artificielles explicables	8
3	Echantillon du corpus d'explications	13

1 Introduction

1.1 Contexte et motivations

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 modèles 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.

Pouvoir interpréter un modèle permettrait de le rendre 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.

Dans le cadre de la RTE, les modèles obtenant les meilleurs taux de réussite, décrit dans la section , sont non interprétables. Le problème qui se pose avec l'utilisation de modèles non interprétables est qu'ils fournissent uniquement les probabilités de chaque étiquette. Comment savoir, de manière humainement compréhensible, pourquoi le modèle a associé une étiquette particulière à une certaine entrée ?

Avoir un modèle digne de confiance permettrait au modèle d'être d'avantage utilisé : en effet, il a été observé, par exemple, que le fait de fournir des explications augmente l'acceptation des recommandations de films [2]. De plus, cela permettrait également au développeur de choisir un modèle parmi ceux qu'il a implémenté, car le taux de réussite n'est pas le seul critère à prendre en compte : un modèle peut fournir l'étiquette attendue en se basant sur de mauvaises informations.

1.2 Tâches

Les concepts sémantiques d'inférence et de contradiction sont au coeur de tous les aspects de la compréhension de texte en TAL. Ainsi, l'inférence textuelle en langage naturel (Natural Language Inference : NLI), caractérisant ces relations, est essentielle dans des tâches telles que la recherche d'information, le raisonnement de bon sens, ou encore le système de question/réponse.

Soit une paire de phrase prémisses/hypothèse, la RTE se voit comme objectif de détecter si la seconde phrase est en contradiction, se déduit ou bien est neutre par rapport à la première phrase. Il y a donc trois étiquettes permettant d'illustrer la relation entre la prémisse et l'hypothèse :

- Contradiction : l'hypothèse contredit la prémisse. Par exemple :
Prémisse : "Le chat est entièrement blanc."
Hypothèse : "Le chat est entièrement noir."
- Neutre : l'hypothèse est possible dans le contexte de la prémisse. Par exemple :
Prémisse : "Le chat dort sur la banquette."

Hypothèse : "Le chat aime le chocolat."

- Inférence : l'hypothèse est déduite de la prémisse. Par exemple :
Prémisse : "Le chat aimerait manger la souris."
Hypothèse : "Le chat a faim."

Dans ce travail, nous proposons d'interpréter les prédictions des modèles axés sur la RTE avec la méthode Best Adversarial eXemple for Interpretability (BAXI), expliquée en section 4. Nous allons comparer ses performances avec la méthode Local Interpretable Model-agnostic Explanations (LIME) de Ribeiro et al., expliquée dans la section 3.3, grâce à un corpus d'explications que nous avons créé à l'aide de six annotateurs. Les résultats sont visibles en section 6.

Nous allons tout d'abord lister nos hypothèses scientifiques, puis énoncer l'état de l'art pour la RTE ainsi que pour l'interprétabilité de prédictions de modèle en général. Puis, nous allons décrire notre méthode ainsi que le cadre expérimental. Enfin, nous allons décrire nos résultats obtenus puis conclure sur les différentes perspectives.

2 Hypothèse scientifique

3 Etat de l'art

3.1 Dans le domaine de la RTE

Les premiers travaux sur la RTE ont été formés sur de très petits ensembles de données avec des méthodes conventionnelles, telles que les méthodes peu profondes [3] ou encore les méthodes de logique naturelle [4].

Ces dernières années, on dénote une nette amélioration pour la tâche de la RTE, en particulier grâce à la publication du corpus SNLI (the Stanford Natural Language Inference) qui contient 570K paires de phrases annotées. Cela a permis d'entraîner des réseaux de neurones complexes, puisqu'ils nécessitent une quantité relativement importante de données. En effet, la recherche sur l'apprentissage automatique dans ce domaine a été limitée par le manque de corpus assez grand.

On peut distinguer deux types d'approches parmi ces modèles : les modèles basés sur l'encodage des phrases et les modèles axés sur l'attention inter-phrases. Les premiers types de modèles encodent les phrases puis un classifieur (un perceptron multi-couches) décide de la relation entre ces deux phrases encodées. Différents encodeurs ont été proposés, tels que les Long-Short Term Memory (LSTM) [1], les Gated Recurrent Unit (GRU) [5], les réseaux de neurones à convolution (Convolutional Neural Network : CNN) [6], les Bidirectionnal Long-Short Term Memory (BiLSTM) et ses variantes [7] [8] [9], et des réseaux neuronaux plus complexes [10] [11].

L'avantage de ces modèles est que les encodeurs transforment les phrases en vecteur de taille fixe, ce qui peut aider à un large éventail de tâches de transfert [12]. Cependant, cette architecture ignore l'interaction locale entre deux phrases, pourtant nécessaire [4].

Les seconds modèles ont alors été proposés pour éviter ce problème. Dans ce

cadre, l'information d'inférence locale est collectée par le mécanisme d'attention puis introduite dans des réseaux de neurones pour composer des vecteurs de taille fixe avant la classification finale. Beaucoup suivent cette route. Parmi eux, Rocktäschel et al. [13] ont été les premiers à proposer des réseaux basés sur l'attention pour la RTE. Chen et al. [14] ont proposé un modèle d'inférence séquentielle amélioré qui est l'un des meilleurs modèles à ce jour (88,6% de taux de réussite).

3.2 Dans l'interprétabilité pour le TAL

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 modèles 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 dernières années, de grandes avancées ont été effectuées pour tenter d'interpréter les modèles utilisés jusqu'alors comme des boîtes noires.

Nous étudions dans ce travail la méthode LIME qui permet d'expliquer les prédictions de n'importe quel classifieur ou régresseur. L'objectif global de LIME est d'identifier un modèle interprétable parmi le voisinage de l'entrée x .

Tout d'abord, on distingue les features utilisées aux représentations interprétables des features. Par exemple, les features sont les embeddings des mots et la représentation interprétable de ces features est un vecteur binaire qui indique la présence ou l'absence des mots.

LIME définit une explication par un modèle $g \in G$, où G est la classe des modèles interprétables tels que les modèles linéaires ou les arbres de décisions. Vu que les modèles interprétables n'ont pas tous la même difficulté à être interprétés, LIME définit $\Omega(g)$ qui est une mesure de la complexité d'interpréter g . En prenant l'exemple des arbres de décisions, $\Omega(g)$ est la profondeur.

On dénote par $f : \mathbb{R}^d \rightarrow \mathbb{R}$ le modèle utilisé comme une boîte noire. $f(x)$ est la probabilité que l'entrée x appartienne à une certaine classe.

LIME va alors se baser sur la représentation interprétable des données en retirant un ou plusieurs mots au hasard. On dénote cette nouvelle entrée par z . LIME définit la localité de x avec $\pi_x(z)$ qui est une mesure de proximité entre z et x . C'est un noyau se basant sur la similarité cosinus.

Enfin, on définit $\mathcal{L}(f, g, \pi_x)$ qui est une mesure pour savoir à combien g est infidèle à f dans la localité défini par π_x . Pour préserver à la fois l'interprétabilité et la fidélité locale, LIME minimise $\mathcal{L}(f, g, \pi_x)$ avec $\Omega(g)$ assez petit pour être interprétable par les humains. L'explication de LIME est donc la suivante :

$$\mathcal{E}(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (1)$$

Cette formule peut être utilisée par différents modèles $g \in G$, fonctions de fidélité $\mathcal{L}(f, g, \pi_x)$, et mesure de complexité $\Omega(g)$.

LIME peut alors donner les K mots les plus importants de l'entrée x pour

tout label.

La figure ci-dessous est un exemple illustrant le principe de LIME :

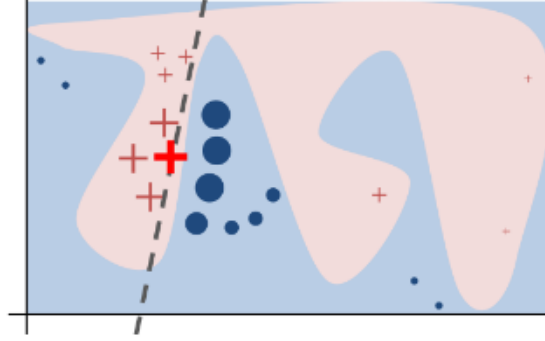


FIGURE 1 – Exemple présentant l'intuition de LIME.

La décision de la boîte noire f , inconnue par LIME, est représentée par le fond bleu et rose. La croix rouge en gras est l'entrée x que l'on veut expliquer. LIME crée des entrées modifiées, utilise f pour avoir la probabilité de ces entrées pour le label y , et les pondère par leur proximité par rapport à x (les poids sont représentés par la taille). La droite pointillée est l'explication apprise qui est localement fidèle.

3.3 Dans l'interprétabilité de la RTE

Ce n'est que très récemment que Silva et al. [15]
/*PARLER ARTICLE LREC*/

4 Définition d'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 [16]. 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

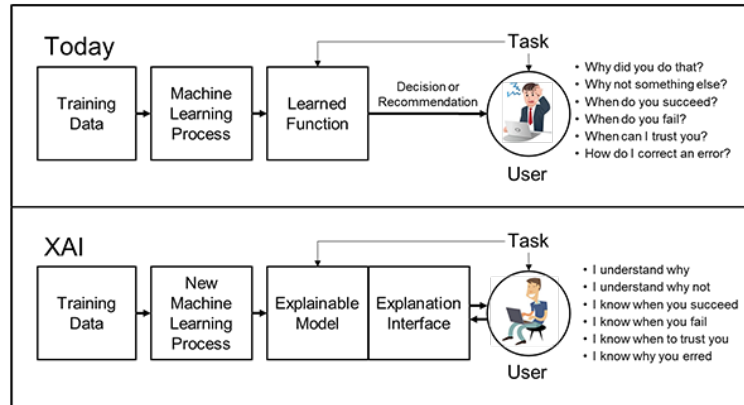


FIGURE 2 – Concept des intelligences artificielles explicables (eXplainable Artificial Intelligences : XAI) [17].

[18], 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.

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 changée 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.

Explication sélective C'est une explication qui doit être courte. Généralement, on peut expliquer un phénomène par plusieurs facteurs. Il faut en donner peu, à savoir deux ou trois raisons, même si les explications peuvent être plus complexes que cela.

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.

5 Systèmes implémentés

Nous avons implémenté 3 systèmes différents avec la librairie DyNet [19] en C++. Nous utilisons les LSTMs et les BiLSTMs.

5.0.1 Systèmes basés sur l’encodage des phrases

Premier système Le premier système passe la prémisse et l’hypothèse au LSTM pour avoir une représentation pour chacune de ces deux phrases. On les concatène pour les envoyer ensuite à une couche de décision :

$$y = \text{softmax}(W \times [LSTM(\text{prémisse}) ; LSTM(\text{hypothèse})] + b) \quad (2)$$

où y est un vecteur contenant les probabilités de chaque label, W est la matrice de poids, $LSTM(\text{prémisse})$ et $LSTM(\text{hypothèse})$ sont respectivement la représentation de la prémisse et de l’hypothèse, b est le biais, et $[\cdot]$ dénote la concaténation.

Deuxième système Le deuxième système effectue le même mécanisme que le premier système pour avoir une représentation de la prémisse et de l’hypothèse. On compare les deux représentations pour envoyer la comparaison à une couche de décision :

$$y = \text{softmax}(W \times (LSTM(\text{prémisse}) \times LSTM(\text{hypothèse})^T) + b) \quad (3)$$

où T dénote la transposée.

5.0.2 Système avec mécanisme d’attention

Troisième système Le troisième système est inspiré de la méthode KIM [20]. On représente les mots de la prémisse et de l’hypothèse en les passant dans un BiLSTM : 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 concaténés pour obtenir une représentation du mot :

$h_t = [h_t^{\rightarrow}; h_t^{\leftarrow}]$, où h_t^{\rightarrow} est l’état caché généré par le LSTM forward à l’instant t , h_t^{\leftarrow} 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 :

$$e_{ij} = (p_i^s)^T h_j^s \quad (4)$$

où p_i^s est la représentation du $i^{\text{ème}}$ mot de la prémisse, h_j^s est celle du $j^{\text{ème}}$ mot de l’hypothèse.

Avec cette matrice, nous pouvons alors construire les vecteurs de contexte p^c et h^c suivants pour la prémisse et l’hypothèse :

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^N \exp(e_{ik})}, p_i^c = \sum_{j=1}^N \alpha_{ij} h_j^s \quad (5)$$

$$\beta_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^M \exp(e_{kj})}, h_j^c = \sum_{i=1}^M \beta_{ij} p_i^s \quad (6)$$

où M est la longueur de la prémisse, N est la longueur de l'hypothèse, $\alpha \in \mathbb{R}^{M \times N}$ est un $\text{softmax}(e)$ sur la prémisse, et $\beta \in \mathbb{R}^{M \times N}$ est un $\text{softmax}(e)$ sur l'hypothèse. Ceci permet à la prémisse de voir le contexte de l'hypothèse et vice-versa.

Avec ces nouvelles représentations pour les mots, on effectue du mean-pooling :

$$pool_p = \frac{\sum_{i=1}^N p_i^c}{N}, pool_h = \frac{\sum_{i=1}^M h_i^c}{M} \quad (7)$$

puis on effectue une concaténation du mean-pooling de la prémisse et de l'hypothèse pour l'envoyer à une couche de décision :

$$y = \text{softmax}(W \times [pool_p; pool_h] + b) \quad (8)$$

6 Description de l'approche BAXI

Cette section décrit notre technique pour interpreter une prédiction d'une entrée. L'objectif est de donner des explications pour chaque étiquette en donnant les α mots les plus importants dans la prémisse et les β mots les plus importants dans l'hypothèse. On veut donc calculer l'importance de chaque mot. L'importance d'un mot correspond à l'importance de sa contribution pour l'étiquette y .

Notre intuition est la suivante : On veut donner une explication pour l'étiquette y et l'entrée x composée de la prémisse et de l'hypothèse. On remplace alors un mot et on demande à notre modèle de nous donner les probabilités de chaque étiquette avec cette nouvelle entrée. Si la probabilité de l'étiquette y a baissé, alors le mot était important : cela veut dire que le mot avait contribué à l'étiquette y . A l'inverse, si elle a augmenté, le mot n'avait donc pas contribué à y . De plus, si la probabilité des autres étiquettes a augmenté, alors le mot a d'autant plus d'importance : en le remplaçant, l'entrée x a basculé vers une autre étiquette.

Pour résumé, lorsque l'on remplace m_i , nous pénalisons l'augmentation de la probabilité de l'étiquette que l'on veut expliquer, et nous encourageons l'augmentation des probabilités des autres étiquettes.

L'importance d'un mot, notée IMP^{total} pour "importance", est une fonction comme suit :

$$impact^{y_{ref}}(m_i) = -p(y_{ref} | x \leftarrow m_i = m_a) + p(y_{ref} | x) \quad (9)$$

$$impact^{y_j}(m_i) = \sum_{j=1, j \neq ref}^{|Y|} p(y_j | x \leftarrow m_i = m_a) - p(y_j | x) \quad (10)$$

$$IMP(m_i) = \max_{m_a} (impact^{y_{ref}}(m_i) + impact^{y_j}(m_i)) \quad (11)$$

où y_{ref} est le label de référence (le label que l'on veut "expliquer"), x est l'entrée, m_i est le mot que l'on retire de l'entrée x , y_j est un label différent de y_{ref} , et $|Y|$ est le nombre de label. Les probabilités sont exprimées en log-probabilités pour des soucis de précision et d'optimisation du code. On parle donc plutôt de score.

Cette formule suit notre intuition de base et également celle de Robnik-Sikonja et Kononenko [21]. Nous avons rajouté la prise en compte de l'impact sur les scores des autres labels quand on retire le mot m_i .

Un problème persiste tout de même : que se passe-t-il vraiment lorsque l'on retire un mot ? Nous ne pouvons pas simplement le supprimer de l'entrée, car la phrase n'aurait plus de sens et le modèle n'est peut-être pas entraîné à faire face aux bruits. On ne peut pas vraiment le remplacer par un mot générique (le mot "UNK" pour "mot inconnu", par exemple) car un modèle remplaçant ce mot par un embedding à 0 et un modèle remplaçant ce mot par un embedding particulier auront des réponses très différentes. Or on voudrait que notre méthode soit réalisable par n'importe quel modèle. La solution est donc de remplacer le mot que l'on veut retirer par un autre :

$$p(y|x \setminus m_i) = \sum_{s=1}^{|S|} p(y|x \leftarrow m_i = m_s) \quad (12)$$

où $|S|$ est le nombre de mot que l'on peut mettre à la place du mot m_i et m_s est le mot que l'on met à la place de m_i .

Plus il y a de mots de remplacements, plus on a une meilleure estimation de l'importance du mot m_i , mais dans ce cas la complexité en temps augmente. Il faut donc trouver un compromis entre l'estimation de l'importance d'un mot et le temps que met le programme pour donner une explication. Nous expliquons comment et par quoi nous avons remplacé les mots dans la section 6.1.

Cette méthode permet donc d'avoir une explication contrastée, puisque l'on compare notre entrée de base avec des entrées créées artificiellement en remplaçant un mot par un autre. De plus, elle est également sélective puisque l'on sélectionne les K mots, respectivement dans la prémisse puis dans l'hypothèse, ayant le plus contribué au label y_{ref} . Enfin, le programme surligne les mots les plus importants pour chaque label, ce qui permet d'avoir une visualisation pour faciliter la compréhension. Nous respectons donc les trois règles d'une bonne explication citées dans la section 3.

/* METTRE UN SCREEN D'UNE EXPLICATION */

7 Cadre expérimental

7.1 Corpus SNLI et représentation de mots

Nous utilisons les corpus SNLI composés d'un fichier d'entraînement, de validation et de test. Ces corpus ont été réalisés par cinq annotateurs à l'aide d'une image accompagnée d'un texte bref -la prémisse- présentant la dite image. Les

annotateurs ont alors écrit une phrase étant neutre par rapport à la prémisse, une autre étant en contradiction et une autre phrase qui pouvait être déduite de la prémisse : ils ont donné une hypothèse et une étiquette. Pour que le corpus ne soit pas trop subjectif, les annotateurs ont eu accès à quelques paires prémisse/hypothèse sans étiquette. Chacun d’entre eux a donné une étiquette, celle ayant eu le plus de voix a été décidé comme l’étiquette gold de la paire observée. Ainsi, certaines paires n’ont pas d’étiquette car les annotateurs n’ont pas trouvé de consensus : nous ne prenons pas en compte ce genre d’entrée. La table ci-dessous est un échantillon du corpus SNLI :

A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

TABLE 1 – **Echantillon de 5 paires du corpus de développement de SNLI** présentant à gauche la prémisse, à droite l’hypothèse, et au centre les étiquettes des 5 annotateurs (C pour Contradiction, N pour Neutral, et E pour Entailment). L’étiquette en gras est celle qui a eu le plus de voix, et est donc l’étiquette gold de la paire prémisse/hypothèse.

Pour mesurer les taux de réussite des explications, nous utilisons un échantillon du corpus de test de SNLI composé de 19 paires prémisse/hypothèse avec leur étiquette associée. La table ci-dessous montre les taux des étiquettes présentes dans ce petit corpus :

% Inférence	% Neutre	% Contradiction
36,84%	26,31%	36,84%

TABLE 2 – **Taux des étiquettes présentes dans le corpus de l’échantillon de test.**

Pour la représentation des mots, nous utilisons des *words embeddings* pré-entraînés 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.

7.2 Mots de remplacement

Nous pensons que nous pouvons choisir automatiquement, pour chaque mot m_i d’une phrase, les mots pouvant les remplacer (avec des outils tel que Word-

Net, par exemple). Cependant, nous avons voulu les choisir nous-même pour assurer la qualité des mots et essayer d’avoir les meilleurs exemples de comparaisons possible. Nous avons donc annoté trois fichiers suivant le label que l’on veut "expliquer". Lorsque l’on parle "d’expliquer un label", cela signifie que l’on donne les K mots -dans la prémisse et dans l’hypothèse- qui ont le plus contribué au label que l’on veut expliquer.

L’idée pour l’annotation des mots de remplacement est la suivante : si le label que l’on veut expliquer se trouve être le label gold de la paire prémisse/hypothèse, on cherche alors à remplacer les mots des phrases par des mots pouvant baisser fortement le score du label.

A l’inverse, si le label que l’on veut expliquer n’est pas le label gold, on cherche alors à remplacer les mots de la phrase par des mots pouvant augmenter le score du label expliqué (et par la même occasion, faire baisser le score des autres labels).

/* FAIRE TABLE EXEMPLE DES MOTS DE REMPLACEMENT */

7.3 Corpus d’explications

La méthode habituelle pour ce genre d’expériences est de montrer à plusieurs personnes les paires de phrases prémisse/hypothèse avec leurs explications associées pour connaître leur avis sur la qualité de l’explication.

Nous proposons une autre méthode qui permet d’évaluer automatiquement la qualité d’une explication : nous avons montré un échantillon de 43 exemples issus du fichier de test de SNLI à six personnes pour qu’elles puissent annoter ce qu’elles pensent être une explication correcte, c’est-à-dire les mots de la prémisse et les mots de l’hypothèse qui conduisent au label associé. Nous avons récolté leurs réponses pour en faire un corpus d’explications de références, au format CSV.

La figure ci-dessous est un échantillon de ce corpus d’explications :

pair ID	label	premise	hypothesis	important words in the premise	important words in the hypothesis
4923271886.jpg#2r1e	entailment	Male in a blue jacket decides to lay in the grass .	The guy wearing a blue jacket is laying on the green grass	Male 0 blue 3 jacket 4 lay 7 grass 10	guy 1 blue 4 jacket 5 laying 7 grass 11
4923271886.jpg#2r1n	neutral	Male in a blue jacket decides to lay in the grass .	The guy wearing a blue jacket is laying on the green grass taking a nap .	-1	taking 12 nap 14
3551170666.jpg#3r1c	contradiction	A golden retriever nurses puppies .	A cat nurses puppies	golden 1 retriever 2 nurses 3 puppies 4	cat 1 nurses 2 puppies 3

FIGURE 3 – Echantillon du corpus d’explications. La première colonne est l’identifiant de la paire prémisse/hypothèse, la deuxième est l’étiquette, la troisième est la prémisse, la quatrième est l’hypothèse, la cinquième contient les mots de la prémisse expliquant l’étiquette, et la dernière colonne contient les mots de l’hypothèse expliquant l’étiquette. -1 signifie qu’il n’y a aucun mot important. Chaque numéro à droite d’un mot important correspond à sa position dans la phrase, en commençant par 0.

7.4 Métrique

/* A FINIR */

7.5 Paramètres

Tous les *embeddings* sont mis-à-jour par le réseau. La taille des *batches* est de 16.

Concernant les RNNs utilisés, il n’y a qu’une seule couche, la dimension des états cachés est de 100, et le *dropout* est à 0,3.

8 Résultats

Tout d’abord, les tables ci-dessous montrent les différents résultats des taux de réussite des étiquettes prédites avec les trois systèmes décrits dans la section 5 :

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

TABLE 3 – Résultats des tests de la RTE pour le fichier de validation.

Système	% Inférence	% Neutre	% Contradiction	% Dev
1	test%	test%	test%	test%
2	86,10%	76,14%	77,70%	80,07%
3	test%	test%	test%	test%

TABLE 4 – Résultats des tests de la RTE pour le fichier de test.

La table ci-dessous montre les taux de réussite des étiquettes prédites pour les trois systèmes décrits en section <> sur l’échantillon du corpus de test de SNLI :

Système	% Inférence	% Neutre	% Contradiction	% Total
1	71,43%	60,00%	42,86%	57,89%
2	42,86%	80,00%	57,14%	57,89%
3	71,43%	40,00%	42,86%	52,63%

TABLE 5 – Résultats des taux de réussite des étiquettes prédites pour le corpus de l’échantillon de test.

La table ci-dessous illustre les différents résultats obtenus pour le taux de réussite des explications des trois systèmes décrits en section <> sur l’échantillon du corpus de test de SNLI et sur le corpus d’explication :

Système	Méthode	% Inférence	% Neutre	% Contradiction	% Total
1	BAXI	61,70%	57,14%	51,61%	57,61%
	LIME	test	test	test	test
2	BAXI	76,60%	57,14%	54,84%	66,30%
	LIME	test	test	test	test
3	BAXI	61,70%	57,14%	51,61%	57,61%
	LIME	test	test	test	test

TABLE 6 – Résultats des mesures de BAXI et LIME pour le fichier de l'échantillon de test et pour le corpus d'explications.

9 Discussion

10 Conclusion

11 Remerciements

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] Oren Glickman, Ido Dagan, and Moshe Koppel. Web based probabilistic textual entailment. In *Proceedings of the PASCAL Challenges Workshop on Recognizing Textual Entailment*, 2005.
- [4] Bill MacCartney and Christopher D Manning. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 2007.
- [5] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *arXiv preprint arXiv :1511.06361*, 2015.
- [6] Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. Natural language inference by tree-based convolution and heuristic matching. In *ACL (Volume 2 : Short Papers)*, 2016.
- [7] Yang Liu, Chengjie Sun, Lei Lin, and Wang Xiaolong. Learning natural language inference using bidirectional lstm model and inner-attention. In *arXiv preprint arXiv :1605.09090*, 2016.
- [8] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *arXiv preprint arXiv :1703.03130*, 2017.
- [9] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Recurrent neural network-based sentence encoder with gated attention for natural language inference. In *arXiv preprint arXiv :1708.01353*, 2017.
- [10] Samuel R. Bowman, Jon Gauthier, Abhinav Rastogi, Raghav Gupta, Christopher D. Manning, and Christopher Potts. A fast unified model for parsing and sentence understanding. In *ACL*, 2016.
- [11] Tsendsuren Munkhdalai and Hong Yu. Neural tree indexers for text understanding. In *arXiv preprint arXiv :1607.04492*, 2016.
- [12] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*, 2017.
- [13] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Thomas Kocisky, and Phil Blunsom. Reasoning about entailment with neural attention. In *arXiv preprint arXiv :1509.06664*, 2015.
- [14] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced lstm for natural language inference. In *arXiv preprint arXiv :1609.06038*, 2017.

- [15] Vivian S. Silva, André Freitas, and Siegfried Handschuh. Building a knowledge graph from natural language definitions for interpretable text entailment recognition. In *LREC*, 2018.
- [16] Tim Miller. Explanation in artificial intelligence : Insights from the social sciences. *arXiv preprint arXiv :1706.07269*, 2017.
- [17] David Gunning. Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), 2017.
- [18] Christoph Molnar. *Interpretable Machine Learning*. Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, 2018.
- [19] 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.
- [20] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Diana Inkpen, and Si Wei. Natural language inference with external knowledge. *arXiv preprint arXiv :1711.04289*, 2017.
- [21] Marko Robnik-Sikonja and Igor Kononenko. Explaining classifications for individual instances. In *IEEE Transactions on Knowledge and Data Engineering*, 2008.