Filtrage de relations pour l'extraction d'information non supervisée

Wei Wang[1] Romaric Besançon[1] Olivier Ferret[1] Brigitte Grau[2]
(1) CEA, LIST, 18 route du Panorama, BP 6, 92265 Fontenay-aux-Roses
(2) LIMSI, UPR-3251 CNRS-DR4, Bat. 508, BP 133, 91403 Orsay Cedex
wei.wang@cea.fr, romaric.besancon@cea.fr, olivier.ferret@cea.fr, brigitte.grau@limsi.fr

Résumé. Le domaine de l'extraction d'information s'est récemment développé en limitant les contraintes sur la définition des informations à extraire, ouvrant la voie à des applications de veille plus ouvertes. Dans ce contexte de l'extraction d'information non supervisée, nous nous intéressons à l'identification et la caractérisation de nouvelles relations entre des types d'entités fixés. Un des défis de cette tâche est de faire face à la masse importante de candidats pour ces relations lorsque l'on considère des corpus de grande taille. Nous présentons dans cet article une approche pour le filtrage des relations combinant méthode heuristique et méthode par apprentissage. Nous évaluons ce filtrage de manière intrinsèque et par son impact sur un regroupement sémantique des relations.

Abstract. Information Extraction have recently been extended to new areas, by loosening the constraints on the strict definition of the information extracted, thus allowing to design more open information extraction systems. In this new domain of unsupervised information extraction, we focus on the task of extracting and characterizing new relations between a given set of entity types. One of the challenges of this task is to deal with the large amount of candidate relations when extracting them from a large corpus. We propose in this paper an approach for filtering such candidate relations, based on heuristic and machine learning methods. We present an evaluation of this filtering phase and an evaluation of the impact of the filtering on the semantic clustering of relations.

Mots-clés : Extraction d'information non supervisée, filtrage, apprentissage automatique, clustering.

Keywords: Unsupervised information extraction, filtering, machine learning, clustering.

1 Introduction[1]

Les années récentes ont vu se développer de nouveaux paradigmes dans le domaine de l'extraction d'information (EI), parmi lesquels la notion d'EI non supervisée. Cette approche prend comme point de départ des entités ou des types d'entités et se fixe comme objectif de mettre en évidence les relations intervenant entre ces entités, sans connaissance *a priori* de leur type. Cette mise en évidence est éventuellement suivie d'un regroupement de ces relations en fonction de leurs similarités pour en faire la synthèse. Les travaux effectués dans ce champ de recherche s'envisagent selon trois points de vue. Le premier est l'acquisition de connaissances, que ce soit des connaissances sur le monde collectées à vaste échelle à partir du Web, comme avec le concept d'*Open Information Extraction* développé dans (Banko *et al.*, 2007), ou dans des domaines plus spécialisés, comme le domaine biologique, où cette extraction est le moyen d'ajouter de nouveaux types de relations entre entités à une ontologie existante (Ciaramita *et al.*, 2005). Le deuxième se situe dans le cadre d'applications d'EI, où ce type d'approche correspond à la volonté d'offrir aux utilisateurs des modes d'extraction de l'information plus souples et plus ouverts quant à la spécification de leur besoin informationnel. L'approche *On-demand information extraction* (Sekine, 2006), préfigurée dans (Hasegawa *et al.*, 2004) et concrétisée par les travaux sur la *Preemptive Information Extraction* (Shinyama & Sekine, 2006), vise ainsi à induire l'équivalent d'un *template* à partir d'un ensemble de documents représentatifs des informations à extraire, obtenus par le biais d'un moteur de recherche, par le regroupement des relations qui en sont extraites (Rosenfeld & Feldman, 2007). Enfin, l'EI non supervisée peut aussi servir à compléter l'EI supervisée, qui dépend de corpus annotés qui ne sont généralement pas de grande taille, étant donné la complexité des tâches considérées. Les résultats d'une approche non supervisée peuvent alors être utilisés pour élargir la couvertures des modèles appris (Banko & Etzioni, 2008; González & Turmo, 2009).

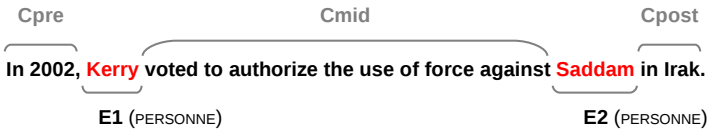
Dans cet article, nous nous plaçons dans le cadre du deuxième point de vue exposé ci-dessus, celui d'une extrac-

[1]Ce travail a été partiellement réalisé dans le cadre du projet FILTRAR-S soutenu par le programme CSOSG 2008 de l'ANR.

tion d'information plus souple, en y intégrant la problématique du filtrage des relations extraites telle qu'elle est abordée dans (Banko *et al.*, 2007), mais avec un point de vue différent. Nous présentons ainsi un travail visant à déterminer si deux entités nommées cooccurrant dans une phrase entretiennent ou non une relation sans fixer d'*a priori* sur la nature de cette relation. Compte tenu de notre perspective globale, nous évaluons dans un second temps l'impact d'un tel filtrage sur le regroupement des relations extraites.

2 Vue d'ensemble

Le travail que nous présentons ici s'inscrit dans un contexte plus large visant à développer un processus d'extraction d'information non supervisée susceptible de répondre à des problématiques de veille telle que « suivre tous les événements faisant intervenir les sociétés X et Y ». À la base de ce processus, nous nous restreignons comme les travaux ci-dessus, hormis (Banko *et al.*, 2007), aux relations présentes entre deux entités nommées. Plus formellement, les relations extraites des textes sont caractérisées par deux grandes catégories d'information permettant tout à la fois de les définir et de fournir les éléments nécessaires à leur regroupement : un couple d'entités nommées (E1 et E2) et une caractérisation linguistique de la relation (*Cpre*, *Cmid*, *Cpost*).



Cpre, *Cmid* et *Cpost* correspondent respectivement à la partie de la phrase précédant la première entité E1, située entre les deux entités et suivant la seconde entité E2. L'ensemble recouvre l'expression linguistique de la relation. Le plus souvent *Cmid* exprime la relation proprement dite tandis que *Cpre* et *Cpost* fournissent plutôt des éléments de contexte pouvant être utiles dans la perspective de son regroupement avec d'autres relations.

Le processus d'extraction d'information non supervisée défini autour de cette notion de relation s'effectue avec les étapes suivantes : pré-traitement linguistique des textes, extraction de relations candidates, filtrage des relations candidates et regroupement des relations selon leur similarité. Le pré-traitement linguistique des textes permet de mettre en évidence les informations nécessaires à la définition des relations. Ce pré-traitement comporte donc une reconnaissance des entités nommées pour les types d'entités visés, une désambiguïsation morpho-syntaxique des mots ainsi que leur normalisation. Ces traitements s'appuient sur les outils d'OpenNLP. Nous nous concentrons sur six types de couples d'entités : ORG - LIEU, ORG - ORG, ORG - PERS, PERS - LIEU, PERS -ORG, PERS - PERS.

3 Filtrage des relations

Les relations candidates, extraites de 18 mois du *New York Times* issus du corpus AQUAINT-2, sont constituées par tout couple d'entités nommées (EN) dont les types correspondent aux types ciblés, avec pour seules restrictions la cooccurrence de ces entités dans une même phrase et la présence d'au moins un verbe entre les deux. Leur examen montre qu'un nombre très significatif des relations ainsi extraites ne sont pas valides. Cette stratégie basique d'extraction, intéressante en domaine de spécialité, n'est pas suffisamment sélective en domaine ouvert. Nous avons donc cherché à la compléter par un processus de filtrage visant, comme dans (Banko & Etzioni, 2008), à déterminer si deux entités dans une phrase sont liées par une relation, sans *a priori* sur la nature de cette relation.

3.1 Filtrage heuristique

Compte tenu du nombre important de relations invalides, nous avons d'abord défini un nombre restreint d'heuristiques pour réaliser un filtrage à gros grain. Ces heuristiques sont au nombre de trois :

– suppression des relations comportant entre leurs deux entités un verbe exprimant un discours rapporté ;
– limitation à 10 du nombre de mots entre les deux entités. Au-delà de cette limite empirique, le nombre des relations effectives entre les deux entités devient en effet très faible ;
– limitation à 1 du nombre des verbes entre les deux entités, sauf si ce sont des auxiliaires.

L'application de ces heuristiques aux relations extraites a globalement pour conséquence de réduire leur volume d'environ 50%. Nous avons choisi au hasard 50 relations de chaque type et nous avons procédé à une annotation manuelle de leur validité. Le taux de fausses relations constaté parmi les relations filtrées étant encore assez élevé, 50,7% pour les 6 types considérés, nous avons mis en œuvre un second filtrage, à grain plus fin.

3.2 Filtrage par apprentissage

Cette seconde étape de filtrage repose sur un classifieur statistique décidant si une relation extraite est véritablement sous-tendue par une relation effective entre ses entités. Nous avons construit pour ce faire un corpus d'entraînement et de test en annotant manuellement un ensemble de 200 relations sélectionnées au hasard et annotées par les types d'EN. L'annotation distinguait les relations correctes, les relations incorrectes du fait d'un problème de reconnaissance des EN et les relations fausses du fait de l'absence de relation. Les relations incorrectes du fait des EN (20%) ont été écartées pour l'entraînement et le test des classifieurs. Le corpus résultant se compose donc de 964 relations, 531 étant correctes et 433 étant fausses, ce qui constitue un ensemble suffisamment équilibré pour l'apprentissage des modèles statistiques.

Plusieurs de ces modèles ont été testés en se concentrant d'abord sur des modèles exploitant un ensemble de caractéristiques non structurées[2]. Les différents classifieurs ont été entraînés en utilisant le même ensemble de caractéristiques, classiques pour l'extraction de relations, à l'instar de (Banko & Etzioni, 2008) :

– le type des entités nommées E1 et E1 ;
– la catégorie morpho-syntaxique des mots situés entre les deux entités avec un trait binaire pour chaque couple (*position dans la séquence*, *catégorie*), les bigrammes de catégories morpho-syntaxiques entre E1 et E2 avec un trait binaire pour chaque triplet (*position i*, cat_i, cat_{i+1}) ;
– la catégorie morpho-syntaxique des deux mots précédant E1 et des deux mots suivant E2, à la fois en tant qu'unigrammes et en tant que bigrammes ;
– la séquence des catégories morpho-syntaxiques entre E1 et E2. Chaque séquence possible de 10 catégories est encodée comme une caractéristique binaire ;
– le nombre de mots entre E1 et E2 ;
– le nombre de signes de ponctuation (virgule, guillemet, parenthèse ...) entre E1 et E2.

Nous avons également testé un classifieur prenant en compte la notion de séquence en nous appuyant sur les *Champs Conditionnels Aléatoires* (CRF). Dans ce cas, la tâche considérée prend la forme d'un étiquetage cherchant à associer à chaque mot d'une phrase l'une des quatre étiquettes suivantes : O pour un mot de la phrase en dehors d'une relation, ENT pour une EN définissant une relation potentielle (E1 ou E2), B-REL pour le premier mot d'une relation suivant E1, I-REL pour un mot faisant partie d'une relation. Dans ce schéma, une relation est jugée correcte lorsque l'étiquetage suit une configuration de type O – ENT – B-REL – I-REL* – ENT – O tandis qu'elle est jugée fausse si l'étiquetage produit une configuration O – ENT – O* – ENT – O. Ce modèle à base de CRF linéaires s'appuie sur l'ensemble de caractéristiques suivant :

– la catégorie morpho-syntaxique du mot courant, du mot précédent et du mot suivant ;
– les bigrammes de catégories morpho-syntaxiques $< \text{cat}_{i-1}, \text{cat}_i >$, avec i=-1,0,1 (0 est le mot courant) ;
– le type d'entité nommée du mot courant et de chacun des 6 mots le précédant et le suivant. Ce type peut avoir une valeur NIL lorsque le mot ne fait pas partie d'une entité nommée.

Une validation croisée a été faite pour évaluer ces différents classifieurs en découpant le corpus en 10 parties égales. Le tableau 1 montre les résultats obtenus par le SVM, le meilleur des premiers classifieurs testés, et les CRF. L'intégration par ces derniers d'un modèle de séquence leur confère une légère supériorité par rapport au SVM. C'est donc ce modèle que nous avons retenu pour le filtrage des relations. On notera également un certain équilibre entre la précision et le rappel. Enfin, comme le montre la dernière ligne du tableau 1, ces résultats se comparent favorablement à ceux de (Banko & Etzioni, 2008) sur le même sujet mais sur un corpus différent constitué de documents Web. Dans ce dernier cas, la précision est plus forte que la nôtre mais le rappel très largement inférieur. Il faut néanmoins préciser que dans (Banko & Etzioni, 2008), les relations extraites peuvent faire intervenir des entités plus générales que des entités nommées, ce qui est *a priori* un facteur de difficulté. En revanche, le corpus d'apprentissage de (Banko & Etzioni, 2008) est constitué de relations sélectionnées sur la

[2]Nous avons ainsi entraîné quatre types de classifieurs : bayésien naïf, maximum d'entropie (MaxEnt), arbre de décision et Machines à Vecteurs de Support (SVM).

Modèle	Exactitude	Précision	Rappel	F1-mesure
SVM	0,732	0,740	0,798	0,767
CRF	0,745	0,762	0,782	0,771
(Banko & Etzioni, 2008)	/	0,883	0,452	0,598

TAB. 1 – Évaluation des classifieurs statistiques

base d'heuristiques appliquées aux résultats d'un analyseur syntaxique. La gamme des expressions linguistiques apprises est donc limitée dans leur cas par les possibilités de l'analyseur syntaxique utilisé, problème auquel nous n'avons pas à faire face et qui peut expliquer pour partie leur faible rappel.

3.3 Application du filtrage des relations

L'extraction des relations se compose de 3 étapes : (1) une extraction initiale, en ne posant comme contraintes que la cooccurrence dans une phrase d'entités nommées de types donnés et la présence d'au moins un verbe entre les deux ; (2) le filtrage heuristique permettant d'écarter avec une bonne précision un grand nombre de relations fausses ; (3) l'application du filtrage statistique permettant de discriminer plus finement les relations correctes.

Type des relations	ORG-LIEU	ORG-ORG	ORG-PERS	PERS-LIEU	PERS-ORG	PERS-PERS
extraction initiale	71 858	77 025	73 895	152 514	126 281	175 802
heuristiques	33 505 (47%)	37 061 (48%)	32 033 (43%)	72 221 (47%)	66 035 (52%)	78 530 (45%)
classifieur CRF	16 700 (23%)	17 025 (22%)	12 098 (16%)	55 174 (36%)	50 487 (40%)	42 463 (24%)
dédoublonnage	15 226 (21%)	13 704 (18%)	10 054 (14%)	47 700 (31%)	40 238 (32%)	38 786 (22%)

TAB. 2 – Volume et pourcentage, relativement à leur nombre initial, des relations après chaque étape de filtrage

Le constat de la présence dans nos relations filtrées d'un certain nombre de relations identiques, pour une part issues d'articles sur un même sujet ou d'articles correspondant à des rubriques très formatées, nous a conduit à compléter ce processus de filtrage par un dédoublonnage final visant à éliminer ces relations redondantes. Il est à noter que cette opération de dédoublonnage vient en dernière position, à la fois du fait de son coût, plus important que celui des autres opération de filtrage et de sa dépendance vis-à-vis de l'évaluation de la similarité entre les relations, exploitée ensuite directement pour le regroupement des relations. Si ce filtrage rejette un grand nombre des relations extraites initialement, le volume des relations restantes est *a priori* suffisant pour alimenter les étapes suivantes du processus d'EI. Par ailleurs, nous nous situons dans un contexte de traitement de volumes textuels importants caractérisés par une certaine redondance informationnelle conduisant à privilégier la précision des processus d'extraction afin d'éviter un bruit trop important.

4 Impact du filtrage des relations sur leur regroupement

4.1 Regroupement des relations

À l'instar de beaucoup de travaux dans le domaine de l'extraction d'information non supervisée comme (Shinyama & Sekine, 2006) ou (Rosenfeld & Feldman, 2007), notre objectif final est le regroupement des relations selon leur similarité, en particulier pour en faciliter l'exploration. Mais notre but ici étant d'étudier l'influence du filtrage sur le regroupement, nous nous contenterons d'une approche simple (Hasegawa *et al.*, 2004) visant à rassembler les relations équivalentes sur le plan sémantique. L'équivalence de deux relations est évaluée par la mesure *cosinus* appliquée à une représentation de type « sac de mots » de la caractérisation linguistique des relations. Seule la partie *Cmid* de cette caractérisation est prise en compte.

Pour le regroupement, nous avons choisi l'algorithme *Markov Clustering* (van Dongen, 2000). Cet algorithme partitionne un graphe de similarité en clusters disjoints en réalisant une série de marches aléatoires dans ce graphe.

L'hypothèse est ici qu'un cluster se caractérise par une forte densité de liens entre ses éléments et qu'en « sortir » ne peut donc se faire qu'après un nombre important de pas. Cet algorithme itératif converge en pratique rapidement et se montre capable de faire face à nos graphes composés de quelques dizaines de milliers de nœuds. Par ailleurs, il détermine de manière intrinsèque le nombre de clusters à former et n'est dépendant pour ce faire que d'un seul paramètre – le facteur d'inflation – que nous avons laissé à sa valeur par défaut.

Le *Markov Clustering* s'appuyant sur un graphe de similarité des éléments à regrouper, le problème de son calcul pour plusieurs dizaines de milliers de relations se pose ici. Nous avons eu recours pour ce faire à l'algorithme *All Pairs Similarity Search* (Bayardo *et al.*, 2007) qui permet, moyennant la fixation d'un seuil de similarité minimale, de calculer efficacement et de manière exacte une mesure telle que *cosinus* pour tous les couples d'éléments dont la similarité est supérieure au seuil fixé. Cette valeur de similarité minimale a été établie grâce au *Microsoft Research Paraphrase Corpus*, qui rassemble un ensemble de couples de phrases[3] associées à un jugement de paraphrase. Le calcul de la mesure *cosinus* pour tous ces couples nous a permis de retenir une valeur minimale de 0,45 pour les expérimentations de la section suivante, seuil correspondant aux trois-quarts des valeurs calculées.

4.2 Évaluation de l'impact du filtrage des relations

Nous avons cherché à évaluer l'impact du filtrage des relations sur le regroupement de relations en comparant la qualité des regroupements formés avec ou sans filtrage à la suite de la méthode de regroupement décrite ci-dessus. Une classification de référence n'existant pas pour des relations telles que les nôtres, nous avons utilisé des mesures internes d'évaluation des regroupements. Les mesures de ce type permettent d'établir si le regroupement obtenu reflète bien les valeurs de similarités dans l'espace des relations. L'objectif est ici de tester si l'espace des relations après filtrage présente une meilleure distribution des similarités vis-à-vis de sa capacité à regrouper les relations. Parmi les différentes mesures existantes, nous avons retenu la mesure de la densité attendue (*expected density*), qui est évaluée dans (Stein *et al.*, 2003) comme la mesure ayant la meilleure corrélation avec la mesure externe de F-mesure pour le clustering de documents (la mesure usuelle de l'indice de Dunn étant jugée moins stable). Nous avons utilisé dans nos expériences une version modifiée de cette mesure, moins dépendante de la taille de l'ensemble des objets à regrouper, définie par ρ dans l'équation (1). De façon complémentaire, nous avons également utilisé la mesure de connectivité (*connectivity*) (Handl *et al.*, 2005), définie par c dans l'équation (2), qui évalue dans quelle proportion les relations des plus proches voisins ne sont pas coupées par le clustering. Cette mesure est intéressante dans notre cas parce qu'elle s'appuie sur la structure du graphe de similarité que nous utilisons aussi pour la méthode de clustering[4]. Les deux mesures se définissent ainsi à partir d'un graphe pondéré (V, E, w), où V est l'ensemble des nœuds, E l'ensemble des arcs et w, la pondération des arcs.

$$\rho = \sum_{i=1}^{|C|} \frac{|V_i|}{|V|} \frac{\theta_i}{\theta} \quad (1) \qquad\qquad c = \sum_{i=1}^{|V|} \sum_{j=1}^{p} x_{i,nn_i(j)} \quad (2)$$

Pour la définition de ρ, C est l'ensemble des clusters, V_i les nœuds du cluster i, et θ une mesure de densité des objets à regrouper, définie par $\theta = ln(w(G))/ln(|V|)$, avec $w(G) = |V| + \sum_{e \in E} w(e)$, et θ_i est la même mesure pour le graphe restreint aux nœuds du cluster C_i. Pour la définition de c, p est le nombre de voisins considérés, $nn_i(j)$, le $j^{ième}$ plus proche voisin de i et $x_{i,nn_i(j)}$, égal à 0 si i et $nn_i(j)$ sont dans le même cluster et $1/j$ sinon. Ces deux mesures évoluent de façon inverse : plus la mesure ρ est élevée, plus le clustering est jugé de bonne qualité, alors qu'un c plus bas est signe d'un meilleur clustering.

Les résultats de ces mesures sont présentés dans le tableau 3. Les deux mesures utilisées montrent que le clustering est dans la plupart des cas de meilleure qualité après le filtrage des relations, ce qui montre son utilité pour le regroupement des relations. Les deux couples d'entités pour lesquels cette tendance n'est pas vérifiée pour la mesure de densité attendue, ORG – LIEU et PERS – LIEU, ont en commun de faire intervenir des entités de type lieu. Ce constat s'explique peut-être par une spécificité de ces entités. En effet, lorsqu'une entité de type lieu est présente dans une phrase et qu'elle n'est pas en relation avec l'autre entité considérée dans la phrase, il est fréquent qu'elle soit inclue dans un complément circonstanciel de lieu. Or, avec la mesure de similarité choisie, les formes des compléments circonstanciels de lieu peuvent induire une similarité entre les phrases valide du point de vue d'un regroupement global et donc donner un bon score de clustering. Ces premiers résultats montrant

[3] Le corpus distingue un ensemble d'apprentissage et un ensemble de test que nous avons fusionnés dans le cas présent.

[4] Cette mesure est également dépendante de la taille du corpus : pour pallier ce problème, cette mesure est calculée pour un sous-ensemble de relations choisies aléatoirement, en l'occurrence 5000 relations présentes avant et après filtrage dans les résultats présentés.

	Expected density		*Connectivity* ($p = 20$)	
	avant filtrage	après filtrage	avant filtrage	après filtrage
ORG – ORG	1,06	**1,13**	5335,7	**3450,8**
ORG – LIEU	**1,13**	1,02	4458,7	**2837,6**
ORG – PERS	1,09	**1,17**	3025,4	**1532,4**
PERS – ORG	1,02	**1,06**	5638,0	**4620,0**
PERS – LIEU	**1.08**	1,07	5632,5	**4571,3**
PERS – PERS	1,13	**1,15**	3892,7	**2569,2**

TAB. 3 – Résultats de l'évaluation interne du regroupement des relations. Les résultats en gras sont les meilleurs scores (la mesure *expected density* doit être maximisée, la mesure *connectivity* doit être minimisée)

l'impact positif du filtrage pour le regroupement des relations ne donnent qu'une première tendance et doivent être confirmés par une évaluation s'appuyant sur une classification de référence.

5 Conclusion

Dans cet article, nous avons présenté un travail sur le filtrage de relations semi-structurées dans le contexte de l'extraction d'information non supervisée visant à déterminer si deux entités nommées cooccurrant dans une phrase sont en relation, sans *a priori* sur la nature de cette relation. Ce filtrage est réalisé par la combinaison d'heuristiques pour éliminer les cas les plus simples et d'un classifieur appris à partir d'exemples. Concernant ce dernier, les évaluations ont montré une légère supériorité des CRF sur les SVM. Dans la perspective du processus d'extraction non supervisée que nous développons, nous avons également caractérisé l'intérêt de ce filtrage pour le regroupement sémantique des relations par l'utilisation et l'adaptation de mesures internes d'évaluation de la qualité des regroupements formés.

Références

BANKO M., CAFARELLA M. J., SODERLAND S., BROADHEAD M. & ETZIONI O. (2007). Open Information Extraction from the Web. In *IJCAI-07*, p. 2670–2676.

BANKO M. & ETZIONI O. (2008). The tradeoffs between open and traditional relation extraction. In *48^{th} Annual Meeting of the ACL : Human Language Technologies (ACL-08 : HLT)*, p. 28–36.

BAYARDO R. J., MA Y. & SRIKANT R. (2007). Scaling up all pairs similarity search. In *16^{th} international conference on World Wide Web*, p. 131–140.

CIARAMITA M., GANGEMI A., RATSCH E., SARIC J. & ROJAS I. (2005). Unsupervised learning of semantic relations between concepts of a molecular biology ontology. In *IJCAI 2005*, p. 659–664.

GONZÁLEZ E. & TURMO J. (2009). Unsupervised relation extraction by massive clustering. In *Ninth IEEE International Conference on Data Mining (ICDM 2009)*, p. 782–787.

HANDL J., KNOWLES J. & KELL D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**(15), 3201–3212.

HASEGAWA T., SEKINE S. & GRISHMAN R. (2004). Discovering relations among named entities from large corpora. In *42^{nd} Meeting of the Association for Computational Linguistics (ACL'04)*, p. 415–422.

ROSENFELD B. & FELDMAN R. (2007). Clustering for unsupervised relation identification. In *Sixteenth ACM conference on Conference on information and knowledge management (CIKM'07)*, p. 411–418.

SEKINE S. (2006). On-demand information extraction. In *COLING-ACL 2006*, p. 731–738.

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

STEIN B., SVEN & WISSBROCK F. (2003). On Cluster Validity and the Information Need of Users. In *Proc. 3rd IASTED International Conference on Artificial Intelligence and Applications (AIA'03)*, p. 404–413.

VAN DONGEN S. (2000). *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht.