TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Une approche faiblement supervisée pour l'extraction de relations à large échelle

Ludovic Jean-Louis Romaric Besançon Olivier Ferret Adrien Durand
CEA, LIST, Laboratoire Vision et Ingénierie des Contenus
Fontenay-aux-Roses, F-92265, France.
{ludovic.jean-louis,romaric.besancon,olivier.ferret,adrien.durand}@cea.fr

Résumé. Les systèmes d'extraction d'information traditionnels se focalisent sur un domaine spécifique et un nombre limité de relations. Les travaux récents dans ce domaine ont cependant vu émerger la problématique des systèmes d'extraction d'information à large échelle. À l'instar des systèmes de question-réponse en domaine ouvert, ces systèmes se caractérisent à la fois par le traitement d'un grand nombre de relations et par une absence de restriction quant aux domaines abordés. Dans cet article, nous présentons un système d'extraction d'information à large échelle fondé sur un apprentissage faiblement supervisé de patrons d'extraction de relations. Cet apprentissage repose sur la donnée de couples d'entités en relation dont la projection dans un corpus de référence permet de constituer la base d'exemples de relations support de l'induction des patrons d'extraction. Nous présentons également les résultats de l'application de cette approche dans le cadre d'évaluation défini par la tâche KBP de l'évaluation TAC 2010.

Abstract. Standard Information Extraction (IE) systems are designed for a specific domain and a limited number of relations. Recent work has been undertaken to deal with large-scale IE systems. Such systems are characterized by a large number of relations and no restriction on the domain, which makes difficult the definition of manual resources or the use of supervised techniques. In this paper, we present a large-scale IE system based on a weakly supervised method of pattern learning. This method uses pairs of entities known to be in relation to automatically extract example sentences from which the patterns are learned. We present the results of this system on the data from the KBP task of the TAC 2010 evaluation campaign.

Mots-clés : extraction d'information, extraction de relations.

Keywords: information extraction, relation extraction.

1 Introduction

Dans le cadre de l'extraction d'information, l'extraction de relations est un processus dont l'objectif est de déterminer l'existence d'un lien sémantique entre deux entités et lorsque cela est possible, de caractériser la nature de ce lien. Nous nous intéressons plus particulièrement dans cette étude à l'extraction de relations entre entités nommées en vue de la collecte et de la construction d'une base de connaissances à large échelle. En effet, on trouve dans des sources d'informations ouvertes, en particulier dans le contexte du Web sémantique, un grand nombre d'informations disponibles sous forme semi-structurée : par exemple, l'encyclopédie Wikipédia contient des informations qui peuvent être structurées sous forme d'une base de données, comme le montre le projet DBpedia[1] (Bizer *et al.*, 2009). Cette structuration première des informations semi-structurées peut alors être complétée par l'extraction automatique de relations entre entités à partir de texte brut.

Les travaux ayant pour objet l'extraction de relations peuvent être considérés selon l'angle du degré de supervision qu'ils requièrent. Au degré le plus faible, que l'on qualifie d'approche non supervisée, le type des relations à extraire n'est pas défini *a priori*, que ce soit par le biais d'exemples ou d'un modèle. Tout au plus peuvent être fixées certaines contraintes sur les entités reliées, comme leur type par exemple. Le type des relations extraites est quant à lui défini *a posteriori*, en regroupant les relations jugées similaires. Une telle approche est mise en œuvre dans (Shinyama & Sekine, 2006) ou dans (Banko & Etzioni, 2008) par exemple. À l'autre extrême de cette échelle, le type des relations visées mais aussi les moyens de les extraire à partir des textes sont définis *a priori*. Cette approche dite supervisée se caractérise soit par la donnée d'un modèle élaboré manuellement, typiquement sous la forme de règles, soit par l'association d'un ensemble d'exemples de relations en contexte issus de l'annotation d'un corpus et d'un algorithme d'apprentissage permettant d'en construire automatiquement un modèle. Cette seconde option est dominée par les modèles d'apprentissage statistique, qui se focalisent sur la prise en compte d'un large spectre de caractéristiques de différents types (lexicales, syntaxiques, sémantiques ...) (Zhou *et al.*, 2005) et sur l'élaboration de fonctions noyaux permettant de prendre en compte ces caractéristiques, en particulier lorsqu'elles ont des structures complexes comme celles produites par l'analyse syntaxique (Zhou *et al.*, 2007).

Entre ces deux pôles se trouvent les approches dites faiblement supervisées, vocable recouvrant l'idée que des exemples ou un modèle sont fournis pour le développement du système d'extraction de relations mais que cette seule contribution n'est pas suffisante pour la réalisation d'un système pleinement opérationnel. De ce fait, elle doit être étendue de manière automatique, généralement en exploitant un corpus non annoté. Les travaux existant en la matière font apparaître deux cas de sous-détermination de la contribution initiale, cas pouvant être éventuellement associés :

– une sous-détermination liée au volume de cette contribution. Seul un petit ensemble de relations exemples ou un modèle incomplet sont fournis ;
– une sous-détermination liée à la nature de la contribution initiale, ce qui se produit lorsque les exemples ou le modèle doivent être instanciés pour être utilisés.

Le premier cas de figure est typiquement traité suivant la méthodologie initiée par Hearst (1992) grâce à un mécanisme d'amorçage exploitant le petit ensemble initial d'exemples de relations ou de règles d'extraction pour acquérir de nouveaux exemples à partir d'un corpus et venir ainsi enrichir progressivement le modèle des relations visées au fil de cycles successifs d'application de ces deux étapes. (Agichtein & Gravano, 2000) en est un représentant typique pour les relations entre entités nommées. Bien qu'opérant dans un champ différent – l'extraction de structures qualia – (Claveau & Sébillot, 2004) offre un autre exemple d'amorçage pour l'induction de patrons linguistiques en combinant deux systèmes aux caractéristiques différentes.

Le second cas de figure est quant à lui illustré par la notion récente de « Distant supervision », introduite formellement par (Mintz *et al.*, 2009) mais déjà présente dans certains travaux sur l'amorçage. Les exemples sont ici donnés sous une forme sous-déterminée puisque réduite à un couple d'entités : ils sont donc à la fois privés de contexte et de caractérisation linguistiques. Le développement de ce type d'approches est favorisé par la mise à disposition de larges bases de connaissances extraites de ressources telles que Wikipédia.

Dans cet article, nous présentons un système d'extraction d'information à large échelle fondé sur un apprentissage faiblement supervisé de patrons d'extraction de relations reposant sur des exemples sous la forme de couples d'entités. Ces couples sont projetés dans un corpus de référence pour constituer la base d'exemples de relations à partir

[1] `http://dbpedia.org/About`

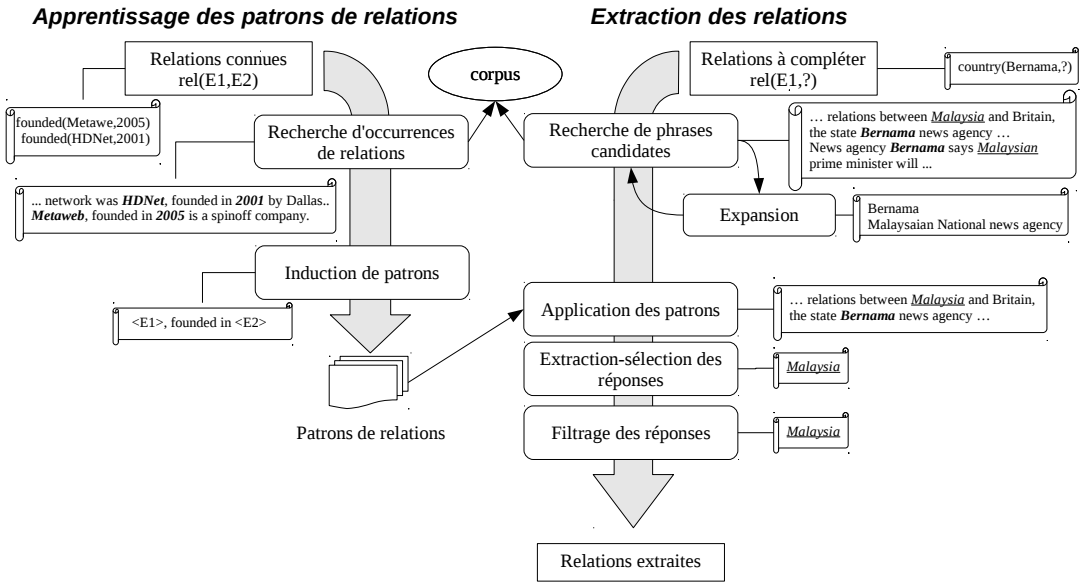

FIG. 1 – Architecture générale du système

de laquelle les patrons d'extraction sont appris. Ce travail se rattache donc au concept de « Distant supervision ». Nous présentons également les résultats de l'application d'une telle approche dans le cadre d'évaluation défini par la tâche KBP (Knowledge Based Population) de l'évaluation TAC 2010 (Text Analysis Conference).

2 Présentation de l'approche

Nous nous concentrons dans notre approche sur l'extraction de relations à large échelle en supposant la préexistence d'une base de connaissances partiellement remplie, extraite automatiquement à partir de données semi-structurées. Nous nous limitons ici aux relations entre entités nommées étant donné que, n'intervenant pas en domaine de spécialité où la recherche des entités peut être guidée par une terminologie existante, nous avons volontairement choisi de nous focaliser sur des entités aisément identifiables. La notion de « large échelle » se décline quant à elle selon plusieurs dimensions. La première réside dans le grand nombre de types de relations différents considérés, induisant une mise en œuvre difficile pour une approche à bases de règles écrites manuellement. La deuxième est liée à la prise en compte initiale d'un grand nombre de relations existantes (c'est-à-dire l'association de deux valeurs d'entités à un type de relation) ; ces relations fournissent un bon ensemble de départ pour l'apprentissage automatique d'un modèle de ces types de relations. Enfin, le corpus dans lequel de nouvelles relations sont recherchées est lui-même important, ce qui implique l'utilisation de techniques d'indexation et de recherche pour extraire des bons candidats (on ne peut pas envisager l'application directe de patrons sur toutes les phrases du corpus). Cette approche, illustrée par la figure 1, s'articule en deux phases : une phase d'*apprentissage de patrons* à partir d'occurrences de relations connues et une phase d'*extraction de relations* pour la découverte de nouvelles relations. La première phase part des relations connues *R(E1,E2)* pour trouver des occurrences de ces relations dans un corpus, c'est-à-dire les différentes expressions de cette relation dans les textes et utiliser ces occurrences pour induire des patrons de reconnaissance pour le type de relation concerné. La seconde phase part de relations incomplètes *R(E1,x)*, où l'entité source *E1* est connue et l'entité cible *x* est à trouver, cherche des occurrences de relations impliquant *E1* dans un corpus, puis extrait l'entité *x* en utilisant les patrons induits dans la première phase. Ces deux phases sont détaillées dans les sections suivantes.

2.1 Apprentissage des patrons

L'apprentissage des patrons de relations repose sur l'induction (ou généralisation) de patrons lexicaux à partir de phrases exemples contenant des occurrences des relations considérées. L'objectif de cet apprentissage est de

capturer les différentes expressions d'une relation sémantique entre deux entités. Par exemple, les deux extraits de phrases ci-dessous contiennent des occurrences de relations pour le type *founded_by*, instancié pour les couples d'entités (Charles Revson ; Revlon Cosmetics) et (Mayer Lehman ; Lehman Brothers investment).

The glamourous cabaret chanteuse reportedly had had a romantic liaison with <source>Charles Revson</source>, the founder of <cible>Revlon Cosmetics</cible> ... – Lehman was a great-grandson of <source>Mayer Lehman</source>, a founder of the <cible>Lehman Brothers investment</cible> house ...

Plusieurs travaux présentent des algorithmes de généralisation de patrons lexicaux (Ravichandran, 2005; Schlaefer *et al.*, 2006; Ruiz-Casado *et al.*, 2007). Notre approche est similaire à celle de (Pantel *et al.*, 2004) et reprend plus directement encore la méthode de (Embarek & Ferret, 2008). L'idée générale de l'approche est de trouver, dans le contexte entre les entités cible et source, des points communs entre deux phrases exprimant la relation que l'on veut capturer. Ici, nous cherchons ces points communs parmi trois niveaux d'information linguistique : forme de surface, lemme et catégorie morpho-syntaxique. Ces informations linguistiques sont mises en évidence grâce à l'outil OpenNLP[2], qui est plus globalement également utilisé pour la reconnaissance des entités nommées. La présence de ces trois niveaux d'information donne une plus grande expressivité aux patrons construits et permet ainsi de trouver un compromis intéressant en termes de niveau de généralisation entre la spécificité des éléments lexicalisés et le caractère plus général des catégories morpho-syntaxiques.

L'induction d'un patron à partir de deux occurrences de relation est plus précisément composée des trois étapes suivantes :

– le calcul de la distance d'édition entre les deux phrases exemples, c'est-à-dire le nombre minimal d'opérations d'éditions (insertion, suppression, substitution) à effectuer pour passer d'une phrase à l'autre. Toutes les opérations ont ici le même poids ;
– l'alignement optimal des phrases exemples à partir de la matrice des distances entre sous-séquences issue du calcul de la distance d'édition. L'algorithme classique pour trouver un tel alignement est ici étendu en permettant la mise en correspondance de deux mots lors d'une substitution selon les trois niveaux d'information possibles ;
– construction des patrons en complétant si nécessaire les alignements par des opérateurs *jokers (*s*)*, représentant 0 ou 1 mot quelconque, et *(*g*)*, représentant exactement un mot quelconque.

Le tableau 1 montre un exemple d'induction de patron pour le type de relation *founded_by* à partir des deux extraits de phrases ci-dessus. On peut noter la présence de la catégorie *DET* (déterminant) comme généralisation pour *(a|the)*, ce qui rend le patron pertinent pour d'autres extraits tels que *"Charles Kettering, another founder of DELCO ..."*.

Charles Revson	,	the	founder	of		Revlon Cosmetics
Mayer Lehman	,	a	founder	of	the	Lehman Brothers investment
<source>	**,**	***DET***	***founder***	***of***	***(*s*)***	*<cible>*

TAB. 1 – Exemple d'induction de patron de relation

Cet exemple illustre également le fait que la généralisation peut aller jusqu'à l'utilisation de jokers pouvant se substituer à n'importe quel mot. Comme il est toujours possible de généraliser deux phrases en un patron ne contenant que des jokers, il est nécessaire de fixer une limite supérieure au nombre de jokers pouvant être introduits dans une opération de généralisation pour conserver un niveau de spécificité raisonnable des patrons. Par ailleurs, travaillant en domaine ouvert et avec des entités nommées assez générales, nous souhaitons plutôt induire un nombre important de patrons spécifiques qu'un ensemble restreint de patrons très généraux, ceci afin de privilégier la précision. C'est également pour cette raison que nous ne cherchons pas à généraliser les patrons en leur réappliquant la procédure d'induction décrite. Dans l'évaluation présentée en section 3, le nombre maximal de jokers dans un patron est donc fixé à 1.

Dans le contexte de supervision distante dans lequel nous nous plaçons, les phrases exemples ne sont pas directement fournies en tant que telles mais résultent de la projection dans un corpus de relations se présentant sous la forme de couples d'entités (par exemple le couple (Ray Charles, Albany) pour le type de relation *city_of_birth*). Plus concrètement dans notre cas, elles sont récupérées en soumettant à un moteur de recherche des requêtes conte-

[2] `http://opennlp.sourceforge.net/index.html`

nant des couples d'entités pour un type de relations donné et en restreignant les résultats du moteur aux phrases contenant effectivement les deux valeurs des entités. On peut souligner que la nature des restrictions appliquées a un impact direct sur la quantité et la précision des patrons induits. Plus on impose de contraintes, moins on obtient de phrases exemples, mais meilleurs seront les patrons induits. Par exemple, les auteurs de (Agirre *et al.*, 2009) ne retiennent que les phrases exemples dans lesquelles les paires d'entités apparaissent dans un voisinage de zéro à dix mots.

Il est important de noter que le processus d'induction de patrons s'effectue en comparant les phrases exemples deux à deux. Il peut donc être coûteux (en temps de calcul) lorsque le nombre de phrases exemples est important : pour 10 000 exemples, on a environ 50 millions de couples distincts de phrases à comparer ($n(n-1)/2$ exactement). Pour traiter ce problème, la solution immédiate consiste à réduire de façon drastique le nombre de phrases exemples en amont du processus d'induction, la conséquence étant une réduction de la couverture des différentes forme d'expression des types de relations. Une autre solution consiste à faire une réduction sélective du nombre de couples de phrases exemples à généraliser en évitant de considérer les couples de phrases dont la distance est visiblement trop grande pour induire des patrons intéressants. Même si la distance utilisée pour cette induction est une distance d'édition, donc tenant compte de l'ordre des mots, il est évident qu'un faible recoupement des phrases en termes de mots conduira à une valeur élevée de la distance d'édition. Le filtrage *a priori* des couples de phrases peut donc se fonder sur une mesure s'appliquant à une représentation de type « sac de mots », telle que la mesure *cosinus*, en fixant une valeur minimale en dessous de laquelle la généralisation des couples de phrases n'est pas réalisée. Or, la mesure *cosinus* peut être évaluée de manière efficace, soit avec une bonne approximation, comme dans le cas du *Local Sensitive Hashing* (Gionis *et al.*, 1999), soit de manière exacte mais en fixant un seuil de similarité minimale, ce qui correspond à notre cas de figure. Nous avons donc retenu pour notre filtrage l'algorithme *All Pairs Similarity Search* (APSS), proposé dans (Bayardo *et al.*, 2007), qui calcule la mesure *cosinus* pour les seules paires d'objets considérés – ici, les phrases exemples – dont la similarité est supérieure ou égale à un seuil fixé *a priori*. Cet algorithme se fonde plus précisément sur une série d'optimisations dans l'indexation des objets tenant compte des informations recueillies sur leurs caractéristiques et d'un tri appliqué à ces objets en fonction de ces mêmes caractéristiques.

Notons que lors de l'induction de patrons à partir d'un grand volume de phrases exemples, on retrouve de nombreux doublons, soit parce que la même phrase exemple se trouve dans plusieurs documents, soit parce que l'on retrouve la même forme d'expression d'un type de relations avec des valeurs différentes (*Obama's height is 1.87m* ; *Sarkozy's height is 1.65m*). Ainsi, nous proposons de filtrer les phrases exemples à deux niveaux : d'abord avec un seuil de similarité fort afin d'identifier et éliminer les phrases identiques ; puis avec un seuil de similarité faible pour s'assurer d'un niveau minimal de similarité entre les phrases en vue du processus d'induction.

2.2 Extraction des relations

L'extraction de nouvelles relations se fait à partir des types de relations existants et d'entités connues : on cherche à compléter une base de connaissances existante en complétant les informations concernant les entités qu'elle contient. La première étape de l'extraction de relations est la recherche de phrases candidates pouvant contenir l'expression d'une relation. Elle prend comme point de départ des requêtes contenant une entité nommée associée à son type et le type de l'information recherchée. La recherche proprement dite est réalisée, comme dans le cas de l'apprentissage de patrons, grâce à un moteur de recherche ayant préalablement indexé le corpus cible pour l'extraction des relations. Nous nous sommes appuyés dans notre cas sur le moteur Lucène[3], avec une indexation adaptée aux caractéristiques de notre recherche : les documents initiaux sont découpés en unités d'indexation de petite taille, trois phrases, grâce à une fenêtre glissante et au sein de ces unités, sont indexés les mots pleins sous leur forme normalisée et les entités nommées, avec leur type. L'interrogation du corpus présente en outre la particularité d'inclure une phase d'expansion de l'entité source. En effet, on retrouve souvent dans les documents des formes plus ou moins développées des entités nommées : par exemple *Bill Clinton* est généralement utilisé au lieu de *William Jefferson Blythe III Clinton*. Il est donc intéressant de savoir que ces deux mentions d'entités sont équivalentes et associées à la même entité, en particulier lors de la recherche de documents. Nous utilisons donc une étape d'expansion des entités visant à associer à une entité donnée les formes alternatives lui faisant référence. Pour l'entité "Barack Obama", on a ainsi : *{B. Hussein Obama, Barack H. Obama Junior, Barack Obama Jr, Barack Hussein Obama Jr., etc.}*. L'intérêt est de pouvoir augmenter les chances de retrouver des phrases candidates

[3] `http://lucene.apache.org/java/docs/index.html`

liées à l'entité puisque l'on considère tous les documents dans lesquels apparaissent ses différentes expressions. Une base d'expansion des entités a été constituée de façon automatique à partir du corpus Wikipédia[4] en collectant pour chaque entité les formulations extraites des pages de redirection de Wikipédia vers cette entité. Au total, la base d'expansion contient des formes étendues pour environ 2,4 millions d'entrées.

Nous appliquons ensuite sur les phrases candidates sélectionnées les patrons induits lors de la phase d'apprentissage. Les entités cibles extraites par ces patrons sont cumulées pour ne retenir finalement que les plus fréquentes : notre hypothèse est que les entités cibles les plus pertinentes apparaissent plus souvent dans les documents que les moins pertinentes. Pour les relations mono-valuées (ex. : date de naissance), une seule valeur est conservée. Pour les relations multi-valuées (ex. : lieux de résidence), un nombre arbitraire de trois valeurs sont conservées à défaut de connaissances fournies *a priori* ou extraites des textes quant à ce point. Enfin, un dernier filtre est appliqué sur les entités cibles pour vérifier la compatibilité des valeurs obtenues avec les contraintes relatives au type d'information recherché qu'elle représentent, définies par des listes de valeurs ou d'expressions régulières : on vérifie par exemple que le pays de naissance d'une personne fasse bien partie de la liste des pays connus.

3 Évaluation

Nous présentons dans cette section les résultats de l'évaluation de notre système en utilisant les données de la tâche *Slot Filling* de la campagne d'évaluation TAC-KBP 2010 (TAC-KBP, 2010). Nos expérimentations ont donc été réalisées pour l'anglais. La tâche *Slot Filling* correspond aux spécifications de notre contexte de travail telles que nous les avons définies à la section 2 : son objectif est d'extraire à partir d'un vaste corpus l'entité cible d'une relation ayant comme source une entité présente dans une base de connaissances abritant un ensemble important d'exemples du type de relation visé. Les types de relations considérés dans ce cadre sont au nombre de 42, répartis en 16 relations pour des entités de type ORGANISATION (ORG) et 26 relations pour les entités de type PERSONNE (PERS). La liste des types de relations traités est présentée dans le tableau 2. Nous précisons que les expériences ont été réalisées sur un cluster de 24 nœuds (4 processeurs/nœud) avec une parallélisation par type de relations.

3.1 Cadre d'évaluation

Les données d'évaluation issues de TAC-KBP sont les suivantes :

- un corpus de textes composé d'environ 1,8 millions de documents (1 780 980 exactement) répartis en 0,04% de transcriptions (conversations téléphoniques, journaux radio, conversations radio), 72,24% d'articles de presse et 27,72% de pages Web ;
- une base de connaissances *(KB)* reposant sur une image de Wikipédia d'octobre 2008. Un identifiant unique et un type d'entité sont attribués à chaque page contenant des *infobox*. Le type d'entité *personne*, *organisation*, *entité géopolitique* ou *inconnu* est associé à chaque page en fonction des champs contenus dans les infobox. Typiquement, les infobox *Infobox_Actor* sont ainsi liées à des personnes. Au final 818 741 entités ont été retenues pour former la KB, chacune d'elles étant associée à un ensemble de propriétés (champs des infobox) ainsi qu'à un texte la décrivant. Ainsi les relations sont représentées dans la KB par des tuples (identifiant, type infobox, nom, type, propriété, valeurs), ex. : (E0000437 ; Infobox_Actor ; Julia Roberts ; PER ; birthplace ; Atlanta) ;
- une table de correspondance entre les propriétés issues de Wikipédia et les types de relations retenus pour l'évaluation. Par exemple, *Infobox_Actor*:*birthplace* est convertie en *per*:*city_of_birth*. Cette correspondance permet de prendre en compte une certaine hétérogénéité de désignation des propriétés dans Wikipédia ;
- une liste de 100 entités sources pour lesquelles on cherche à extraire toutes les entités en relation pour tous les types de relations considérés. On dénombre parmi ces entités 15 entités présentes dans la KB et 85 inconnues de la KB. Par ailleurs, toutes les relations considérées ne trouvent pas d'entités cibles dans le corpus pour ces 100 entités. Dans le cadre de cette étude, nous nous focalisons uniquement sur les relations pour lesquelles il existe une entité cible dans le corpus[5], ce qui représente au total 2069 relations. Le détail par type de relations est présenté dans la colonne *Nb Ref.* du tableau 2.

[4] Plus précisément, la version mise à disposition par l'université de New York : `http://nlp.cs.nyu.edu/wikipedia-data`

[5] Les entités cibles existantes dans le corpus sont établies par la référence fournie par les organisateurs de la campagne, construite à partir des résultats des participants.

Types de relations	Type de cible	Couv. Doc.	Couv. Rel.	Nb Appr.	Nb Test	Nb Induc.	Nb Patrons	Couv. Patrons	Nb Ref.
org:alternate_names	ORG	89,17%	33,33%	20 013	10 006	214	6 007	66,10%	120
org:city_of_headquarters	LOC + liste	90,12%	59,26%	6 847	3 423	4 553	2 010 749	65,52%	81
org:country_of_headquarters	LOC + liste	91,04%	55,22%	18 401	9 200	2 110	185 158	69,56%	67
org:dissolved	DATE	100%	25%	532	266	87	775	0%	4
org:founded_by	ORG/PER	95,45%	31,82%	1 954	977	197	4 385	77,87%	28
org:founded	DATE	92,86%	53,57%	13 688	6 844	127	22 482	77,34%	22
org:member_of	ORG	100%	100%	7 951	3 976	102	103	70%	2
org:members	ORG	77,78%	11,11%	531	265	183	552	86%	9
org:number_of_employees_members	regexp + liste	90,48%	23,81%	7 173	3 586	216	3 109	100%	21
org:parents	ORG	96,67%	43,33%	22 361	11 181	3 013	485 947	69,04%	30
org:political_religious_affiliation	ORG	78,57%	64,29%	3 427	1 713	406	3 250	55,36%	14
org:shareholders	ORG/PER	66,67%	33,33%	3	2	0	0	0%	3
org:stateorprovince_of_headquarters	LOC + liste	92,65%	63,24%	9 672	4 836	1 422	148 610	69,93%	68
org:subsidiaries	ORG	82,69%	28,85%	5 588	2 794	498	3 764	56,48%	52
org:top_members_employees	PER	91,48%	37,22%	40 929	20 464	108	1 010	70,57%	223
org:website	regexp	78,26%	30,43%	30 813	15 407	32	28	0%	23
per:age	regexp + liste	85,32%	32,11%	157	79	3	1	0%	109
per:alternate_names	PER	61,63%	11,63%	18 115	9 057	68	2 818	82,58%	86
per:cause_of_death	liste	100%	0%	1	1	0	0	0%	2
per:charges	liste	61,54%	0%	184	92	0	0	0%	13
per:children	PER	72%	16%	2 010	1 005	147	238	0%	25
per:cities_of_residence	LOC + liste	77,59%	34,48%	3 631	1 815	722	14 297	77,88%	58
per:city_of_birth	LOC + liste	69,23%	15,38%	4 745	2 373	2 252	62 455	63,34%	13
per:city_of_death	LOC + liste	100%	100%	1 631	816	505	2 860	70,27%	1
per:countries_of_residence	LOC + liste	73,53%	20,59%	8 098	4 049	2 181	205 344	80,08%	34
per:country_of_birth	LOC + liste	82,35%	5,88%	11 085	5 542	11 192	9 145 385	65,02%	17
per:country_of_death	LOC + liste			2 873	1 436	1 068	22 374	62,89%	0
per:date_of_birth	DATE	90%	20%	11 689	5 845	30	22	0%	20
per:date_of_death	DATE	100%	0%	4 692	2 346	54	63	33,33%	1
per:employee_of	ORG	84,21%	29,32%	24 762	12 381	2 435	704 833	71,13%	133
per:member_of	ORG	82,42%	36,26%	27 523	13 761	3 901	740 999	57,25%	91
per:origin	liste	81,58%	42,11%	37 626	18 813	2 710	276 653	74,41%	76
per:other_family	PER	86,67%	33,33%	4	2	0	0	0%	30
per:parents	PER	78,13%	9,38%	1 314	657	37	604	77,78%	64
per:religion	liste	85,71%	57,14%	1 468	734	515	1 575	80%	7
per:schools_attended	ORG + liste	87,50%	37,50%	2 246	1 123	67	170	4,17%	16
per:siblings	PER	78,26%	20,29%	4	2	0	0	0%	69
per:spouse	PER	80%	35,56%	5 385	2 693	3 094	314 329	80%	45
per:stateorprovince_of_birth	LOC + liste	80%	50%	7 047	3 523	2 097	60 782	75,42%	10
per:stateorprovince_of_death	LOC + liste	100%	100%	1 616	808	278	911	66,67%	1
per:states_or_provinces_of_residence	LOC + liste	84,21%	50%	4 980	2 490	1 166	115 418	77,90%	38
per:title	liste	84,55%	52,77%	31 574	15 787	8 797	1 573 512	49,07%	343

TAB. 2 – Résultats des différentes étapes, pour tous les types de relations

Type de cible : mécanisme utilisé pour retrouver l'entité cible. *Couv. Doc.* : couverture des documents de référence dans les résultats de la recherche de phrases. *Couv. Rel.* : couverture des phrases candidates de référence. *Nb Appr.* : nombre de relations pour l'apprentissage des patrons. *Nb Test* : nombre de relations pour l'évaluation des patrons. *Nb Induc.* : nombre de phrases contenant des occurrences de relations pour l'induction des patrons. *Nb Patrons* : nombre de patrons induits à partir des occurrences de relations. *Couv. Patrons* : couverture des patrons induits. *Nb Ref.* : nombre de relations de référence.

3.2 Évaluation de l'apprentissage des patrons

Les patrons servent à confirmer/infirmer la présence d'une relation entre deux entités. Il est donc important de vérifier que les patrons appris aient une couverture suffisamment large pour retrouver le plus possible de variantes parmi les occurrences de relations. Pour évaluer la qualité des patrons, nous avons séparé les relations connues en deux ensembles : un ensemble d'apprentissage (2/3 des relations) et un ensemble de test (1/3 des relations). Nous mesurons la qualité de la couverture des patrons en calculant le pourcentage des occurrences de relations de l'ensemble de test que l'on retrouve en appliquant les patrons appris à partir des occurrences de relations de l'ensemble d'apprentissage. Le corpus utilisé pour réaliser cette évaluation est le corpus TAC-KBP 2010 décrit ci-dessus. Précisons que l'utilisation de ce corpus pour évaluer l'extraction des relations n'empêche pas son utilisation pour l'apprentissage des patrons, les relations étant différentes pour les deux tâches.

Nous indiquons dans le tableau 2 le nombre de relations de l'ensemble d'apprentissage et de l'ensemble de test respectivement dans les colonnes *Nb. Appr* et *Nb. Test*. Le nombre de phrases trouvées contenant des occurrences des relations du corpus d'entraînement, et qui ont donc servi pour l'induction des patrons, est indiqué dans la colonne *Nb. Induc*. Le nombre de patrons générés à partir de ces phrases candidates est indiqué dans la colonne *Nb. Patrons* de ce même tableau.

Par exemple, pour le type de relation *org:alternate_names*, à partir des 20 013 relations de l'ensemble d'apprentissage, seules 214 phrases candidates contenant l'expression d'une de ces relations sont sélectionnées. Ces 214 phrases servent à générer 6 007 patrons, qui ont une couverture de 66,10% (*i.e.* on retrouve 66,10% des phrases contenant des occurrences des 10 006 relations de test). L'écart conséquent entre les 20 013 relations et les 214 phrases trouvées est dû à deux facteurs :

– une contrainte réductrice imposée lors de la sélection des phrases candidates. Seules les phrases dont tous les mots des entités nommées sont correctement identifiés sont en effet conservées. Or, les entités peuvent être partiellement (ou mal) reconnues lors des traitements linguistiques ;
– la nature des documents du corpus : 72% des documents sont des articles de presse édités entre janvier 2008 et août 2009, ce qui explique le peu de documents, voir aucun, concernant certaines personnes ou organisations pourtant présentes dans la KB.

Les résultats de la couverture des patrons sont présentés dans le tableau 2 pour chaque type de relations dans la colonne *Couv. Patrons*. À titre indicatif, le temps d'induction des patrons pour le type de relations *per:country_of_birth* (11 192 phrases exemples à comparer) passe de 690mn et 5s pour la version sans filtrage à 0mn et 30s pour la version avec filtrage[6], ce qui illustre l'intérêt de celui-ci en termes de temps de calcul.

3.3 Évaluation de l'extraction des relations

L'extraction des relations comprenant plusieurs étapes, chacune d'entre elles peut influer sur le résultat global. Nous proposons donc de faire une évaluation séparée de la recherche des phrases candidates et de l'extraction des relations proprement dite.

3.3.1 Recherche des phrases candidates

Une condition nécessaire pour des extraire relations pertinentes est de s'assurer que le moteur de recherche renvoie suffisamment de documents pertinents pour nous permettre de retrouver des entités cibles. Nous avons donc mesuré la couverture en documents de notre recherche de phrases candidates, à savoir le pourcentage de documents renvoyés par l'index que l'on retrouve effectivement dans la référence. Nous avons testé de ce point de vue différentes stratégies en faisant varier des paramètres comme le nombre de résultats retournés et l'utilisation ou non de l'expansion pour la requête. Les résultats de cette évaluation nous ont ainsi conduit à utiliser les entités sources et leurs formes étendues pour interrogation de l'index et prendre en compte les 1000 premiers résultats retournés : ces paramètres permettent de retrouver 84,24% des documents de référence. Le résultat détaillé par type de relations est donné par la colonne *Couv. Doc* du tableau 2.

À partir des documents ainsi sélectionnés, les phrases candidates à l'extraction d'une relation pour un type donné

[6]La version avec filtrage étant parallélisée, le temps donné est une somme des temps comptabilisés au niveau de chaque processeur.

sont extraites en retenant les phrases contenant à la fois l'entité source et le type de l'entité cible. La qualité et la quantité des phrases candidates sont largement influencées par la qualité de la reconnaissance des entités nommées. Comme nous ne disposons pas d'annotation de référence pour les entités nommées du corpus, il n'est pas possible de mesurer les pertes causées par la mauvaise reconnaissance des entités. En revanche, nous avons évalué la proportion de documents de référence dans lesquels nous retrouvons des phrases candidates. Cette donnée permet de fixer une borne maximale pour le pourcentage de relations qu'il serait possible d'extraire si les étapes à la suite se déroulaient idéalement. Nous obtenons au total une couverture de 37,55% des phrases appartenant aux documents de référence. Le détail par type de relations est présenté à la colonne *Couv. Rel* du tableau 2.

3.3.2 Extraction de relations

Pour évaluer les relations extraites, nous avons réutilisé les mesures et les outils d'évaluation fournis par la campagne TAC-KBP[7] sans nous limiter aux seuls documents présents dans la référence pour accepter une relation correcte[8]. Le tableau 3 fournit les résultats de cette évaluation en agglomérant tous les types de relations et en caractérisant l'impact du filtrage *a posteriori* des entités cibles sur les relations extraites en termes de rappel *(R.)*, précision *(P.)* et f1-mesure *(F1.)*. Pour mémoire, ce filtrage consiste à s'assurer que l'entité cible valide des expressions régulières et/ou une liste fermée de valeurs. Nous indiquons dans la colonne *Type de cible* du tableau 2 le mécanisme utilisé pour chaque type de relations.

Les résultats du tableau 3 montrent d'une part, que ce filtrage améliore les performances (en moyenne +2,74% de f1-mesure) et d'autre part, valident l'hypothèse que les patrons induits à partir de l'APSS sont aussi pertinents que ceux induits en considérant tous les exemples de relations deux à deux (dans ce cas, il y a même une amélioration de +1,72% de la f1-mesure en moyenne).

	Avant filtrage			Après filtrage		
	R. (%)	P. (%)	F1. (%)	R. (%)	P. (%)	F1. (%)
Tous les couples d'entités	16,28	11,20	13,26	18,07	13,66	15,56
APSS	16,90	12,76	14,54	18,67	16,87	17,72

TAB. 3 – Évaluation de l'impact du filtrage des réponses

Le tableau 4 présente les résultats de différents systèmes sur deux corpus très similaires, les corpus KBP 2009 et KBP 2010, ce dernier ajoutant au premier des documents Web et des transcriptions, *a priori* plus difficiles. Bien que ces chiffres ne portent que sur les relations effectivement présentes dans le corpus, ils intègrent la contrainte pour les systèmes ayant participé à la tâche *Slot Filling* de devoir décider si la relation existe ou non dans le corpus, ce que notre système, développé en dehors du contexte de ces campagnes, ne fait pas. Dans ce tableau, les colonnes *2009* et *2010* désignent les scores des trois systèmes les plus et les moins performants de KBP 2009 et 2010. Ji *et al.* (2010) ont montré que sur 492 relations de référence, 60,4% se trouvaient dans la même phrase tandis que les 39,6% restantes dépassaient l'espace phrastique dans leur expression et nécessitaient pour leur extraction la résolution de coréférences ou l'application de mécanismes d'inférence impliquant par exemple la composition de plusieurs relations ou l'utilisation de connaissances *a priori* sur les types de relations. De ce fait, nous avons distingué dans la colonne *2010 (a)* du tableau 4 les scores des systèmes qui nous sont les plus directement comparables, c'est-à-dire ceux se limitant à l'extraction de relations au niveau phrastique.

On peut noter que le meilleur système de KBP 2010 (Chada *et al.*, 2010) se détache très nettement : +36,63% par rapport au deuxième et +4,68% par rapport à un annotateur humain. Cette prédominance s'appuie à la fois sur l'utilisation d'un corpus annoté manuellement (différent du corpus KBP) de 3 millions de documents et la présence de plusieurs mécanismes d'extraction de relations au niveau inter-phrastique : coréférence pronominale, métonymie entre entités, résolution de dépendances sémantiques entre les mots et les entités, etc. L'utilisation du corpus supplémentaire semble être l'élément déterminant par rapport aux systèmes venant à la suite immédiate, ceux-ci se distinguant de systèmes plus médians par la prise en compte des relations inter-phrastiques. Les plus mauvais résultats, plus faibles en 2010, sont dûs pour une bonne part à des systèmes en cours de développement.

[7] `http://nlp.cs.qc.cuny.edu/kbp/2010/scoring.html`

[8] La référence du point de vue des documents n'étant constituée qu'à partir des résultats des participants à l'évaluation TAC-KBP, elle n'est en effet pas complète.

Concernant notre système, le tableau 4 permet de situer nos résultats dans la moyenne des résultats obtenus par les participants de l'évaluation KBP 2010 et parmi les trois premiers systèmes pour les approches faisant de l'extraction de relations au niveau de la phrase. Dans ce dernier cas, l'approche la plus performante (29,15% de f1-mesure) (Byrne & Dunnion, 2010) utilise des règles construites manuellement permettant d'atteindre un score de précision (66,55%) équivalent au meilleur score de la campagne (66,80%) et un score de rappel (18,67%) se situant dans la moyenne de la campagne (15,33%). Ce fort déséquilibre entre précision et rappel est d'ailleurs assez symptomatique des approches manuelles.

Systèmes TAC KBP	2009	2010	2010 (a)
Nb. soumissions (N) / participants	N=16 / 8	N=31 / 15	N=18
Annotateur humain	58,99%	61,10%	61,10%
1^{er} score	34,35%	65,78%	29,15%
$2^{ème}$ score	25,05%	29,15%	14,22%
$3^{ème}$ score	18%	28,29%	14,13%
(N-2)^{ème} score	5,90%	0,55%	0,55%
(N-1)^{ème} score	2,60%	0,19%	0,19%
N^{ème} score	1,75%	0,08%	0,08%
Notre système	–	17,72%	17,72%
Moyenne	13,43%	17,49%	9,71%
Médiane	13,93%	14,13%	12,27%

TAB. 4 – Résultats sur les données TAC-KBP (f1-mesure)

4 Travaux associés

L'extraction de relations à large échelle, au sens où nous l'avons définie à la section 2, est une problématique encore récente. Néanmoins, au travers notamment des évaluations TAC-KBP, elle a été l'objet d'un certain nombre de travaux proposant différentes approches. Concernant spécifiquement l'extraction des relations, les travaux se répartissent entre l'utilisation de l'apprentissage statistique (Agirre *et al.*, 2009; Li *et al.*, 2009b; Chen *et al.*, 2010b), l'induction de patrons lexicaux (Li *et al.*, 2009a; de Pablo-Sánchez *et al.*, 2009; McNamee *et al.*, 2009; Chen *et al.*, 2010b) et enfin, l'adaptation de systèmes existants pour la détection de relations (Schone *et al.*, 2009; Bikel *et al.*, 2009). On note pour KBP 2010 l'introduction d'approches à base de règles, par exemple (Byrne & Dunnion, 2010), et d'approches reposant sur le principe de « Distant supervision » à partir de classifieurs, dont celle de (Surdeanu *et al.*, 2010). Notre approche relève de l'induction de patrons lexicaux et fait l'hypothèse, comme (Mintz *et al.*, 2009), que la seule présence d'un couple d'entités dans une phrase est suffisante pour marquer la présence effective d'une relation entre ces entités. Ce n'est cependant pas toujours le cas et nous pensons ainsi qu'il est important de filtrer en amont les exemples utilisés pour l'induction des patrons, à l'instar de ce que propose (Riedel *et al.*, 2010).

Comme notre système, ceux élaborés pour KBP 2009 n'exploitent pas les liens de dépendance entre les types de relations, à l'image du lien entre la date de naissance et l'âge par exemple. Dans (Chen *et al.*, 2010a), les auteurs montrent que les résultats obtenus dans (Li *et al.*, 2009a) (31,96% de f1-mesure) peuvent être améliorés (ils obtiennent 34,81% de f1-mesure) par l'intégration des dépendances entre les relations en utilisant des règles d'inférence fondées sur une extension de la logique du premier ordre. Plus généralement, Chada *et al.* (2010) ont montré dans le cadre de KBP 2010 une augmentation très significative des performances en intégrant des mécanismes permettant d'extraire des relations au-delà de la phrase. Sur un autre plan, Li *et al.* (2009a) se distinguent dans KBP 2009 en utilisant deux étapes d'extraction de relations : la première vise à retrouver dans les documents du corpus des entités cibles potentielles en utilisant des patrons de relations ; la seconde applique le même processus à une version récente de Wikipédia pour trouver des entités cibles potentielles supplémentaires qui n'auraient pas été identifiées lors de la première étape. Les entités cibles ainsi acquises ne sont finalement conservées que si elles sont retrouvées dans un document du corpus. Cette récupération d'entités améliore les performances de façon significative (+9% de f1-mesure par rapport à (Bikel *et al.*, 2009)) mais ajoute l'utilisation d'un corpus externe que l'on peut considérer comme trop lié à la KB. Les résultats sur KBP 2010 ont d'ailleurs montré que

les performances globales pouvaient être améliorées sans cette ressource supplémentaire et que son impact sur les résultats est plus limité que pour KBP 2009 (une baisse des résultats a même été observée).

5 Conclusion et perspectives

Dans cet article, nous avons présenté un système d'extraction d'information à large échelle permettant d'extraire des relations de type attributive entre entités nommées. Le qualificatif « à large échelle » recouvre à la fois la prise en compte d'un grand nombre de types de relations et la recherche de ces relations dans un large corpus. Ce système se fonde sur une approche faiblement supervisée dans laquelle les exemples se limitent à des couples d'entités en relation. L'extraction des relations s'effectue par l'application de patrons lexico-syntaxiques caractéristiques des types de relations considérés et appris à partir de phrases issues de la projection des couples d'entités exemples dans un corpus. Nous avons évalué les résultats de cette approche en utilisant le cadre d'évaluation offert par la tâche *Slot Filling* de l'évaluation KBP en nous concentrant sur la problématique de l'extraction des relations proprement dite, sans nous attacher à la détection de l'absence d'une relation dans un corpus. Les résultats obtenus dans ce contexte se situent dans la moyenne des résultats obtenus par les participants de l'édition 2010, ce que nous pouvons considérer comme un point de départ intéressant dans la mesure où notre système repose sur une approche volontairement générique et n'exploite que très faiblement les spécificités des types de relations traités. Nous avons aussi pu montrer que des techniques permettant de prendre en compte certains aspects d'un passage à une « large échelle », comme le filtrage des couples de phrases exemples à généraliser par l'utilisation de l'APSS, ne dégradent pas les performances et peuvent même contribuer à les améliorer.

Nous travaillons par ailleurs sur l'amélioration de notre système en conservant l'idée de garder une certaine généricité par rapport au type des relations considérées. Pour ce faire, nous nous focalisons particulièrement sur l'apprentissage des patrons d'extraction. Un premier pas dans cette direction vise à disposer à la fois d'un nombre plus important d'exemples mais également d'exemples de meilleure qualité. Ces deux points sont liés dans la mesure où l'obtention d'un ensemble plus large d'exemples passe par le relâchement des contraintes touchant la sélection des phrases exemples. Or, si l'on peut espérer qu'un tel relâchement permettra l'obtention de nouveaux bons exemples, il sera aussi source de nouveaux mauvais exemples. Nous souhaitons donc coupler un tel relâchement avec l'utilisation d'un module de filtrage de relations qui, à l'instar de (Banko & Etzioni, 2008), est capable de déterminer si une phrase contient une relation entre deux entités sans *a priori* sur la nature de cette relation.

Références

AGICHTEIN E. & GRAVANO L. (2000). Snowball : Extracting relations from large plain-text collections. In *5^{th} ACM International Conference on Digital Libraries*, p. 85–94, San Antonio, Texas, USA.

AGIRRE E., CHANG A., JURAFSKY D., MANNING C., SPITKOVSKY V. & YEH E. (2009). Stanford-UBC at TAC-KBP. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA.

BANKO M. & ETZIONI O. (2008). The tradeoffs between open and traditional relation extraction. In *ACL-08 : HLT*, p. 28–36, Columbus, Ohio.

BAYARDO R., MA Y. & SRIKANT R. (2007). Scaling up all pairs similarity search. In *16^{th} International Conference on World Wide Web (WWW'07)*, p. 131–140, Banff, Alberta, Canada.

BIKEL D., CASTELLI V., RADU F. & JUNG HAN D. (2009). Entity Linking and Slot Filling through Statistical Processing and Inference Rules. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA.

BIZER C., LEHMANN J., KOBILAROV G., AUER S., BECKER C., CYGANIAK R. & HELLMANN S. (2009). DBpedia - A crystallization point for the Web of Data. *Journal of Web Semantics*, **7**, 154–165.

BYRNE L. & DUNNION J. (2010). UCD IIRG at TAC 2010 KBP Slot Filling Task. In *Third Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA.

CHADA D., ARANHA C. & MONTE C. (2010). An Analysis of The Cortex Method at TAC 2010 KBP Slot-Filling. In *Third Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA.

CHEN Z., TAMANG S., LEE A., LI X., PASSANTINO M. & JI H. (2010a). Top-down and Bottom-up : A Combined Approach to Slot Filling. In *6th Asia Information Retrieval Symposium on Information Retrieval Technology*, Gaithersburg, Maryland, USA : Springer-Verlag.

CHEN Z., TAMANG S., LEE A., LI X., SNOVER M., PASSANTINO M., LIN W.-P. & JI H. (2010b). CUNY-BLENDER TAC-KBP2010 Slot Filling System Description. In *Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA.

CLAVEAU V. & SÉBILLOT P. (2004). From efficiency to portability : acquisition of semantic relations by semi-supervised machine learning. In *20^{th} International Conference on Computational Linguistics (COLING 2004)*, p. 261–267, Geneva, Switzerland.

DE PABLO-SÁNCHEZ C., PEREA J., SEGURA-BEDMAR I. & MARTÍNEZ P. (2009). The UC3M team at the Knowledge Base Population task. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA.

EMBAREK M. & FERRET O. (2008). Learning patterns for building resources about semantic relations in the medical domain. In *6^{th} Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

GIONIS A., INDYK P. & MOTWANI R. (1999). Similarity search in high dimensions via hashing. In *25^{th} International Conference on Very Large Data Bases (VLDB'99)*, p. 518–529, Edinburgh, Scotland, UK.

HEARST M. (1992). Automatic acquisition of hyponyms from large text corpora. In *14^{th} International Conference on Computational linguistics (COLING'92)*, p. 539–545, Nantes, France.

JI H., GRISHMAN R. & TRANG DANG H. (2010). Overview of the TAC 2010 Knowledge Base Population Track. In *Third Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA.

LI F., ZHENG Z., BU F., TANG Y., ZHU X. & HUANG M. (2009a). THU QUANTA at TAC 2009 KBP and RTE Track. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA.

LI S., GAO S., ZHANG Z., LI X., GUAN J., XU W. & GUO J. (2009b). PRIS at TAC 2009 : Experiments in KBP Track. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA.

MCNAMEE P., DREDZE M., GERBER A., GARERA N., FININ T., MAYFIELD J., PIATKO C., RAO D., YAROWSKY D. & DREYER M. (2009). HLTCOE Approaches to Knowledge Base Population at TAC 2009. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA.

MINTZ M., BILLS S., SNOW R. & JURAFSKY D. (2009). Distant supervision for relation extraction without labeled data. In *ACL-IJCNLP'09*, p. 1003–1011, Suntec, Singapore.

PANTEL P., RAVICHANDRAN D. & HOVY E. (2004). Towards terascale knowledge acquisition. In *20th International Conference on Computational Linguistics (COLING'04)*, p. 771–777, Geneva, Switzerland.

RAVICHANDRAN D. (2005). *Terascale knowledge acquisition*. PhD thesis, Faculty of the Graduate School University of Southern California, Los Angeles, CA, USA.

RIEDEL S., YAO L. & MCCALLUM A. (2010). Modeling relations and their mentions without labeled text. In J. BALCÁZAR, F. BONCHI, A. GIONIS & M. SEBAG, Eds., *Machine Learning and Knowledge Discovery in Databases*, volume 6323 of *Lecture Notes in Computer Science*, p. 148–163. Springer Berlin / Heidelberg.

RUIZ-CASADO M., ALFONSECA E. & CASTELLS P. (2007). Automatising the learning of lexical patterns : An application to the enrichment of wordnet by extracting semantic relationships from wikipedia. *Data Knowledge Engineering*, **61**, 484–499.

SCHLAEFER N., GIESELMANN P., SCHAAF T. & WAIBEL A. (2006). A pattern learning approach to question answering within the ephyra framework. In P. SOJKA, I. KOPECEK & K. PALA, Eds., *Text, Speech and Dialogue*, volume 4188 of *Lecture Notes in Computer Science*, p. 687–694. Springer Berlin / Heidelberg.

SCHONE P., GOLDSCHEN A., LANGLEY C., LEWIS S., ONYSHKEVYCH B., CUTTS R., DAWSON B., MACBRIDE J., MATRANGOLA G., MCDONOUGH C., PFEIFER C. & URSIAK M. (2009). TCAR at TAC-KBP 2009. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA.

SHINYAMA Y. & SEKINE S. (2006). Preemptive information extraction using unrestricted relation discovery. In *HLT-NAACL 2006*, p. 304–311, New York City, USA.

SURDEANU M., MCCLOSKY D., TIBSHIRANI J., BAUER J., CHANG A., SPITKOVSKY V. & MANNING C. (2010). A Simple Distant Supervision Approach for the TAC-KBP Slot Filling Task. In *Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA.

TAC-KBP (2010). Preliminary task description for knowledge-base population at TAC 2010.

ZHOU G., SU J., ZHANG J. & ZHANG M. (2005). Exploring various knowledge in relation extraction. In *43^{rd} Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, p. 427–434, Ann Arbor, USA.

ZHOU G., ZHANG M., JI D. & ZHU Q. (2007). Tree kernel-based relation extraction with context-sensitive structured parse tree information. In *EMNLP - CoNLL'07*, p. 728–736, Prague, Czech Republic.