Une expérience de fusion pour l'annotation d'entités nommées

Caroline Brun(5), Nicolas Dessaigne(1), Maud Ehrmann(5), Baptiste Gaillard(4), Sylvie Guillemin-Lanne(3), Guillaume Jacquet(5), Aaron Kaplan(5), Marianna Kucharski(3), Claude Martineau(2), Aurélie Migeotte(1), Takuya Nakamura(2), Stavroula Voyatzi(2)

(1) Arisem – 1-5 rue Carnot- 91883 Massy cedex
{Nicolas.Dessaigne, Aurelie.Migeotte}@arisem.com

(2) IGM-LabInfo Université Paris-Est – 77454 Marne-la-Vallée Cedex 2
{claude.martineau, takuya.nakamura, stavroula.voyatzi}@univ-mlv.fr

(3) Temis –Tour Gamma B -193-197 rue de Bercy, 75582 Paris Cedex
{sylvie.guillemin-lanne, marianna.kucharski}@temis.com

(4) Thales Communication – 1-5 Avenue Carnot, 91883 Massy
Baptiste.gaillard@fr.thalesgroup.com

(5) XRCE – 6 chemin de Maupertuis, 38240 Meylan
{**Caroline.Brun, Maud.Ehrmann, Guillaume.Jacquet, Aaron.Kaplan**}**@xrce.xerox.com**

Résumé Nous présentons une expérience de fusion d'annotations d'entités nommées provenant de différents annotateurs. Ce travail a été réalisé dans le cadre du projet Infom@gic, projet visant à l'intégration et à la validation d'applications opérationnelles autour de l'ingénierie des connaissances et de l'analyse de l'information, et soutenu par le pôle de compétitivité Cap Digital « Image, MultiMédia et Vie Numérique ». Nous décrivons tout d'abord les quatre annotateurs d'entités nommées à l'origine de cette expérience. Chacun d'entre eux fournit des annotations d'entités conformes à une norme développée dans le cadre du projet Infom@gic. L'algorithme de fusion des annotations est ensuite présenté ; il permet de gérer la compatibilité entre annotations et de mettre en évidence les conflits, et ainsi de fournir des informations plus fiables. Nous concluons en présentant et interprétant les résultats de la fusion, obtenus sur un corpus de référence annoté manuellement.

Abstract In this paper, we present an experiment aimed at merging named entity annotations provided by different annotators. This work has been performed as part of the Infom@gic project, whose goal is the integration and validation of knowledge engineering and information analysis applications, and which is supported by the pole of competitiveness Cap Digital « Image, MultiMédia et Vie Numérique ». We first describe the four annotators, which provide named entity annotations that conform to guidelines defined in the Infom@gic project. Then we present an algorithm for merging the different annotations. It uses information about the compatibility of various annotations and can point out conflicts, and thus yields annotations that are more reliable than those of any single annotator. We conclude by describing and interpreting the merging results obtained on a manually annotated reference corpus.

Mots-clés : Entités nommées, fusion d'annotations, UIMA
Keywords: Named entities, fusion of annotations, UIMA

1 Introduction : le projet Infom@gic

Infom@gic est un projet de Recherche et Développement dans le domaine de l'Analyse de l'Information, cofinancé par des entreprises, des universités et des laboratoires d'Ile-de-France, ainsi que par la Délégation Générale des Entreprises, dans le cadre du Pôle de Compétitivité CAP DIGITAL[1]. Porté par Thales Communication, il regroupe un consortium de 29 partenaires. Son objectif, sur trois ans, est d'identifier et de développer les technologies clés dans le domaine de l'analyse de l'information et de l'extraction de connaissances, afin de fournir des outils permettant de naviguer aisément dans de gros volumes de données multimédia. Etant donnée l'approche multimodale du projet, l'ensemble des composants d'annotation, d'extraction, d'analyse, de transcription et de recherche a pour vocation d'être couplé dans des chaînes de traitement de l'information. Ces chaînes sont fédérées au sein d'outils standardisés et spécialisés constituant une plate-forme logicielle d'intégration de traitement de l'information. L'interopérabilité est assurée au moyen d'une plateforme commune UIMA (cf. section 3).

Parmi les différents cas d'usage du projet INFOM@GIC, le cas d'usage *Annotation et fusion des entités nommées* entre dans le sous projet d'extraction d'information dans les documents. Il réunit Arisem, IGM, TEMIS et XRCE qui ont mis en commun leurs annotateurs d'extraction d'entités nommées, leur expérience et leur savoir-faire. Les annotateurs identifient dans les documents issus du web des entités nommées et transfèrent ces annotations aux composants auxquels la plateforme est connectée. Suite au succès de cette étape, il s'est avéré utile de bâtir un outil de fusion des annotations d'entités nommées afin, d'une part, de gérer les doublons d'annotations issues des différents annotateurs et, d'autre part, de régler certains « conflits » d'annotations.

Cet article rend compte de notre expérience de fusion d'annotation d'entités nommées (EN dans la suite du document). Nous revenons tout d'abord sur les travaux existants pour la fusion d'annotation (section 2) puis décrivons rapidement la plateforme UIMA (section 3) ainsi que les différents annotateurs impliqués (section 4). Nous précisons ensuite la méthode de fusion adoptée (section 5) et présentons, enfin, les résultats obtenus lors de l'évaluation de ce retour d'expérience (section 6).

2 La fusion d'annotations : état de l'art

En apprentissage automatique, il est désormais bien établi que la combinaison d'un ensemble de méthodes ou de systèmes d'apprentissage permet souvent d'obtenir de meilleurs résultats qu'avec une méthode ou un système pris isolément. L'un des algorithmes très prometteurs qui fait actuellement l'objet de travaux de recherche importants est l'algorithme *Adaboost* (Schwenk, 1999) qui consiste à entraîner une cascade de systèmes similaires, chacun étant chargé de traiter les erreurs laissées par les systèmes précédents.

Ce principe trouve ses origines dans les expériences ROVER (Recognizer Output Voting Error Recognition) (Fiscus, 1997) qui ont été effectuées pendant les campagnes d'évaluation américaines pour la reconnaissance automatique de la parole. Il consiste à combiner les données produites par les systèmes à l'aide d'une simple stratégie de vote pour diminuer le

[1] Ce travail a été labellisé par le pôle de compétitivité CAP DIGITAL et financé en partie par la DGE et la Mairie de Paris.

nombre d'erreurs. Plus précisément, les votes s'appuient sur les fréquences maximales d'occurrence et une valeur seuil du score de confiance (les sorties des systèmes de transcription avaient été annotées avec un score de confiance (Chase, 1997)).

Pour ce qui est de l'annotation morpho-syntaxique, le principe de la combinaison de systèmes a été utilisé dans le passé par (Padro et Marquez, 1998) qui ont combiné deux systèmes d'annotation pour marquer un corpus et par (Tufis, 1999) qui a proposé d'utiliser plusieurs versions d'un même système, chacune entraînée sur des données de type différent. Cette approche a déjà connu de nombreux succès en désambiguïsation du sens des mots (Pedersen, 2000), et en analyse syntaxique (Henderson et Brill, 1999, Monceaux et Robba, 2003). Par ailleurs, le projet « Passage » (Paroubek *et al.*, à paraître) vise à utiliser plusieurs chaînes de traitement pour produire des annotations syntaxiques sur un corpus français d'au moins cent millions de mots, pour combiner ces annotations à l'aide de techniques de vote par majorité et pour utiliser ces annotations combinées pour des tâches d'acquisition de connaissances lexicales.

Enfin, pour ce qui est de la reconnaissance des EN, la méthode de vote par majorité a été utilisée pour l'anglais par (Borthwick *et al.* 1998) et, plus récemment, par (Kozareva *et al.*, 2007) qui ont combiné trois systèmes s'appuyant sur différentes méthodes d'apprentissage automatique pour annoter un corpus espagnol. Le système composite résultant a eu de meilleures performances (en précision) que n'importe lequel de ses modules élémentaires.

Notre expérimentation se situe dans le même esprit que les deux derniers travaux. Une des singularités de notre approche est d'avoir mis en place une plateforme commune à tous les partenaires où les quatre annotateurs ainsi que le module de fusion sont intégrés. Cette plateforme a été développée en vue d'une exploitation par d'autres outils d'extraction ou de recherche d'information. La chaîne de traitement implémentée traite du corpus web (avec tous les problèmes que cela peut poser). Enfin, nous rendons compte dans notre algorithme de fusion de cas d'inclusion ou de chevauchement d'annotations, qui constitue un apport nouveau vis-à-vis des précédents travaux.

3 La plateforme UIMA

UIMA (*Unstructured Information Management Architecture*) est une plateforme pour la gestion, l'organisation, et la coordination de l'information non-structurée (Ferrucci, Lally, 2004). Son développement a commencé chez IBM, et continue aujourd'hui en "open source" sous l'autorité de la fondation Apache (http://incubator.apache.org/uima/). Cette plateforme a été précisément développée pour la gestion d'outils de TAL ; l'exemple donné par Ferrucci et Lally est notamment de faciliter une intégration rapide d'un analyseur syntaxique avec un système de reconnaissance d'entités nommées. L'objectif déclaré de cette plateforme est d'accélérer les progrès scientifiques en permettant une combinaison rapide de technologies de traitement d'information non-structurée, dites UIM (*Unstructured Information Management*). UIMA a été choisi dans le cadre du projet Infom@gic pour faciliter l'interaction entre les technologies développées par les différents partenaires du projet. Chaque partenaire décrit dans la section suivante son annotateur qu'il a adapté au format UIMA afin d'interagir au sein de la même chaîne de traitement et d'y intégrer un nouveau module de fusion.

4 Les différents annotateurs d'EN

Dans cette section, les annotateurs d'entités nommées des différents partenaires sont présentés avec leurs spécificités. Ces systèmes ont été mis à jour selon les directives d'annotation mises au point dans le cadre du projet Infom@gic (c.f. livrable Infom@gic D2.11-5)

4.1 L'Annotateur d'Arisem

Arisem utilise une méthode symbolique pour la détection des entités nommées et de leurs relations. Des grammaires locales décrivent des règles lexico-syntactico-sémantiques faisant référence à des informations issues de dictionnaires et d'ontologies. Ces ressources sont exploitées par le moteur d'analyse HST (High Speed Transducer) qui les compile en automates afin d'optimiser le processus d'analyse sémantique.

L'annotateur Arisem se distingue par sa capacité à prendre en compte des connaissances métiers disponibles sous forme d'ontologies. Il est également capable d'utiliser ces connaissances pour résoudre des cas d'ambiguïtés sémantiques sur les entités nommées. Dans le cadre du projet, une normalisation des expressions numériques a également été développée (ex. : "trois mille dollars" => num:3.0e+3USD). L'annotateur Arisem supporte à ce jour le français, l'anglais, l'italien, l'espagnol, le portugais et l'allemand.

4.2 L'Annotateur de l'IGM

L'annotateur de l'IGM utilise l'environnement de développement *Open Source* multilingue et multiplateforme Unitex (Paumier 2003). Il utilise une méthode symbolique et dispose de deux types de ressources. D'une part des dictionnaires électroniques au format DELA (Courtois 1990) généraux et spécialisés, notamment toponymiques (Maurel *et al.* 1996) et, d'autre part de grammaires locales faisant appel à ces dictionnaires. Toutes ces grammaires utilisent le formalisme RTN (Recursif Transition Network) qui permet de créer aisément des sous-grammaires dédiées au traitement de chaque type d'Entités Nommées recherchées.

L'annotateur IGM permet de reconnaître et identifier les EN dans les textes, extraire certains attributs d'EN (ex. : la fonction ou l'ethnonyme associés à une EN de type « Person ») afin de les reformuler ou les normaliser. La tâche de normalisation touche plus particulièrement les EN mettant en jeu des données numériques. Une date ou une monnaie pouvant s'exprimer diversement, il est souvent utile d'en garder une représentation unique : sa forme normalisée. Ainsi, les EN *12 février 2009* et *35 €* ont pour formes normalisées respectives *2009-02-12* (norme ISO8601) et *35 EUR* (norme ISO4217). Ces normes facilitent la comparaison, le tri et les requêtes sur des données numériques.

4.3 L'annotateur de Temis Insight Discoverer™ Extractor

Contraction de Text Mining Solutions, TEMIS conçoit des applications dédiées à l'analyse et à la fouille de données textuelles. TEMIS dispose d'un serveur d'extraction d'information Insight Discoverer™ Extractor couplé à des Skill Cartridge™ L'ensemble fournit des applications dédiées à la veille stratégique et concurrentielle, à la gestion de la relation clients, de la connaissance, des savoir-faire et des ressources humaines.

L'information à extraire est modélisée et organisée selon une hiérarchie de composants de connaissance modulaires intégrables à différents domaines d'activité et/ou langues, appelée Skill Cartridge™. Un composant de connaissance peut avoir la forme d'un ou de plusieurs dictionnaire(s) ou d'un ensemble de règles d'extraction. L'objectif est de construire des patrons d'extraction (Yangarber et Grishman, 1997) suivant une approche guidée par le but (Appelt, 1993) (Poibeau, 2002). Le module d'extraction utilise la technologie des transducteurs (Hobbs 1997). Les Skill Cartridge™ développées au sein de TEMIS sont thématiques (intelligence économique, reconnaissance d'entités nommées), ou spécialisées par domaine d'activité (industrie pharmaceutique) (Aubry et al, 2002), l'information extraite peut servir d'entrée à des applications spécifiques métier (Guillemin-Lanne et Six, 2006).

4.4 L'annotateur de XRCE

XIP (Ait-Mokthar et al. 2002) est un analyseur développé au centre de recherche XRCE, dont l'objectif est d'extraire des dépendances syntaxiques profondes de façon robuste. Le formalisme proposé par XIP permet d'exprimer un large éventail de règles, de la désambiguïsation catégorielle à la construction de dépendances (Sujet, Objet, modifieurs, attributs etc.), en passant par la constitution de syntagmes noyaux.

Un système de détection des EN a été développé au sein de l'analyseur XIP. Il permet de détecter les types « standards » d'entités nommées (dates, lieux, personnes, organisations, monnaies …). Il s'agit d'un système symbolique, consistant en un ensemble de règles locales qui utilisent de l'information lexicale combinée à de l'information contextuelle sur les catégories syntaxiques. Ces règles locales sont très similaires à des règles d'identification de syntagmes noyaux, mais opérant au niveau du nom. Ce système de détection des entités nommées est intégré aux différentes grammaires syntaxiques. Récemment, les travaux conduits au sein du XRCE se sont orientés vers des méthodes de désambiguïsation des EN et de résolution de métonymie des EN (Brun, Ehrmann et Jacquet 2007).

5 La fusion d'annotations

L'objectif de cette fusion est de coordonner l'ensemble des annotations provenant des différents partenaires afin d'obtenir un annotateur unique bénéficiant des qualités spécifiques à chaque annotateur. De nombreuses difficultés sont à surmonter. La première, évoquée en introduction de la section 4, consiste à harmoniser les annotations : l'ensemble des partenaires doit se mettre d'accord sur un schéma d'annotation commun. Cette étape a fait l'objet d'un livrable (c.f. livrable Infom@gic D2.11-5). Toutefois l'annotation manuelle de notre corpus d'évaluation nous a confirmé, si besoin était, la difficulté de cette harmonisation. La deuxième difficulté consiste à éviter les doublons dans l'annotation fusionnée. D'apparence triviale, les cas de doublons correspondent tout de même à six cas différents répertoriés dans l'algorithme ci-dessous (1.a, 1.b, 2.a, 2.b, 3.a, 3.b). Enfin, la résolution de conflits constitue une autre difficulté (cf. section 5.2).

5.1 Algorithme

L'algorithme de fusion a été mis au point par les différents partenaires, suite à une étude tenant compte des différents cas de figures concernant les annotations, à savoir :

1. EN ayant le même offset de début et de fin (A = B) :
    a. Même type pour tous les annotateurs → Fusion directe.
    b. EN de types différents mais dans la même hiérarchie sans conflit de sous type → Le type le plus fin est sélectionné.
    c. EN de types différents mais dans la même hiérarchie avec conflit de sous type → Le type commun de plus haut niveau est sélectionné.
    d. Types différents mais pas dans la même hiérarchie → Supprimer les deux entités.
2. Inclusion de l'EN A dans l'EN B (A= Giscard, B=Giscard D'Estaing ou A= D'Estaing, B=Giscard D'Estaing ou A= Giscard, B=Valery Giscard D'Estaing)
    a. Même type pour tous les annotateurs → Seule l'entité B est conservée.
    b. Types différents mais dans même hiérarchie → Seule l'EN B est conservée avec son type
    c. Types différents mais pas dans même hiérarchie→ Les EN A et B sont conservées.
3. Chevauchement de AB et BC (AB= Valery Giscard, BC=Giscard D'Estaing)
    a. Même type pour tous les annotateurs→ Fusion de ABC.
    b. Types différents mais dans même hiérarchie → Fusion de ABC avec le type de plus haut niveau.
    c. Types différents mais pas dans même hiérarchie→ Les EN AB et BC sont conservées.

Dans les résultats ci-dessous, nous souhaitons privilégier une fusion qui améliore la précision plutôt que le rappel, c'est pourquoi nous avons choisi de ne garder que les annotations provenant d'au moins deux annotateurs différents. Cependant, les résultats de la fusion ne sont pas, de facto, égaux ou meilleurs que le meilleur annotateur. Par exemple, dans la phrase « Cette ville que, d'après *Cartier*, les Iroquois nomment Canada. », si deux annotateurs annotent *Cartier* en tant qu'<Organisation> et qu'un seul annotateur fourni l'annotation <Personne>, alors la fusion donnera l'annotation <Organisation>.

5.2 Exemples et discussion

Nous proposons d'illustrer quelques cas de conflits. Considérons le cas 1.b : A et B ont les mêmes offsets, et A ⇔ <Location>Lyon</Location>[2], B ⇔ <City>Lyon</City>. Dans ce cas, nous choisissons l'annotation la plus fine <City>Lyon</City>. En revanche dans le cas 1.c, A et B ont les mêmes offsets et les annotations sont A ⇔ <Country>Andorre</Country> et B ⇔ <City>Andorre</City>. Dans ce cas, les deux annotations sont conflictuelles, mais elles ont un parent en commun (différent de <NamedEntity>) : <Location>. C'est cette dernière annotation qui sera conservée.

Dans le cas 1.d, les annotations de A et B (dans les faits un ensemble d'annotations) sont conflictuelles et plusieurs choix s'offrent à nous : conserver les deux annotations en gardant l'information qu'il y a un cas de conflit entre annotateurs, supprimer les deux annotations en considérant que l'une des deux est nécessairement fausse et qu'il n'est pas possible de savoir laquelle. Nous avons choisi une troisième stratégie qui est le vote majoritaire : si plus de la moitié des annotateurs s'exprimant proposent la même annotation, celle-ci est gardée, sinon nous supprimons les deux annotations A et B (toutes les annotations conflictuelles). Ce choix va aussi dans le sens de privilégier la précision au détriment du rappel.

[2] La liste complète des types d'annotation utilisés se trouve dans le Tableau 1.

Le cas 2.c. peut-être illustré par l'expression « Association couleurs de Chine ». Dans ce cas, il est possible d'avoir l'annotation <Organisation> pour « Association couleurs de Chine » et l'annotation <Country> pour « Chine ». Dans un cadre applicatif, il nous semble intéressant de conserver les deux annotations. En revanche, dans le cadre de notre évaluation, nous n'avons pas annoté les entités incluses dans une autre entité (seule <Organisation> a été retenue).

6 Evaluation

Afin d'évaluer les performances de l'annotateur de fusion, les partenaires ont constitué un corpus de référence annoté manuellement selon les critères d'annotation définis dans le cadre d'Infom@gic. Dans la mesure où l'objectif est de pouvoir intégrer les annotateurs à un moteur de recherche sémantique (cf. livrable Infom@gic D1-11.22), ce corpus est constitué de données issues du web. Le schéma d'annotation se conforme à une hiérarchie de types constituée par les partenaires, laquelle correspond à une synthèse de l'expression des besoins applicatifs des différents partenaires ainsi que ceux du projet Infom@gic. Le Tableau 1 en est une simplification :

Annotations générales	**Annotations fines correspondantes**
ContactInformation	Address ; Email ; TelFax ; URL
Event	MilitaryEvent ; SportEvent ; CulturalEvent
Location	Continent ; Country ; Region ; City ; Microtoponym ; Hydronym ; Building
NumericalExpression	Area ; Length ; Money ; Percent ; Temperature
Organisation	Company ; SportOrganisation ; PoliticalOrganisation
Person	
TemporalExpression	Date ; Interval ; Time
Work	Product. Sous-type de Product : Vehicle

Tableau 1: synthèse de la hiérarchie de types

Le corpus de référence contient 3548 entités ; il est constitué de 4 fichiers, deux articles de Wikipédia (corpus « endurance » et « canada » ci-dessous) et deux pages de sites touristiques (corpus « lyon-evian » et « sete »).

Afin de bien interpréter les résultats de l'évaluation, il importe de souligner deux points. Le premier est relatif à la nature des corpus : issus du web, ces derniers sont au format XHTML et ne sont par conséquent pas aussi « propres » que les corpus sur lesquels les annotateurs travaillent habituellement. Malgré un ensemble de prétraitements (conversion de format, validation, correction/suppression de balises) réalisés automatiquement en amont au sein de la plateforme UIMA, les annotateurs ont été confrontés à des difficultés spécifiques liées au format des corpus. Le second point est relatif à l'annotation manuelle des corpus : réalisée par les partenaires eux-mêmes, elle n'est par conséquent peut-être pas totalement « objective ». Conscients de ce biais, nous avons cherché à le contrôler, en complétant l'évaluation de la fusion de l'ensemble des annotateurs sur l'ensemble des fichiers par une évaluation où, pour chaque fichier, nous avons exclu de la fusion l'annotateur provenant du partenaire ayant annoté manuellement ce même fichier (cf. colonne « fusion test » du Tableau 2).

Nous avons calculé, pour chaque fichier ainsi que globalement, les taux de précision, de rappel et la F-mesure. Nous présentons les résultats obtenus pour les types supérieurs de la hiérarchie (niveau « général », Tableau 2), et ceux obtenus pour les sous-types plus spécifiques (niveau « fin », Tableau 3). L'évaluation peut être considérée comme stricte dans la mesure où si un annotateur fournit la bonne annotation mais fait une erreur de frontière,

l'annotation est considérée comme fausse. Concernant l'évaluation sur les annotations générales, si un annotateur fourni une annotation fine, celle-ci est rapportée à l'annotation générale dont elle est le sous-type. Pour l'évaluation sur les annotations fines, les annotations générales ne sont pas évaluées : elles sont ignorées dans le corpus de test tout comme dans les réponses des annotateurs.

			Annot. 1	Annot. 2	Annot. 3	Annot. 4	**Fusion**	Fusion test
niveau d'annotation général	fichier 1 : canada	P	74.80	73.28	76.79	74.95	**81.80**	82.42
		R	61.58	41.00	51.61	66.88	65.76	60.29
		F2	67.55	52.58	61.73	70.69	**72.91**	69.64
	fichier 2 : endurance	P	72.52	**79.94**	66.96	67.83	79.50	78.21
		R	62.40	54.84	44.38	49.03	68.41	62.60
		F2	67.08	65.06	53.38	56.92	**73.54**	69.54
	fichier 3 : sete	P	63.36	57.06	74.24	74.89	**78.11**	79.29
		R	59.52	42.77	50.95	68.03	62.68	61.81
		F2	61.38	48.89	60.43	**71.30**	69.55	69.47
	fichier 4 : lyon-evian	P	69.94	53.73	71.66	81.72	**85.07**	83.09
		R	62.91	43.67	46.45	75.91	69.15	59.62
		F2	66.24	48.18	56.36	**78.71**	76.29	69.42
	TOTAL Pour niveau "général"	P	67.53	61.85	73.24	75.23	**80.10**	80.23
		R	60.85	44.36	49.38	66.35	65.11	61.30
		F2	64.02	51.67	58.99	70.51	**71.83**	69.50

Tableau 2: résultats de la fusion pour le niveau d'annotation "général"

			Annot. 1	Annot. 2	Annot. 3	Annot. 4	**Fusion**
niveau d'annotation fin	Continent	P	71.79	70.37	**93.33**	76.47	82.61
		R	93.33	63.33	46.67	43.33	63.33
		F2	**81.16**	66.67	62.22	55.32	71.70
	City	P	68.31	66.36	0.00	82.58	**85.13**
		R	40.79	53.81	0.00	80.34	66.09
		F2	51.08	59.43	0.00	**81.44**	74.41
	Microtoponym	P	0.00	46.43	0.00	0.00	**66.67**
		R	0.00	34.21	0.00	0.00	10.53
		F2	0.00	**39.39**	0.00	0.00	18.18
	Region	P	0.00	50.51	0.00	39.29	**50.70**
		R	0.00	29.07	0.00	44.77	20.93
		F2	0.00	36.90	0.00	**41.85**	29.63
	Country	P	84.96	79.35	77.46	86.12	**93.72**
		R	81.36	30.93	80.08	89.41	88.56
		F2	83.12	44.51	78.75	87.73	**91.07**
	Hydronym	P	0.00	53.97	0.00	**64.06**	63.33
		R	0.00	66.67	0.00	80.39	74.51
		F2	0.00	59.65	0.00	**71.30**	68.47
	Building	P	0.00	21.05	0.00	49.25	**56.10**
		R	0.00	4.94	0.00	40.74	28.40
		F2	0.00	8.00	0.00	**44.59**	37.70
	tous les sous-types de "Location"	P	70.56	60.13	73.94	72.38	**78.46**
		R	49.10	43.76	32.25	65.98	61.69
		F2	57.91	50.66	44.91	69.04	**69.07**
	TOTAL Pour niveau "fin"	P	69.14	59.27	73.04	72.87	**77.60**
		R	53.59	45.05	37.22	64.70	62.82
		F2	60.38	51.19	49.31	68.54	**69.43**

Tableau 3: résultats de la fusion pour le niveau d'annotation "fin"

Pour évaluer l'apport de la fusion, nous avons adopté le principe suivant : l'annotateur de fusion est considéré comme meilleur si sa précision est supérieure à celle du meilleur annotateur et si sa F-mesure n'est pas inférieure à celle du meilleur annotateur.

Le Tableau 2 nous montre que, pour chaque fichier, la précision de la fusion est toujours supérieure à celle du meilleur annotateur (ou équivalente pour le fichier 2) et que la F-mesure reste à plus ou moins deux points. La colonne « fusion test » montre les résultats obtenus sur chaque fichier en excluant de la fusion l'annotateur du partenaire ayant annoté manuellement ce fichier. Nous pouvons constater que les résultats restent stables : la fusion reste meilleure en précision et en F-mesure que les trois annotateurs restants et les résultats d'ensemble sont comparables à ceux de la fusion classique. Cela montre une certaine robustesse des résultats de cette fusion. Si l'on examine (cf. Tableau 3) les résultats obtenus à un niveau plus fin (ici, pour le type <Location> et ses sous-types), nous constatons également une meilleure performance de l'annotateur de fusion, avec une hausse de la précision par rapport à celle du meilleur annotateur et une F-mesure au moins équivalente. A ce niveau plus fin, il est intéressant de noter la complémentarité des partenaires : chaque annotateur présente de bonnes performances pour certains sous-types et non pour d'autres, mais la fusion permet d'allier les qualités de chaque annotateur et, au final, d'obtenir de meilleurs résultats[3].

Ainsi, tant pour le niveau d'annotation « général » que « fin », la fusion permet d'améliorer les résultats, avec une précision nettement supérieure à celle du meilleur annotateur (+ 4,87 pts pour l'annotation « générale », + 4,56 pts pour l'annotation « fine »), et une F-mesure légèrement supérieure (respectivement + 1,32 points et +0,89 points).

7 Conclusion

L'objectif de nos travaux consistait à réunir nos annotateurs afin d'améliorer et d'enrichir l'annotation des entités nommées. L'utilisation d'une architecture UIMA nous a permis d'assurer l'interopérabilité entre les différents systèmes ; la réalisation d'un module de fusion nous a ensuite permis de fédérer les annotations. Une des singularités de notre processus a été de traiter les conflits de frontière d'annotation lors de la fusion. L'évaluation finale, opérée sur des documents Web, montre qu'en fédérant les résultats des annotateurs nous améliorons la reconnaissance des entités nommées en tirant parti des atouts spécifiques à chacun. Dans le cadre d'Infom@gic, les annotations d'entités nommées provenant de notre plateforme sont intégrées à un outil de recherche sémantique sur le web, développé par la société Pertimm, dont l'objectif est de travailler sur de très gros volumes de documents textuels.

Références

AÏT-Mokthar S., CHANOD J.P., ROUX C. (2002). Robustness beyond Shallowness: Incremental Dependency Parsing. *Special issue of NLE Journal*.

AUBRY C., GRIVEL L., GUILLEMIN-LANNE S., LAUTIER C. (2002) « Aide à la construction de composants de connaissance pour l'extraction d'information : méthodologie et environnement» CIFT 2002, Hammamet- Tunisie.

BORTHWICK. A., STERLING J., AGICHTEIN E. ET GRISHMAN R. (1998). Exploiting diverse knowledge sources via maximum entropy in named entity recognition, Actes du *6th Workshop on Very Large Corpora,* 152-160.

[3] Pour le sous-type « Microtoponym » les résultats de la fusion sont non-nuls, pourtant un seul annotateur affiche des résultats non nuls. Ceci s'explique par le fait qu'il n'y a qu'un seul annotateur ayant donné des réponses justes, les autres ont fourni les bonnes annotations mais ont fait des erreurs de frontières et ont donc des résultats nuls. Ceci correspond au cas 2.a. de l'algorithme (cf. section 5.1) et explique les résultats de la fusion.

BRUN C., EHRMANN M., JACQUET G. (2007). XRCE-M: A Hybrid System for Named Entity Metonymy Resolution. Actes de *SemEval 2007*.

CHASE L. L. (1997). Word and Acoustic Condence Annotation for Large Vocabulary Speech Recognition, Actes de *Eurospeech 1997*, 815-818.

COURTOIS B. (1990). Un système de dictionnaires électroniques pour les mots simples du français, In Courtois B. et Silberztein M. (éds), *Dictionnaires électroniques du français*, *Langue Française,* 87, Larousse, Paris, 11-22.

FERRUCCI D., LALLY A. (2004). UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 327-348.

FISCUS G. J. (1997). A Post-Processing System To Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER), Actes de *IEEE Workshop on Automatic Speech Recognition and Understanding*, 347-354.

GUILLEMIN-LANNE S., SIX A. (2007) La normalisation : nouveau challenge en extraction d'information. Actes de *VSST 2006*.

HENDERSON C. J. et BRILL E. (1999). Exploiting Diversity in Natural Language Processing: Combining Parsers. Actes de *EMNLP-99*, 187-194.

HOBBS J. R. ET AL. (1997). FASTUS : A Cascaded Finite-State Transducers for Extracting Information from Natural-Language Text. dans E. Roche et Y. Schabes (eds.), *Finite-State Language Processin*g. Cambridge MA: MIT Press. (1997)

INFOM@GIC – Livrable D1-11.22. (2008). Intégration des annotateurs sémantiques.

INFOM@GIC – Livrable D2-11.5. (2006). Spécifications d'une norme de représentation des entités nommées.

KOZAREVA Z., FERRANDEZ O., MONTOYO A., MUÑOZ R. et SUAREZ A. (2007). Combining Data-Driven Systems for Improving Named Entity Systems, *Data & Knowledge Engineering*, 61:3, Elsevier Science Publishers B.V., Amsterdam, The Netherlands, 449-466.

MAUREL D., BELLEIL C., EGGERT E. et PITON O. (1996). Le projet PROLEX : réalisation d'un dictionnaire électronique relationnel des noms propres du français, Actes de *GDR-PRC Communication Homme-Machine Séminaire Lexique*, 164-175.

MONCEAUX L. et ROBBA I. (2002), Les analyseurs syntaxiques : atouts pour une analyse des questions dans un système de question-réponse ?, Actes de *TALN'02*, 195-204.

PADRO L. et MARQUEZ L.. (1998). On the evaluation and comparison of taggers: the effect of noise in test corpora, Actes de *COLING/ACL'98,* Canada.

PAROUBEK P., DE LA CLERGERIE E., LOISEAU S., VILNAT A. et FRANCOPOULO G. (à paraître). The PASSAGE Syntactic Representation, Actes de *LT7*.

PAUMIER S. (2003). *De la reconnaissance de formes linguistiques à l'analyse syntaxique*, Thèse de doctorat, Université de Marne-la-Vallée.

PEDERSEN T. (2000). A Simple Approach to Building Ensembles of Naive Bayesian Classifiers for Word Sense Disambigusation, Actes de *NAACL*, 63-69.

POIBEAU T. (2002) : Extraction d'information à base de connaissances hybrides, Thèse de Doctorat, Université Paris XIII.

SCHWENK H. (1999). Using boosting to improve a hybrid Hmm/Neural Network speech recognizer, Actes de *ICASSP*, 1009-1012.

TUFIŞ D. (1999). Tiered Tagging and Combined Classifiers , In Jelinek F. et E. Nöth (éds.), *Text, Speech and Dialogue, Lecture Notes in AI*, vol. 1692, Springer, 28-33.

YANGARBER R., GRICHMAN R. (1997), Customisation of Information Extraction Systems. Dans Pazienza M. T., editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, Springer Verlag, Heildelberg, 1-11.