Synthèse Bibliographique
Extraction de Relations

Matthieu RÉ et Yaohui WANG

January 8, 2017

Abstract

Nous présentons dans cette synthèse un résumé, une description de l'état de l'art, des principales méthodes (en les comparant) répondant aux problématiques d'extraction de relations.

Mots-clefs Extraction de relation, Traitement Automatique de la Langue

1 Introduction

L'extraction de relation est une discipline relative à l'extraction d'information de textes et à l'analyse sémantique des informations qu'ils contiennent. C'est une problématique très importante puisque connaitre les relations qui composent le contenu d'un texte non structuré peut s'avérer être un moyen rapide et précis d'extraire une information que l'on recherche, mais également de définir une structure au texte.

Nous reviendrons en détail sur ce qu'est une relation, mais pour illustrer la problématique, prenons un exemple. Admettons qu'il soit nécessaire de constituer un arbre généalogique des personnages présents dans le chapitre d'un livre. L'extraction de relations, et notamment des relations généalogiques nous permettra de faire ressortir une liste, d'une longueur égale au nombre de relations imaginées entre les différents protagonistes, et composées de listes de couples (ou pas forcément, selon les types de relations !) d'individus.

Ces analyses peuvent se faire entre des mots, des entités, des concepts, et d'autres notions dont l'unicité sémantique est à définir selon la relation. Nous allons commencer dans cette synthèse par un retour sur les thèmes et concepts importants que nous utiliserons par la suite, puis nous continuerons par un état de l'art orienté sur deux types de méthodes d'extraction de relations, avant de finir par les comparaisons de ces outils et une ouverture vers des pistes et des concepts envisageables pour le développement du sujet.

2 Présentation des notions, des problématiques de structuration de connaissances

2.1 Relation, dites-vous ?

Commençons par définir ce qu'est une relation en prenant un exemple.

"Gérard, qui est marié à Françoise, travaille à regret chez Cisco, à Lille, depuis qu'il a déménagé à Roubaix, il y a 20 ans."

De cet exemple on peut extraire plusieurs relations qui pourraient être intéressantes : par exemple, que Gérard est marié à Françoise (*Marié(Gérard, Françoise)*), que Gérard habite à Roubaix (*Habite(Gérard, Roubaix)* ou que Gérard a un travail chez Cisco à Lille depuis 20 ans (*Travail(Gérard, Cisco, Lille, depuis 20 ans)*).

L'exemple précédent fait par ailleurs ressortir le fait qu'une relation peut considérer 2 entités (on parle de relations binaires) mais également plus de deux (on parle de relations n-aires).

On peut extraire différents types de relations de cet exemple. On identifiera une liste non exhaustive de ces relations dans les prochains points.

Relation dans la phrase Le premier type de relations que l'on peut évoquer, est celle des relations grammaticales, dans une phrase. On peut ainsi considérer l'exemple de cette façon :

"Gérard, qui est marié à Françoise, travaille à regret chez Cisco, à Lille, depuis qu'il a déménagé à Roubaix, il y a 20 ans."

Vu ainsi, par rapport à la forme verbale qu'est "travaille", on peut identifier une forme nominale "Gérard" qui en est le sujet. Ainsi on peut définir des relations sur les fonctions syntaxiques des entités d'une phrase.

Relation entre mots ou entités clefs On peut également traiter la phrase de façon à identifier des lieux, des personnes, des dates, etc. Dans l'exemple :

"Gérard, qui est marié à Françoise, travaille à regret chez Cisco, à Lille, depuis qu'il a déménagé à Roubaix, il y a 20 ans."

On assimile donc les termes de la phrase à des tokens, des catégories sémantiques auxquels ils appartiennent (des personnes, des lieux, des institutions, des dates) qui nous permettent d'identifier des relations intraclasses (mariage entre personnes, hiérarchie entre personne, proximité entre des lieux, etc.), ou interclasses (personne salariée dans une institution, personne habitant dans un lieu, etc.).

Relations d'impression entre entités On peut également imaginer des relations plus complexes entre des entités, telles que l'affection que l'une porte à l'autre. Par exemple ici :

"Gérard, qui est marié à Françoise, travaille à regret chez Cisco, à Lille, depuis qu'il a déménagé à Roubaix, il y a 20 ans."

On voit ici que Gérard ne porte pas forcément Cisco dans son coeur, grâce à l'occurence du complément verbal à regret.

2.2 Formalisme sur les relations

Au travers de l'exemple que nous venons d'aborder, nous avons pu identifier qu'une relation peut prendre différents aspects. Tout d'abord, elle peut relier deux, ou plusieurs notions. Ensuite, elle peut relier plusieurs types d'informations : des entités (personnes, lieux, dates, etc.), des groupes de mots ayant la même fonction syntaxique (groupes verbaux, groupes nominaux, compléments d'objets, de lieu, de manière, etc.) et cette liste n'est pas exhaustive ! Nous aborderons dans les prochaines parties d'autres types de *termes* dont il est possible d'extraire des relations.

Si dans l'état actuel de la recherche les formes de relations binaires sont celles auxquelles on s'intéresse le plus, nous avons vu que des relations ternaires, ou plus longues peuvent être extraites. Les trois niveaux de relation que nous avons évoqué sont loin d'être les seuls auxquels nous nous intéresserons : on peut relever des relations plus lexicales (antonymes, synonymes), etc.

Nous pouvons ainsi, pour conclure cette partie, pointer que l'extraction de relation est un problématique qui dépend également de la possibilité d'extraire des informations : des groupes syntaxiques, lexicaux, des entités, etc. Aussi les performances de la relation d'extraction pourront être fortement influencées par celles de l'extraction des informations du texte.

3 Extraction à l'aide de participation humaine, ou de systèmes extérieurs

Une des complexités de l'extraction de relation est de construire un modèle qui soit assez général possible, et ainsi il est nécessaire d'avoir un nombre important de données d'entrainement. C'est pourquoi la construction manuelle, c'est-à-dire l'étiquetage manuel des relations comme cela a pu être réalisé dans des projets tel que Wordnet[1], une base de données de multiples types de relations, s'avère coûteuse et dans la pratique difficilement reproductible.

Nous allons donc dans cette partie nous intéresser à des modèles collaboratifs, qui puisent leur information de données issues de manière collaborative, ou de différents systèmes extérieurs. Dans les deux cas, on s'intéresse à des informations issues de sources différentes, de sorte à créer une sorte de pivot qui constitue une réponse supposée plus fiable.

3.1 Participation indirecte par jeux

Dans le premier article de notre bibliographie[4], il est évoqué un moyen inédit de participer à la création d'une base de données de relation. Le travail manuel était long pour une seule personne, il est proposé à plusieurs personnes, sous la forme d'un jeu. La finalité de ce jeu est de collecter des relations lexicales (synonymie, antonymie, etc.) entre des mots.

La base de données, évolue ainsi non seulement en nombre, mais également en qualité. En effet, plutôt que de ne pas vérifier les réponses jouées, ou de les faire modifier par quelqu'un, ce sont les joueurs qui sont amenés à corriger et améliorer les réponses apportées par les autres joueurs.

Le jeu consiste en la recherche simultanée, par deux personnes dans une partie, d'une série de mots issus d'une consigne (synonymes, contraire, etc.) en rapport à un mot tiré aléatoirement de la base. Deux joueurs jouent indépendamment, mais essayent d'avoir le plus de réponses communes pour gagner plus de points. Lorsque les deux joueurs proposent un mot qui n'existe pas dans la base, alors ce mot est ajouté. C'est ainsi que la base de mot grandit.

Un système de pondération et de réponses tabous, c'est-à-dire des réponses interdites parce que trop communes, permet de contrôler le nombre de réponses différentes (de façon à avoir une plus grande diversité de relations lexicales concernant un mot, sans en avoir trop). Un système accordant de la confiance à un joueur permet que ses réponses interviennent plus dans la base de données.

Conclusion Ce que l'on peut conclure du système, avant d'évoquer ses performances dans la partie de comparaison des méthodes, est qu'il met en oeuvre un moyen ludique de participer à la création d'une base de données de relation plutôt fiable, et qui se base sur un système collaboratif intéressant.

D'autres méthodes permettent de se servir de systèmes extérieurs pour enrichir une base de données de relation. On a vu que le modèle JeuxDeMots permet de construire une base assez petite, mais plutôt fiable ; une idée serait de combiner des modèles pour en former un plus gros, et plus fiable. C'est une approche par *fusion*. Mais cette méthode peut s'avérer plus couteuse que celle qui est abordée pour le modèle suivant.

3.2 Approche par traduction

Le but de cette approche est de construire, à l'aide de données diverses (corpus, Wikipedia, etc.) et multilingues (français, anglais, roumain, tchèque et bulgare), l'équivalent d'un Wordnet en français : WOLF[1].

[1] https://wordnet.princeton.edu/

L'approche utilisée dans cette article est l'approche par *traduction*. Comme son nom l'indique, elle permet de construire une base de données de relations à partir de bases existantes dans différentes langues. Seulement cette approche implique d'avoir à résoudre certains problèmes au préalable. Un de ceux-ci est la polysémie des mots, le fait que les mots puissent avoir plusieurs sens dans ces langues.

L'alignement des traductions, i.e l'identification d'un sens commun des traductions de ce mots dans au moins deux langues, permet de contourner ce problème. Il consiste en la comparaison des traductions d'une expression dans différentes langues pour avoir dans une langue une liste d'expressions pivots, mettant en exergue les différents sens de l'expression. Par chance c'est le cas de peu de mots dans les 5 corpus de langues utilisés, et la lemmatisation (i.e le fait de ne garder qu'une racine des mots) ont permis l'application de cette technique.

Pour améliorer la transformation, les corpus ont été approchés par voisinage linguistique : de deux corpus de langues proches (français-roumain, et tchèque-bulgare) en sont sortis cinq lexiques multilingues. Reste à associer le contenu des lexiques par groupes de synonymes, en les rapprochant par similarité au lexique français. Une seconde approche a été d'utiliser les traductions de termes en anglais-français de Wikipédia pour créer des synsets. Les ensembles obtenus ont été enrichis ensuite par les premières phrases des articles, qui renferment souvent des synonymes. De la même façon, le dictionnaire Wikitionary, et l'encyclopédie des espèces vivantes Wikispecies (et ses termes latins) ont été utilisées. EUROVOC, bibliothèque des activités de l'UE, donc multilingue, a également été fusionné aux autres synsets créés lors des différentes constructions, en conservant certaines informations pointant la qualité des lexèmes définis, et qui servent ensuite à l'amélioration du résultat. Il subsiste néanmoins des lexèmes sans équivalent français, qui sont conservés dans WOLF.

Conclusion Cette approche offre de bons résultats, mais nécessite encore une forte présence de l'humain et est peu évolutive automatiquement (quand ses sources changent, il faut réaliser à nouveau l'approche pour reformer une nouvelle base). C'est pourquoi dans la partie d'après sont évoquées des méthodes automatique d'extraction des relations, qui offrent plus de possibilités pour des performances compétitives.

4 Extraction automatique

Comme nous l'avons vu, une tâche d'automatisation est nécessaire pour réaliser ce travail long et coûteux. L'idée de l'extraction de relation est donc de se baser sur d'autres travail de détection et d'extraction automatique d'information (POS-taging, détection d'entités nommées) de façon à pouvoir ensuite par exemple (nous le verrons dans le cas des SVM) utiliser ces informations pour réaliser de l'extraction automatique de relations.

On peut donc voir l'extraction de relation comme un double travail de classification, supervisée comme non supervisée. Les données d'apprentissage nous permettent d'estimer deux (ou plus) entités entrent dans une quelconque relation entre elle, pour ensuite permettre de regrouper des couples (pour les relations binaires) ou des n-uplets (pour les relations mettant en jeu plus de deux entités) dans des clusters de couples/n-uplets de même relation.

Nous allons maintenant présenter deux de ces méthodes qui permettent de réaliser l'extraction automatique de relations : l'une basée sur des SVM, l'autre sur des réseaux de neurones convolutifs.

4.1 SVM

Comme rappelé plus tôt, nous réalisons ici deux tâches indépendantes l'une après l'autre : la détection d'une quelconque relation et la classification d'une éventuelle relation[3].

Détection de relation Elle consiste à identifier pour chaque pair d'entité (dans le cas de relations binaires auquel on s'intéressera spécifiquement pour toute cette partie) la possibilité d'appartenir à une relation définie (peu importe laquelle pour l'instant).

Relation classification Chaque relation constitue une classe spécifique des relations détectées. Pour chaque donnée (couple d'entité), on associe un certain nombre de features pour réaliser le travail de classification. Une fois ces features calculées, on applique un clustering multi-classes à l'aide de SVM. Nous reviendrons plus tard sur ces features.

He assigns a specific class to each detected relation. To each task, we apply distance linguistic features ranging from lexical tokens to syntactic structures as well as the semantic type information of the entities. We also apply the distance between two entities to make the detection problem easier and to increase the performance of both the relation detection and classification. We use SVM to avoid overfitting.

Définition du problème

$$(e_1, e_2, s) \rightarrow r$$

e_1 et e_2 sont deux entités présentes dans la phrase s, et r est le label de la relation.

Dataset Pour les tests et l'entrainement du modèle qui seront opérés, la solution utilise le corpus ACE[2]. Ce corpus contient 5 types d'entité : des personnes, des organisations, des entités géopolitiques, des lieux et des entreprises. Les relations définies se répartissent dans 5 catégories : les relations de rôle, de participation, d'appartenance, de proximité et des relations sociales. Pour ce qui est du pré-traitement des données, il est en plusieurs étapes, dont la délimitation des phrases en segments uniques (à l'aide de la méthode développée lors de la compétition DUC), de découpage de fragments de phrases à l'aide du Charniak Parser, et de chunking.

Features 10 features sont proposées ici. Chacune apporte une information sur un point essentiel de l'analyse des relations entre deux entités, parmi lesquelles les informations sur les mots (entités elles-mêmes, mots les séparants dans la phrase), type de discours, grammaticale (rôle dans la phrase), topologiques (distance entre les mots), syntaxique, sémantiques (tagging (type d'entité), chunking (technique permettant d'identifier les entités avec des fonctions et par groupe de mots)).

Ces features ne sont pas toutes nécessaires dans les deux types de classification, aussi on choisit de ne prendre que des combinaisons de certaines dans les deux problèmes. Pour le premier problème de détection de relation, il n'est d'ailleurs pas nécessaire d'utiliser des techniques de TAL pour avoir des features satisfaisantes (les mots, le types d'entité, sa grammaire (pronom, forme nominale,...) et sa position dans la phrase).

Finalement, on met en place une classification par SVM, avec un noyau gaussien (RBF). Le noyau gaussien est en général (et dans ce cas là en particulier) le plus efficace. On pourrait penser à utiliser des méthodes plus modernes de combinaison de noyaux, qui pourraient améliorer les performances du modèle.

4.2 Réseaux de Neurones Convolutifs

L'extraction de relation par CNN se fait en trois étapes[2].

Représentation des mots Afin de normaliser les mots, le CNN utilise la méthode d'embedding Word2Vec[3]. Cette méthode donne une projection de chaque mot dans un espace de représentation normalisé de manière automatique (par le biais d'un réseau de neurones).

Feature Extraction Dans cette étape, plutôt que d'utiliser la représentation "brute" donnée par Word2Vect, on calcule des features représentant deux types d'information : une information lexicale et une information sur la phrase.

[2]ACE 2005 Multilingual Training Corpus (https://catalog.ldc.upenn.edu/LDC2006T06)
[3]https://deeplearning4j.org/word2vec

Lexical level feature Ces features syntaxiques concernent :
1. La première entité
2. La seconde entité
3. Les tokens voisins (gauche et droite) de la première entité
4. Les tokens voisins (gauche et droite) de la seconde entité
5. Les hypernymes de ces entités, c'est à dire leur catégorie ou leur concept, calculés en utilisant WordNet

Sentence level feature On calcule cette feature en trois étapes, constituant les couches cachées d'un (autre) réseau de neurones convolutif. D'abord on détermine une fenêtre de "normalisation" dans la phrase, basée sur des features des mots en soi (Word Feature) et de leur position (Position Feature). Ensuite, la représentation passe par une couche cachée de réseau avant de donner en sortie la Sentence Level Feature, après une transformation non-linéaire.

Final Feature On combine finalement ces deux features pour créer la feature unique qui représente les deux entités. À l'aide d'une classification issue d'un softmax, il est finalement possible de mesurer la qualité d'une feature[4] par rapport à une relation. Cela permet de donner un degré de confiance à la décision finale, qui est donnée par le meilleur score de confiance des features par rapport aux relations.

En terme de mise en place cette méthode est plus gourmande que celle vue auparavent. Elle nécessite un plus grand nombre de données d'entrainement, et un traitement supplémentaire (entrainement de chaque mots par Word2Vec, utilisation de nombreux outils de TAL (WordNet & NLTK,...)) qui n'est pas nécessaire par exemple pour la détection de relation à l'aide de SVM.

5 Comparaison des méthodes d'extraction de relations

La méthode utilisée dans les articles du corpus dans ses évaluations de performance revient dans la plupart des cas à comparer les résultats des techniques utilisées à des données qui font office de références. Nous allons par exemple évoquer les quatre techniques qui diffèrent dans chaque texte.

JeuxDeMots La méthodologie d'évaluation du premier article est issu de la comparaison à l'Euro Wordnet Français (EWF)[5]. Si le nombre de relation est beaucoup moins important que ce dernier, il est vraisemblablement croissant avec le nombre de joueur, et sa précision (97%) n'est dûe qu'à des erreurs d'orthographes de joueurs, ou des malentendus. Elles seraient moins nombreuses avec un plus grand nombre de joueurs. Le système n'est donc pas aussi parfait qu'un système d'annotation manuelle, mais ses performances restent pour le moins honnêtes.

WOLF Comparé également à l'EWN, WOLF contient plus de synsets et de lexèmes que ce dernier, et couvre une plus large partie des concepts de base définis dans le projet BCS[6], notamment grâce à sa couverture d'adjectifs et d'adverbes. La comparaison plus poussée à l'EWN français montre que les concepts principaux sont communs aux deux, mais que les synsets correspondant à des polysémies rares sont moins fréquents dans l'EWN.

SVM Les performances de ce systèmes sont mesurées à l'aide de la précision, du rappel et de la F-mesure qui en découle. La précision est le ratio de résultats corrects rendus par rapport aux résultats rendus, et le rappel et le ratio de résultats corrects rendus par rapport aux nombre d'exemples corrects.

Convolutional Neural Network L'évaluation de ce modèle se fait également à l'aide d'un ensemble de données référent, le SemEval-2010 Task 8 dataset. Pour ce qui est des classements et des calculs de performance, ils se font à l'aide d'une F-mesure (la F_1-mesure) sur 9 relations.

[4] On peut assimiler ce score à une probabilité conditionnelle
[5] $http://catalog.elra.info/product_info.php?products_id = 550$
[6] Basic Concept Sets, projet BalkaNet, 2000

Les résultats que l'on obtient sont meilleurs en terme de performance pour les réseaux de neurones, mais la complexité du modèle et sa lenteur lui font défaut pour l'instant. Les SVM sont un moyens pour l'instant plus simple et plus rapide d'obtenir des résultats corrects, en tout cas sur les données de test que nous avons étudiées.

Pour conclure cette partie, nous pouvons noter que l'état de l'art actuel tend à privilégier les réseaux de neurones aux méthodes plus classiques d'apprentissage artificiel. En effet si dans le cas étudié ici le CNN utilise des features encore assez concrètes, l'avantage des réseaux de neurones est de minimiser la perte d'information que la seule réduction des données du problème à une projection sur un espace de feature apporte. Autrement dit, l'effet "boite noire" du réseau de neurones permet, si tant est qu'il soit bien construit, de minimiser la perte d'information que les autres méthodes peuvent induire et donc d'augmenter la précision des prédictions.

6 Conclusion

L'extraction de relation est une problématique très complexe, mais fondamentale. On peut notamment penser au développement d'outils, qui par exemple souhaiteraient répondre à des questions telles que "Qui est la fille de Gérard ?" ou "À quelle heure arrivera le prochain train en partance de Saint-Étienne ?". Les techniques qui existent aujourd'hui, qui sont principalement des méthodes faisant appel à l'homme (coûteuses et longues à mettre en place), ou des méthodes d'apprentissage supervisé et semi-supervisé.

Par le biais des exemple du corpus, nous avons notamment pû revenir sur certain des problèmes qui touchent les méthodes actuelles : la difficulté de représentation des données, de faire des ressources multilingues, de faire trop confiance à l'homme - ou à la machine.

Le développement des réseaux de neurones promet cependant un nouvel essor des techniques d'extraction de relations. Il y a moins d'un mois, Google, par sa filiale DeepMind, mettait au point ce qu'ils ont appelé des *Differential Neural Computers*, permettant un apprentissage non supervisé plus puissant via des réseaux de neurones complexes[7], capable d'enregistrer mais aussi d'apprendre de nouvelles relations sur la base de nouvelles données.

L'état de l'art que nous avons essayé de reproduire via l'études des 4 textes de notre corpus, risque fort de rapidement s'étoffer. Ajoutons pour conclure que si dans les années 2000 les recherches en extraction de relation semblaient démarrer, en se concentraient sur les entités nommées[8], elles continuent toujours activement de nos jours et constituent toujours des sujets intéressants de recherche[9].

References

[1] Darja Fiser Benoît Sagot. Construction d'un wordnet libre du français a partir de ressources multilingue. *Traitement Automatique des Langues Naturelles, Jun 2008, Avignon, France*, TALN 2008, 2008.

[2] Siwei Lai Guangyou Zhou Jun Zhao Daojian Zeng, Kang Liu. Relation classification via convolutional deep neural network. *Proceedings of COLING, Aug 2014, Vancouver, Canada*, COLING 2014:2335–2344, 2014.

[3] Gumwon Hong. Relation extraction using support vector machine. *Proceedings of the Second international joint conference on Natural Language Processing, Oct 2005, Jeju Island, Korea*, IJCNLP 2005:366–377, 2005.

[7] https://deepmind.com/blog/differentiable-neural-computers/

[8] Les quatres textes en référence sont tirés de ce mouvement

[9] SemEval revient fréquemment sur des Tasks relatives à la relation d'extration, comme en 2010 (2010-Task8 SemEval http://www.aclweb.org/anthology/S10-1006), en 2015 (2015-Task17 SemEval http://alt.qcri.org/semeval2015/task17/), etc.

[4] Alain Joubert Mathieu Lafourcade. Jeuxdemots : un prototype ludique pour l'émergence de relations entre termes. *Journées internationales d'Analyse statistiques des Données Textuelles, Mar 2008, France.*, JADT'08:657–666, 2008.