

Titre de l'article : Recursive Deep Models for Discourse Parsing

Auteurs :

Jiwei Li - Computer Science Department, Stanford University,
Rumeng Li - School of EECS, Peking University,
Eduard Hovy - Language Technology Institute, Carnegie Mellon University,

Nombre de pages : 10

Mots clés : analyse du discours, réseaux de neurones récurrents , unités élémentaires du discours (EDU), algorithme, deep learning, classification, représentation vectorielle distribuée.

Structure de l'article :

- 1.introduction
2. Related work :
 - 1 Discourse Analysis and Parsing
 - 2 Recursive Deep Learning
- 3.The RST Discourse TreeBank
4. EDU Model
5. Discours Parsing
 - 1 Binary (Structure) Classification
 - 2 Multi-class Relation Classification
 - 3 Distributed Vector for Spans
 - 4 Cost Function
 - 5 Backward Propagation
 - 6 Additionnal Features
 - 7 Optimization
 - 8 Inference
- 6 Experiments
 - 1 Parameter Tuning
 - 2 Baselines
 - 3 Performance
- 7 . Conclusion

Synthèse :

Comme le rappellent les auteurs de cet article, un texte est dit cohérent lorsqu'il est possible de décrire sans ambiguïté le rôle que joue chaque unité du discours (EDU) vis-à-vis de ce dernier, et ce à tous les niveaux de granularité. C'est cette structure hiérarchique, qu'essaient de reproduire les analyseurs du discours en identifier les relations discursives présentes dans les phrases.

Mais bien que l'on puisse obtenir un sens commun pour deux EDUs partageant une relation et ayant chacun un sens individuel, il semblerait que l'analyse de textes entiers soit plus complexe que l'analyse de simples phrases.

En effet, il n'y a pas toujours une définition syntaxique pour les relations discursives et les membres d'une relation peuvent être placés individuellement dans des phrases ou paragraphes différents. Ces relations peuvent alors être identifiées par des marqueurs linguistiques (temps, usage de pronoms, formes syntaxiques, conjonctions de coordination, etc).

Pour répondre à cette problématique, cet article présente un analyseur de discours s'appuyant

sur l'utilisation de réseaux de neurones récurrents, ayant montré une bonne efficacité dans leur capacité à capturer les éléments pertinents de syntaxe et de sémantique des EDUs à l'échelle de document entiers.

Le modèle d'analyseur de discours présenté repose sur l'usage de réseaux de neurones récurrents décidant si deux unités du discours sont connectées et par quelle relation, si tel est le cas.

Plus précisément, le système décrit prend en entrée un document de longueur variable pour en obtenir une représentation vectorielle distribuée pour chaque phrase.

Ensuite, est établie une classification binaire(réseau de neurones récurrents) pour déterminer la probabilité de fusionner deux EDU adjacents pour former un sous-arbre.

Puis est appliqué un classifieur multiclasse(réseau de neurones récurrents) pour sélectionner la relation discursive(ici la classe) la plus appropriée pour étiquetter le sous-arbre obtenu précédemment, et en calculer la représentation distribuée.

Ainsi se fait graduellement et récursivement une unification des sous-arbres jusqu'à couvrir totalement les phrases, puis le texte par un grand arbre. Il s'agit bien de calculer la structure arborescente décrivant le mieux les relations discursives d'un texte par l'approche «bottom-up».

On précise qu'un tel arbre a pour feuille des EDUs, pour noeuds internes des relations discursives.

En ce qui concerne l'algorithme d'apprentissage utilisé, on distingue les points suivants :

- Deux fonctions de coût pour chaque réseau (binaire et multiclasse)
- Une procédure de «backward propagation» standard (voir formule),
- Présence de features supplémentaires – un apriori (token en début ou fin d'EDU, POS en début ou fin d'EDU, Deux EDU dans la même phrase),
- Optimization: technique diagonale, variante de l'Adagrad très utilisé dans la communauté du deeplearning (voir formule),
- Inférence ou mise en relation des propositions (sous-arbres) : trouver l'arbre discursif parent le plus probable étant donné les sous-arbres fils (EDUs) à l'aide d'une table de programmation dynamique (voir formule).
- Classifieurs binaire & multiclasse entraînés séparément.

En ce qui concerne la technique d'évaluation employée, elle consiste en une comparaison de la structure et des labels des arbres structurels obtenus avec les structures et annotations «gold-standard» de la base de tests. La formule d'évaluation est la suivante : diviser le nombre de similarités dans les constituants des arbres par le nombre de constituants total.

Comme base de donnée est utilisé le Rhetorical Structure Theory Discourse Treebank (RST-DT) : il contient 385 documents (sous forme arborescente) annotés, dont 347 pour la phase d'entraînement et 38 pour de test. Cette base de donnée recense 110 relations de grains fins pour étiquetter les arbres contenus. Les relations non-binaires contenues sont cependant transformées en relations avec un branchement à droite (forme arborescente). Aussi, le système subit une validation croisée avec découpage en 5 tas.

Pour compléter la phase d'évaluation, les auteurs décident de comparer les performances de leur modèle avec d'autres analyseurs de discours concurrents que sont HILDA, celui de Joty et Al, de Feng et Hirst, ou encore de Ji et Eisenstein.

Contrairement aux modèles concurrents réalisant des scores optimaux sur quelques mesures d'évaluation, le modèle présenté dans cet article réalise un score moyen optimal sur l'ensembles des

mesures d'évaluation, sans toutefois atteindre un score optimal sur une seule mesure.

De plus, les auteurs attirent notre attention sur le fait qu'il faut prendre en considération différents types de relation dans le traitement de l'information (convolution) pour éviter de voir les performances s'effondrer.

Comme ouverture, les auteurs prévoient d'étendre l'application de leur modèle à des tâches telles que l'identification de relations implicites dans le discours ou encore l'analyse de dialogue.

Traoré Kalifou R.B.