Synthèse

Introduction

Nous aborderons dans cette synthèse la problématique de l'extraction de relations. Elle est essentielle dans un monde où la quantité d'informations ne fait que croître. L'information que l'on recherche devient alors plus difficile d'accès et les méthodes d'extraction permettent de retrouver plus facilement les liens entre les “entités nommées”. On y trouve différents moyens d'extraire des relations sémantiques dans les documents que l'on va explorer dans le corps de cette synthèse.

L'extraction d'information :

- ne cherche plus à comprendre les textes dans leur ensemble
- vise à extraire d'un texte donné des éléments pertinents

Le type d'information pertinente pour une application donnée est défini à l'avance par le modèle. Seule une partie du texte est considérée (un faible pourcentage de texte utile pour une tâche spécifique).

Définitions

- Extraction de relations

L'extraction de relations consiste à retrouver des différents liens entre entre lestermes du texte, généralement des entités nommées ou groupes nominaux. Par exemple, la phrase “Paris est la capitale de la France” relie la capitale à son pays associé (grâce au motif “A est *la capitale de* B”). L'extraction de relation consiste, en d'autres termes, à extraire des relations sémantiques entre les différents termes et à les associer explicitement (comme par exemple, associer Paris à la France, avec comme nature du lien “la capitale”). En d'autres termes, on associe deux termes en relations dans un triplet (terme1, nature relation, terme2).

- Extraction d'informations

Permet d’extraire une information à une question posée. Elle consiste à générer des données structurées à partir de données textuelles donc semi structurée voire non structurées. Les outils d'extraction d'informations permettent de récupérer des informations dans des documents textuels, des bases de données, des sites Web ou des sources diverses et sont très utiles dans l'analyse de descriptions de produits, analyse de bibliographies, synthèse de textes journalistiques etc.

- Relations contextuelles (thématiques)

On parle de relations contextuelles lorsque deux entités sont reliées entre elles, précédés d’hypothèses sur l'état de leur lien. Plus simplement, un contexte est l’information qui caractérise les interactions entre humains, applications et l’environnement. On ne cherche plus seulement à extraire une relation entre deux entités mais aussi à connaître le cheminement de leur relation et leur nature.

- WordNet (réseau lexical)

WordNet est une grande base de données lexicale en anglais. Les noms, verbes, adjectifs et adverbes

sont regroupés en ensembles de synonymes (appelés aussi synsets), chacun exprimant un concept distinct. Les synsets sont liés entre eux au moyen de relations conceptuelles, sémantiques et lexicales. (WordNet). Par exemple, le mot “idéal” est en relation avec le mot “parfait” car ce sont des synonymes. On peut même pondérer la pertinence de leur relation pour savoir si ces mots sont très proches ou non.

- Entités nommées

Dans l'extraction d'informations, une entité nommée est un objet du monde réel, tel que des personnes, des lieux, des organisations, des produits, etc., pouvant être désigné par un nom propre. Cela peut être abstrait ou avoir une existence physique. Par exemple, des personnes physiques ou fictifs, reconnaissable en général par des noms propres sont un ensemble d'entités nommées de type “PERSON”.

Difficulté de la tâche

L’extraction de relations peut s’avérer ardue à implémenter pour plusieurs raisons. L’une d’entre elles concerne l’ambiguïté du langage qui peut intervenir et changer complètement le sens de la relation entre deux entités. Dans le cadre de la relation lexicale, on ne s’en occupe pas mais au niveau contextuel, la sémantique devient importante. La phrase “j’ai mangé un avocat. Il semblait périmé” désigne le fruit et non la personne exerçant le métier d’avocat. Le verbe “manger” aide à comprendre le sens du terme “avocat”. La relation (manger, avocat, périmé) signifie donc que le fruit est périmé, et non la personne exerçant ledit métier.

L’enjeu de tout ceci est donc d’automatiser ce genre d’extraction en faisant une désambiguïsation sémantique (on associe un sens à une entité nommée) avec l’aide de WordNet qui fournit une longue liste de termes sémantiques.

Les problèmes sont à la fois théoriques et pratiques

- Grammaire complexe (avec des exceptions en pratique dans le monde réel)
- Ambiguïtés non levées trop nombreuses
- Difficultés pour collecter, mais aussi pour manipuler les connaissances sémantiques et pragmatiques suffisantes
- L'approche générique de compréhension est pour l'instant une utopie

Différentes approches

Relations lexicales

Les méthodes d’extraction de relation au niveau lexical omettent le contexte et toute la sémantique par aux termes du texte étudié. On se concentre principalement sur de la relation basé principalement sur le langage lui-même, sa constitution et son lexique (le vocabulaire).

Similarité relationnelle

La similarité relationnelle renseigne sur les relations syntaxiques et structurelles entre les différents

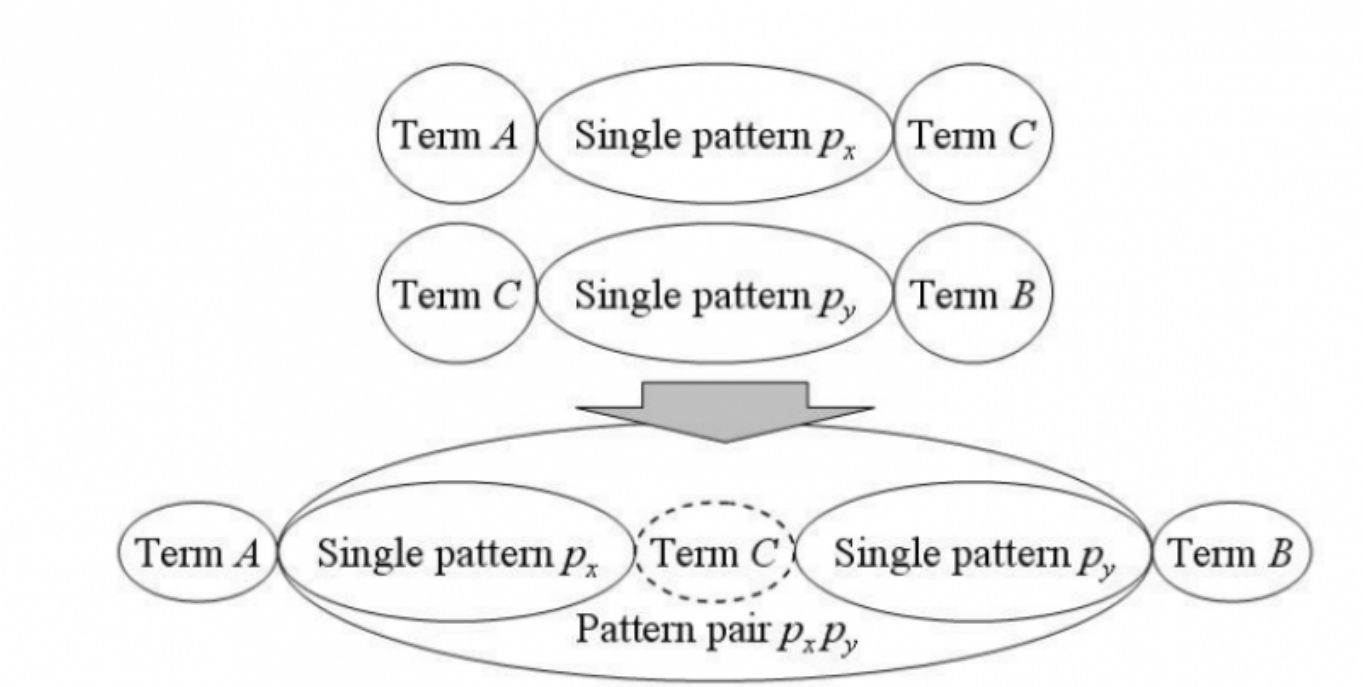
mots du texte. Il y a plusieurs façons de catégoriser les relations. La synonymie est la similarité sémantique entre deux mots. On peut donc dans une phrase assimiler une entité avec une autre. L'hyponymie quant à elle consiste en un l'inclusion d'un mot dans une catégorie plus large ("chat" est un hyponyme de "félin" qui est un hyponyme de "animal"). On peut ainsi plus aisément deviner des relations entre les entités.

L'openIE est également un moyen de caractériser la relation dans une pair d'entités nommées. Par exemple, "La Finlande a battu le Canada" devient "aBattu(Finlande, Canada)".

Extraction de motifs par "Is-a"

On construit une relation entre une paire d'entités à partir d'un pattern (motif). A partir de cela, on va réaliser une extraction de relations sémantiques. On crée pour cela une liste de motifs bien à l'avance dans le programme, puis on récupère la pair de termes satisfaient le motif. Un pattern basique comme "X est un Y" est vérifiée par la phrase "un chien est un animal" qui relie les termes "chien" et "animal".

Le procédé est également transitif comme le montre le schéma ci-dessous.



L'algorithme Espresso (de Pantel et Pennacchiotti, 2006) utilise un algorithme de type bootsrapping qui a de bonnes performances. Cet algorithme extrait des motifs apparaissant entre certains des paires de termes préparés qui satisfont une relation cible spécifiée.

Relations contextuelles

GraphRel et relations en arbres

La contribution sur l'outil est multiple: d'une part, les relations avec les entités sont importants en partie parce que elles sont pondérées, d'autre part, l'outil permet d'extraire des caractéristiques

cachées des paires d'entités. Chaque entité est liée, d'une manière ou d'une autre, dans un arbre entre un arc pondéré. On utilise ici une méthode de deep learning pour une nouvelle approche d'extraction de relation contextuelle dans le but de ressortir le sens de la relation entre chaque entité.

Cause à effets

L'identification d'une relation causale permet de répondre aux question suivantes : Que se passe-t-il ? Pourquoi est-ce arrivé et que va-t'il se passer ensuite ? Pour y répondre, il faudra tout d'abord détecter les indicateurs temporels qui précédent l'effet, donc la cause de l'événement considéré.

La méthode utilisée est une classification supervisés utilisant des relations lexicales et sémantiques. La première étape vers la classification de relation causale est l'identification des paires d'événements candidats. Étant donné un document déjà annoté d'événements, nous prenons en compte chaque combinaison possible d'événements dans une phrase de manière avancée en tant que paires d'événements candidats. Par exemple, si nous avons une phrase «e1, déclenchée par e2, a causé e3», les paires d'événements candidats sont (e1, e2), (e1, e3) et (e2, e3). Nous incluons également, comme paires candidates, la combinaison de chaque événement dans un phrase avec les événements dans la suivante, pour tenir compte de la causalité entre les phrase, dans le cadre de la simplification d'hypothèses que la causalité peut également être exprimée entre événements de deux phrases consécutives.

Evaluations et résultats

La plupart du temps, on évalue les performances d'un système avec le triplet classique: rappel, précision et F1-score. On effectue des tests de performances avec des corpus annotés à l'avance et qui permettent de rendre compte de la précision du programme d'extraction de relation. Cela permet également d'alimenter en données un programme basé sur l'apprentissage automatique.

Conclusion

L'extraction de relations a énormément évolué ces dernières années et ses secteurs d'activités sont divers, partant de l'extraction d'information, à des moteurs de recherche élaborés. De plus, grâce à l'avènement du deep learning ses dernières années, les programmes arrivent à mieux à appréhender les subtilités du langage et sont donc plus performants et plus précis lors du résultat. Cette thématique est en plein essor et va tendre à s'améliorer en partie avec la croissance très rapide des données exploitables pour les systèmes à machine learning.

Bibliographie

- https://www.aclweb.org/anthology/S17-2091.pdf (keyword)
- https://www.aclweb.org/anthology/Y09-1028.pdf (pattern)
- https://www.aclweb.org/anthology/P19-1136.pdf (graphrel)
- https://www.aclweb.org/anthology/C16-1007.pdf (Cause-effet)