Extraction de connaissances pour le renseignement en sources ouvertes.

Laurie Serrano, Bruno Grilhères, Maroua Bouzid, Thierry Charnois

Extraction de connaissances pour le renseignement en sources ouvertes

Laurie Serrano*,**, Bruno Grilheres*, Maroua Bouzid**, Thierry Charnois**

*IPCC, CASSIDIAN
Parc d'Affaires des Portes, 27600 Val de Reuil
**GREYC, Université de Caen
Campus Côte de Nacre, Boulevard du Maréchal Juin, BP 5186 - 14032 Caen

Résumé. Cet article présente un outil d'extraction de l'information pour le renseignement sources ouvertes. Nous détaillons tout d'abord notre modélisation d'une ontologie de domaine destinée à structurer et partager les informations extraites. Puis, nous décrivons notre approche basée sur des techniques linguistiques visant à détecter les entités nommées, évènements et relations d'intérêt. L'implémentation de notre méthode grâce à la plateforme GATE ainsi qu'une évaluation des premiers résultats sont ensuite proposées. Nous concluons cet article en exposant nos perspectives de recherche dans le cadre d'une problématique plus large de capitalisation des connaissances.

1 Introduction

Aujourd'hui, l'abondance des informations accessibles publiquement (dites « sources ouvertes ») a fait émerger le besoin de « fouiller » cette masse de documents afin de repérer, structurer et partager les informations pertinentes et utiles dans un but donné. Les récents efforts en extraction d'information ont donné naissance à des techniques et outils permettant le repérage de ces informations d'intérêt (généralement les entités nommées, relations entre entités et évènements) (Poibeau, 2003). L'élaboration d'un tel système nécessite au préalable de définir la nature de ces informations afin de les partager avec d'autres services de traitement de l'information : l'ontologie de domaine est, à l'heure actuelle, le mode de représentation le plus utilisé dans ce but. Vient ensuite la phase centrale d'analyse textuelle visant à détecter et extraire ces types d'information. Nous pouvons ici distinguer plusieurs approches, parmi lesquelles émergent les techniques linguistiques d'une part et les systèmes statistiques à base d'apprentissage d'autre part. Les informations extraites sont ensuite destinées à peupler l'ontologie de domaine guidant l'extraction.

Nous avons développé un système d'extraction d'information pour l'anglais et le français fondé sur des techniques linguistiques. Nous proposons une extraction d'entités nommées basée sur la construction de grammaires et une approche innovante pour extraire les évènements et relations. Dans un premier temps, nous présentons une synthèse de nos travaux de modélisation et d'extraction de l'information, puis, une évaluation des premiers résultats obtenus. Pour finir, cet article détaille les perspectives de recherche que nous envisageons dans

le domaine de l'extraction d'information et plus généralement de la capitalisation des connaissances.

2 Vers une ontologie du renseignement

La construction de notre système d'extraction a nécessité, au préalable, de définir l'étendue et la nature des informations d'intérêt dans le domaine du renseignement sources ouvertes : c'est-à-dire mettre en place un modèle de connaissances. Nous avons choisi, pour cela, de développer une ontologie de domaine qui servira de guide aux différentes étapes d'extraction. Compte-tenu de la diversité des documents exploités dans le cadre du renseignement sources ouvertes, cette ontologie devra rester assez générale tout en définissant plus amplement les concepts et propriétés relatifs au domaine militaire.

2.1 Travaux existants

Suivant cet objectif, nous avons mené quelques recherches pour faire le point sur les ontologies existantes. En effet, il nous est apparu intéressant de pouvoir éventuellement reprendre tout ou partie d'une modélisation déjà disponible. Nous avons pour cela observé des ontologies générales, dites «de haut niveau» mais également des ontologies du domaine du renseignement.

Nous avons commencé par examiner les ontologies générales les plus connues et utilisées telles que SUMO (Niles et Pease, 2001), PROTON (Terziev et al., 2005), COSMO [1], OpenCyc [2], BFO (Spear, 2006) ou encore DOLCE (Gangemi et al., 2002). Dans un second temps, nous nous sommes intéressés aux ontologies disponibles pour le renseignement. Tout d'abord, nous avons étudié les recommandations de plusieurs standards OTAN. Ceux-ci sont des accords de normalisation ratifiés par les pays de l'alliance définissant des normes pour permettre les interactions entre les différentes armées. Nous avons accordé une attention particulière aux catégories de l'intelligence définies par le STANAG 2433 (NATO, 2005) mais aussi au STANAG 5525 (NATO, 2007). Ces standards restent trop généralistes et techniques mais nous avons pu tout de même nous inspirer des classes principales («pentagramme du renseignement») et de certaines propriétés. Enfin, les ontologies «swint-terrorism» (Mannes et Golbeck, 2005), reprenant les concepts principaux nécessaires au domaine du terrorisme, et AKTiveSA (Smart et al., 2007), dédiée à la description des contextes opérationnels militaires autres que la guerre, ont constitué d'autres exemples de modélisation.

Ces différentes modélisations ne correspondant pas exactement au modèle de connaissances défini précédemment nous avons fait le choix de définir notre propre ontologie en nous basant sur nos observations préalables.

2.2 Une proposition de modélisation

Nous avons fait le choix de baser notre ontologie sur 5 concepts correspondant aux éléments centraux du domaine du renseignement : *Units*, *Equipment*, *Places*, *Biographics*, *Events*.

1. http ://micra.com/COSMO/
2. http ://www.opencyc.org/

Une fois cette taxonomie de base définie, il nous a été nécessaire d'affiner notre représentation, guidés par le contexte du renseignement. Celle-ci comporte à l'heure actuelle environ 60 classes et 50 propriétés de classes. Par ailleurs, sa profondeur moyenne est de 3 niveaux et sa profondeur maximale s'élève à 4 niveaux. Nous avons réservé plus de temps à la modélisation de la classe «Event» (43% des classes et 41% des propriétés de notre ontologie) compte tenu de l'importance de ces entités pour le renseignement et la veille.

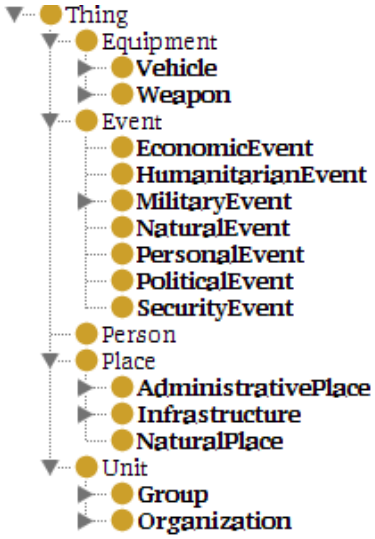

FIG. 1 – *Extrait de notre ontologie du renseignement*

Pour prolonger ces travaux de modélisation, il sera préférable d'interagir avec des opérationnels du renseignement sources ouvertes pour comprendre plus précisément leurs besoins et compléter notre ontologie en conséquence.

3 Extraction d'information pour le renseignement

3.1 Approches existantes

L'on distingue généralement deux types d'approche en extraction d'information : une extraction basée sur des techniques linguistiques d'un côté et des systèmes statistiques à base d'apprentissage de l'autre. Celles-ci se basent, de façon commune, sur des pré-traitements linguistiques «classiques» («tokenization», lemmatisation, analyse morpho-syntaxique).

La première approche exploite les avancées en TAL [3] et repose principalement sur l'utilisation de grammaires formelles construites par la main d'un expert-linguiste. Les pré-traitements cités plus haut servent de base à la construction de règles et patrons linguistiques qui définissent les contextes d'apparition de telle entité ou relation. La seconde approche utilise des techniques

3. Traitement Automatique de la Langue

statistiques pour « apprendre » des régularités sur de larges corpus de textes où les entités-cibles ont été préalablement annotées. Ces méthodes d'apprentissage ou « machine learning » sont plus ou moins supervisées et exploitent des caractéristiques textuelles plus ou moins linguistiques issues des pré-traitements précédemment évoqués.

3.2 Extraction des informations d'intérêt

L'outil que nous présentons ici a été réalisé selon une approche linguistique grâce à la plateforme d'ingénierie textuelle GATE (Cunningham et al., 2002) et vise à extraire les entités nommées, les évènements et les relations. Notre système repose sur le principe des chaines de traitements inhérent à GATE.

3.2.1 Extraction d'entités nommées

Nous avons, tout d'abord, développé une chaine d'analyse de textes anglais et français, permettant le repérage et l'extraction d'entités nommées de type « personne », « organisation », « lieu » et « date ». La plateforme GATE proposant déjà des chaines d'extraction pour l'anglais (ANNIE) et le français, nous avons fait le choix de les réutiliser en y apportant quelques modifications pour améliorer leurs performances et les adapter à notre représentation des connaissances. Pour cela, nous avons fait le choix de privilégier la précision par rapport au rappel. En effet, il nous est apparu plus important dans le contexte de nos travaux d'éviter l'extraction d'informations erronées. Concrètement, cela se traduit par la construction de règles linguistiques dont les résultats sont plus sûrs et la mise à l'écart de règles pouvant entrainer de fausses annotations. De plus, nous avons choisi de réorganiser les différentes phases d'extraction afin de parer à d'éventuelles ambiguïtés entre entités. La réutilisation de la chaine dédiée au français a nécessité de créer nos propres lexiques et règles linguistiques : une simple traduction du système anglais n'aurait pas suffi de par les nombreuses différences syntaxiques et typographiques entre ces deux langues.

3.2.2 Extraction d'évènements

Élément essentiel dans le cadre de la veille stratégique et du renseignement, un évènement peut être défini de façon générale comme une action (un « process ») reliée à un ou plusieurs participants ou circonstants. Avant tout essai d'implémentation, nous avons réfléchi à une nouvelle approche d'extraction d'évènements qui soit la plus générale possible et ceci à différents niveaux. Celle-ci devra, tout d'abord, être applicable à l'analyse de textes en plusieurs langues (français et anglais) et plusieurs domaines, mais aussi à différentes plateformes et environnements de traitement de la langue. De plus, notre méthodologie se voudra adaptée à l'utilisation de plusieurs analyseurs syntaxiques.

Tout d'abord, nous définissons un ensemble de termes considérés comme possibles déclencheurs d'évènement. Ces déclencheurs sont répartis en différentes listes, chacune étant associée à un type d'évènement, autrement dit à une classe d'évènement de notre ontologie. Par la suite, on repère et annote (en précisant la classe de l'ontologie associée) les termes déclencheurs présents dans le texte à analyser. Il nous faut ensuite leur associer les différents participants impliqués dans l'évènement qu'ils déclenchent. Pour cela nous devons repérer les relations entre le déclencheur et d'autres entités de la phrase. Il nous paraît alors judicieux

d'utiliser un analyseur syntaxique donnant les dépendances entre les différents éléments de la phrase. Par observation des sorties de différents analyseurs, nous avons pu établir une méthode générique de détection des participants. La plupart des analyseurs syntaxiques représente une relation de dépendance comme un lien entre la « tête » du syntagme-recteur et la « tête » du syntagme-dépendant. Ainsi, les participants de l'évènement correspondent aux syntagmes dépendants du mot déclencheur. Ceux-ci peuvent être extraits par une analyse en constituants délimitant les différents syntagmes d'une phrase : syntagmes nominaux, verbaux, prépositionnels ou adjectivaux. Une fois ces syntagmes rattachés à l'élément déclencheur par les relations de dépendance, il nous faut leur attribuer un rôle sémantique. Il nous faut, pour cela, réaliser une étude de la structure argumentale des termes déclencheurs : à savoir les rôles sémantiques de leurs différents actants. Pour cela, nous avons choisi de définir des classes argumentales, chacune d'elles correspondant à un type de construction syntaxique. Enfin, pour une meilleure extraction nous devrons prendre en compte d'autres paramètres tels que la voix (passive ou active) du syntagme verbal, la polarité de la phrase (négative ou positive), la modalité mais aussi les phénomènes de valence multiple.

Une fois cette méthodologie établie, nous avons développé une chaine d'extraction d'évènements dans des textes anglais (cf. Fig 2). Notre outil permet, à l'heure actuelle, de repérer les différents types d'évènements définis dans notre ontologie de domaine. Pour chaque événement détecté, nous avons pour objectif d'extraire, s'ils sont présents, les participants / circonstants suivants : la date, le lieu, l'agent et le patient. Nous avons, pour commencer, listé pour chaque type d'évènement un ensemble des termes susceptibles de le réaliser dans un texte. Nous choisissons de nous limiter, pour l'instant, aux déclencheurs verbaux et nominaux et de constituer des listes de lemmes, plus courtes et permettant d'étendre le repérage à toutes les formes fléchies. De plus, pour pouvoir déterminer les rôles sémantiques de chaque argument de l'évènement, nous avons associé à chaque lemme verbal une indication sur sa structure argumentale. Une fois les déclencheurs d'évènements repérés, nous procédons à une analyse syntaxique (grâce au Stanford parser) qui détermine l'ensemble des relations de dépendances. Par la suite, un transducteur JAPE exécute l'ensemble des règles linguistiques que nous avons développées pour extraire les différents arguments de l'évènement et leur attribuer un rôle sémantique.

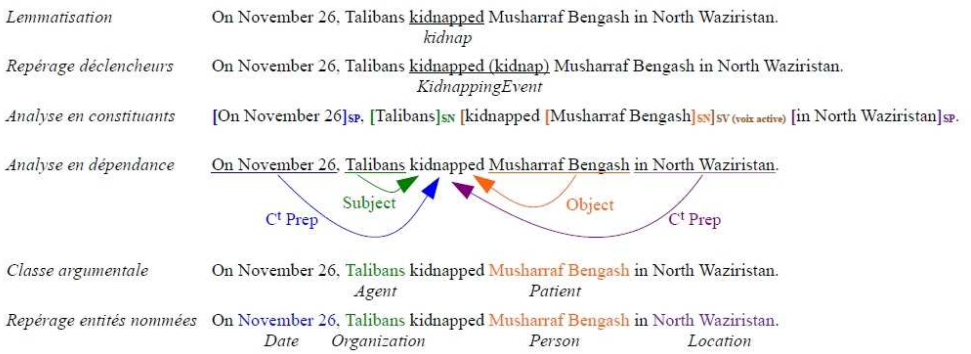

FIG. 2 – *Chaine d'extraction d'évènements pour l'anglais*

A l'heure actuelle, nous obtenons une annotation de type «Event» positionnée sur le déclencheur de l'évènement. Cette annotation présente les différents acteurs de l'évènement ainsi que leur rôle et indique pour chacun d'eux s'il correspond à une entité nommée détectée précédemment.

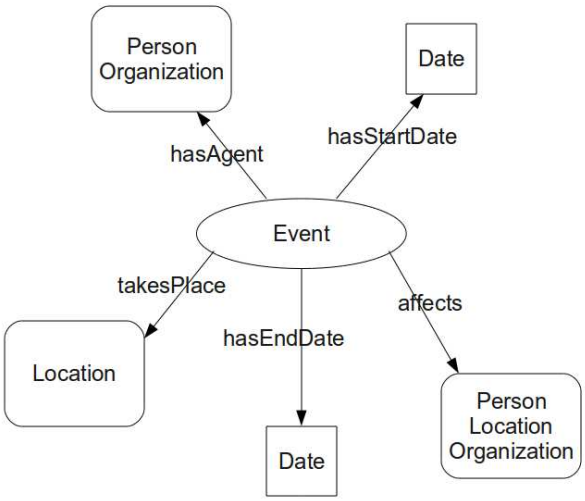

FIG. 3 – *Schéma de l'évènement et ses participants/circonstants*

3.2.3 Extraction de relations

Pour finir, nous nous sommes intéressés à la détection de relations entre entités nommées. Notre méthode consiste à définir, par observation de corpus du domaine visé, un ensemble de mots déclencheurs pour chaque type de relation de l'ontologie. Ceux-ci sont ensuite comparés aux mots du texte à traiter qui, s'il y a correspondance, sont annotés en tant que possibles déclencheurs de relation. Dans un deuxième temps, un ensemble de règles linguistiques teste le contexte de chaque déclencheur pour déterminer la présence d'une relation du type souhaité. Lorsque le contexte concorde, une annotation, dont le type est donné par l'élément déclencheur, est créée sur l'ensemble des éléments de la relation (arguments et lien). Un premier prototype a été réalisé pour l'extraction de relations entre personnes et organisations (appartenance/direction, parenté) dans des textes anglais.

3.3 Premiers résultats

Pour estimer l'efficacité de notre outil, nous avons mené deux types d'évaluation : une évaluation chiffrée de l'extraction d'entités nommées et une évaluation qualitative des deux autres extractions.

Évaluer notre extraction d'entités nommées en anglais et en français a, tout d'abord, nécessité de faire plusieurs choix concernant le protocole à mettre en place. Deux solutions se

sont alors présentées : réutiliser les données d'une campagne d'évaluation existante ou créer notre propre système d'évaluation. Le premier cas impliquait de trouver un corpus du domaine, où les entités nommées ont été préalablement annotées et qui soit accompagné de scripts de « scoring ». Nous avons, pour cela, examiné les données librement diffusées de campagnes d'évaluation telles que MUC (Grishman et Sundheim, 1996), ACE (Doddington et al., 2004) ou ESTER (Gravier et al., 2004) mais n'avons pas trouvé de données répondant à notre besoin. En conséquence, nous avons opté pour la deuxième solution, c'est-à-dire développer notre système d'évaluation. Pour cela, nous avons choisi deux corpus d'évaluation ayant pour thématique le renseignement : le corpus AQUAINT [4] pour l'anglais (une centaine de textes) et un corpus français de même taille composé de dépêches AFP sur le thème de l'Afghanistan. Une fois ces corpus annotés par notre système, les résultats de l'extraction ont été évalués manuellement par quatre spécialistes du « media mining ». Le tableau suivant présente nos résultats.

	Anglais			**Français**		
	Précision	**Rappel**	**F-mesure**	**Précision**	**Rappel**	**F-mesure**
Dates	0,99	0,73	0,84	1,00	0,70	0,83
Lieux	0,98	0,93	0,95	0,99	0,91	0,95
Organisations	0,86	0,73	0,79	0,98	0,78	0,87
Personnes	0,98	0,86	0,92	0,96	0,94	0,95
Entités	0,96	0,86	0,91	0,98	0,85	0,91

TAB. 1 – *Évaluation de l'extraction d'entités nommées*

Tout d'abord, au vu de la F-mesure globale (« Entités »), notre système d'extraction obtient de bons résultats en anglais et en français, la majorité des outils d'extraction d'entités nommées atteignant 90% de F-mesure sur des tâches similaires (Marrero et al., 2009). Nous pouvons dire que notre extraction d'entités nommées atteint l'état de l'art sans oublier pour autant les variations dues aux conditions d'évaluation (corpus, métriques, accord inter-annotateurs, etc.). Par ailleurs, nos résultats s'avèrent globalement équivalents dans les deux langues et, de par notre objectif de départ, notre outil montre une précision supérieure au rappel pour tous les types d'entités. Toutefois, notre système présente encore quelques faiblesses comme en témoignent les scores pour les entités de type « date » et « organisation ».

Notre extraction d'évènements se distingue par une approche originale sur plusieurs points. Tout d'abord, l'utilisation d'un lemmatiseur permet de repérer un évènement ou une relation au travers de ces différentes réalisations linguistiques tout en limitant la taille des lexiques. Par ailleurs, l'analyse syntaxique contribue de façon double à une meilleure extraction des participants : les constituants, d'une part, aident à la délimitation et les dépendances, d'autre part, servent pour le repérage « à distance » et l'attribution des rôles sémantiques. Relevons maintenant les différents points sur lesquels compléter nos extractions d'évènements et de relations. La limite principale concernant les évènements reste l'absence d'implémentation pour le traitement de textes en français. En effet, le manque d'analyseur syntaxique open source et gratuit pour cette langue a constitué un réel obstacle au développement de notre outil. Toutefois, même si un analyseur performant améliore grandement l'extraction d'évènements, notre système peut être amélioré par d'autres moyens. Tout d'abord, il sera nécessaire d'exploiter

4. http ://www-nlpir.nist.gov/projects/aquaint/

toutes les dépendances données par l'analyseur syntaxique pour mieux détecter tous les participants de l'évènement. Il faudra également étudier la construction des déclencheurs nominaux afin de déterminer des classes sur le même principe que pour les verbes. Pour finir, notre outil peut être raffiné en prenant en compte la négation, la modalité et les compléments prépositionnels dans la détection des évènements.

Enfin, notre extraction de relations devra être améliorée sur plusieurs points et implémentée pour les deux langues en question. Nous pourrons pour cela étudier les travaux de Nakamura-Delloye et Villemonte De La Clergerie (2010) proposant de repérer les chemins syntaxiques entre deux entités afin de construire par généralisation un ensemble de patrons de relations syntaxiques spécifique à tel type de relation sémantique.

4 Perspectives : Vers une capitalisation des connaissances orientée utilisateur

Ces premiers travaux nous ont permis de faire le point sur l'efficacité de notre méthode d'extraction et de mettre à jour d'autres problématiques à étudier. Nous souhaitons poursuivre nos réflexions pour tenter de répondre au constat suivant : s'il existe de nombreux systèmes d'extraction d'information, ceux-ci s'avèrent souvent peu fiables et ne prennent pas suffisamment en compte la dimension « connaissance », point essentiel dans ce domaine. Nos recherches viseront donc à définir une approche de capitalisation des connaissances permettant de faciliter et de réduire le travail des opérationnels dans le cadre de la veille économique et stratégique et du renseignement. Pour cela, notre objectif est d'explorer les méthodes d'extraction et de structuration automatique des informations accessibles en sources ouvertes. Les travaux que nous envisageons s'articulent autour de deux axes : « extraction d'information » et « capitalisation des connaissances ».

Au sein de l'axe « extraction d'information », nos recherches viseront à élaborer une approche aussi performante que possible et bien adaptée à nos objectifs. Dans ce but, nous privilégierons une combinaison des meilleures approches actuelles : les observations faites jusqu'à présent nous orientent d'ores et déjà vers le choix d'une approche hybride, combinant des techniques d'extraction à la fois linguistiques et statistiques. Il nous faudra, par ailleurs, tenter de dépasser certaines limites caractéristiques des systèmes d'extraction actuels telles que leur dépendance aux domaine/genre/langue des textes traités ou encore leur coût de développement (construction manuelle de règles/corpus annotés). Un autre objectif sera d'affiner les techniques d'évaluation existantes pour représenter la précision et la confiance à accorder aux informations extraites. Pour cela, nous tenterons de proposer des techniques d'auto-évaluation, prenant en compte l'incertitude ou l'imprécision des règles d'extraction.

L'axe « capitalisation des connaissances » aura pour objectif une structuration automatique des informations, leur capitalisation en base de connaissances mais également leur mise à jour et leur exploitation par des méthodes de raisonnements. Dans un premier temps, il s'agira de structurer les informations extraites en fiches de connaissances ; où une connaissance serait définie non pas comme le contenu d'un document mais bien comme un ensemble organisé d'informations collectées sur plusieurs textes. De plus, afin d'arriver à une réelle structuration et non plus à une simple addition de l'information, nous devrons prendre en compte les

problématiques de continuité de l'information (redondance et contradiction) mais également la temporalité et la modalité discursive exprimée au sein du document.

5 Conclusion

Nous venons de présenter un système d'extraction de l'information pour le renseignement sources ouvertes fondé sur une ontologie de domaine. L'évaluation reportée nous a permis de percevoir plus globalement la qualité de notre extraction mais aussi de mettre en avant certaines limites qui devront être dépassées dans le futur. Notre approche à base de règles contextuelles atteint l'état de l'art pour l'extraction d'entités nommées en anglais et français. Notre méthode d'extraction d'évènements montre le caractère indispensable d'une analyse syntaxique dans le repérage de telles informations. Celle-ci nécessite un analyseur fiable et robuste et nous avons pu constater que de nombreux efforts sont encore à mener dans ce sens en termes de performances et de langues traitées. Par ailleurs, les limites des systèmes linguistiques et statistiques actuels nous orientent vers une future combinaison de ces approches pour une meilleure extraction. Enfin, il s'agira, pour aller plus loin, de donner à l'utilisateur les moyens de tirer partie de ces informations pour acquérir de nouvelles connaissances.

Références

Cunningham, H., D. Maynard, K. Bontcheva, et V. Tablan (2002). Gate : A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA*.

Doddington, G., A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, et R. Weischedel (2004). The Automatic Content Extraction (ACE) program - Tasks, Data, and Evaluation. *Proceedings of LREC 2004*, 837–840.

Gangemi, A., N. Guarino, C. Masolo, A. Oltramari, R. Oltramari, et L. Schneider (2002). Sweetening ontologies with dolce. pp. 166–181. Springer.

Gravier, G., J.-F. Bonastre, E. Geoffrois, S. Galliano, K. M. Tait, et K. Choukri (2004). Ester, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français.

Grishman, R. et B. Sundheim (1996). Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*, Morristown, NJ, USA, pp. 466–471. Association for Computational Linguistics.

Mannes, A. et J. Golbeck (2005). Building a terrorism ontology. In *ISWC Workshop on Ontology Patterns for the Semantic Web*.

Marrero, M., S. Sanchez-Cuadrado, J. M. Lara, et G. Andreadakis (2009). Evaluation of named entity extraction systems. *Advances in Computational Linguistics. Research in Computing Science 41*, 47–58.

Nakamura-Delloye, Y. et E. Villemonte De La Clergerie (2010). Exploitation de résultats d'analyse syntaxique pour extraction semi-supervisée des chemins de relations. In *17e*

Conférence sur le Traitement Automatique des Langues Naturelles - TALN 2010, Montréal Canada.

NATO (2005). The military intelligence data exchange standard - aintp-3(b). Technical report.

NATO (2007). Joint c3 information exchange data model - jc3iedm. Technical report.

Niles, I. et A. Pease (2001). Towards a standard upper ontology. pp. 2–9. ACM Press.

Poibeau, T. (2003). *Extraction automatique d'information : Du texte brut au web sémantique*. Lavoisier.

Smart, P., A. Russell, N. Shadbolt, M. Shraefel, et L. Carr (2007). Aktivesa. *Comput. J. 50*, 703–716.

Spear, A. (2006). Ontology for the twenty first century: An introduction with recommendations. Technical report, The Institute for Formal Ontology and Medical Information Science.

Terziev, I., A. Kiryakov, et D. Mano (2005). Base upper-level ontology (bulo) guidance. Technical report deliverable 1.8.1, SEKT project.

Summary

This paper presents an information extraction tool for « open sources » intelligence. We first detail our domain ontology designed for structuring and sharing the extracted informations. Then, we describe an approach based on linguistic techniques to recognize named entities, events and relationships of interest. A first implementation of our method through the GATE architecture and our early results evaluation are then proposed. Finally, we introduce our planned researches within the knowledge capitalization broader field.