La tâche de reconnaissance des entités nommées : état des lieux

1.1 Aperçu général

La tâche de reconnaissance d'entités nommées s'intéresse à un certain nombre d'unités lexicales particulières, que sont les noms de personnes, les noms d'organisation et les noms de lieux, ensemble auquel sont souvent ajoutés d'autres syntagmes comme les dates, les unités monétaires et les pourcentages[1]. Son objectif est double : il s'agit, d'une part, d'*identifier* ces unités dans un texte, et, d'autre part, de les *catégoriser* en fonction de types sémantiques prédéfinis. Le résultat de ces processus correspond à l'*annotation* des entités, laquelle se matérialise le plus souvent *via* des balises encadrant l'entité. C'est ainsi que, pour la phrase suivante,

> L'ancien premier ministre socialiste Lionel Jospin a confirmé, jeudi 28 septembre, sur RTL, qu'il ne sera pas candidat à l'investiture socialiste pour la présidentielle de 2007.

il importe de reconnaître les entités *Lionel Jospin*, *RTL*, etc., puis de leur attribuer un type sémantique correspondant. L'identification et la catégorisation des entités L'annotation peut alors se présenter comme suit :

> L'ancien premier ministre socialiste <PERS>Lionel Jospin</PERS> a confirmé, <DATE>jeudi 28 septembre</DATE>, sur <ORG>RTL</ORG>, qu'il ne sera pas candidat à l'investiture socialiste pour la présidentielle de <DATE> 2007</DATE>.

[1]Cette énumération correspond peu ou prou à la « définition » traditionnelle des entités nommées ; cette acception sera de mise dans ces premiers paragraphes, avant une discussion plus approfondie de ce point.

La reconnaissance des entités nommées peut être mise en œuvre pour tout types de textes. Si historiquement ce processus a été appliqué sur des corpus journalistiques traitant de sujets géopolitiques (cf. section 1.2), il est aujourd'hui également appliqué sur d'autres types de corpus portant sur des domaines plus spécifiques. Ceux de la biologie et de la médecine sont par exemple fort demandeurs de ce genre d'analyse, la reconnaissance des noms de gènes, de protéines ou de maladies aidant au traitement de l'importante quantité d'information produite par ces communautés. Dans le domaine médical, il s'agit de reconnaître des expressions telles que *maladie de Parkinson* ou *fièvre de Lassa* [Bodenreider et Zweigenbaum, 2000] et, en biologie, des noms tels que *p53* ou *interleukin 1 (IL-1)-responsive kinase* [Fukuda *et al.*, 1998]. Si diverses expressions peuvent ainsi être annotées dans une tâche de traitement des entités nommées, les entités relevées le plus couramment appartiennent aux types sémantiques généraux de PERSONNE, ORGANISATION et LIEU.

La tâche de reconnaissance des entités nommées s'attache donc à extraire et à typer certains éléments informationnels d'un texte, en parfaite dépositaire de l'extraction d'information dont elle est issue. Il importe de retracer cette filiation et, pour ce faire, de considérer la longue tradition de campagnes d'évaluation, américaines et centrées sur ce domaine de l'extraction d'information aux origines, puis s'internationalisant et se diversifiant par la suite.

1.2 Historique

1.2.1 L'extraction d'information

C'est en effet à la faveur du développement de la tâche d'extraction d'information que la tâche de reconnaissance des entités nommées est apparue. La recherche pour la conception de systèmes d'analyse de textes a, depuis les débuts du TAL, exploré diverses voies. C'est dans ce cadre que l'extraction d'information a succédé aux systèmes génériques de compréhension de textes, aux visées sensiblement trop ambitieuses, comme le souligne T. Poibeau et A. Nazarenko [Poibeau et Nazarenko, 1999]. L'extraction d'information, ne cherchant plus à comprendre l'ensemble du texte, vise à extraire d'un texte donné des éléments pertinents d'information, dont la nature a été spécifiée préalablement. Il s'agit ainsi d'identifier des occurrences d'événements particuliers, d'en extraire les arguments impliqués pour ensuite en donner une représentation structurée. L'analyse s'effectue au niveau local et seule une partie du texte est considérée. Cette tâche peut alors se définir, selon la formule de T. Poibeau, comme « l'activité qui consiste à remplir automatiquement une banque de données à partir de

textes écrits en langue naturelle » [Poibeau, 2003, p.13].

Si le principe sous-jacent de l'extraction d'information n'était pas nouveau [Grishman, 1997], cette tâche a gagné en maturité et s'est singulièrement précisée grâce à la série des conférences MUC (*Message Understanding Conferences*[1]). Ce cycle de conférences, organisé par diverses institutions américaines et financé par la DARPA (*Defense Advanced Research Projects Agency*), s'est déroulé de 1987 à 1998, motivant de la sorte de nombreuses équipes de recherche pendant plus d'une décennie. Comme leur nom l'indique, l'objectif de ces conférences était à l'origine d'encourager la recherche autour de la compréhension automatique de messages militaires. Baptisées « conférences », ces dernières sont en réalité des campagnes d'évaluation, au cours desquelles un certain nombre de participants se voient remettre, dans un premier temps, un corpus d'entraînement et des instructions précises sur les informations à en extraire automatiquement, puis, dans un second temps, un corpus de test sur lequel ils doivent appliquer leurs systèmes. Les résultats sont ensuite évalués et présentés lors de la conférence finale, à laquelle seuls les participants à l'évaluation ont le droit d'assister. L'histoire de ces conférences est désormais bien connue ; [Grishman, 1997, Hirschman, 1998, Poibeau, 2003] permettent d'en apprécier l'évolution de façon détaillée. Nous en retraçons ici les grandes lignes afin de mieux situer l'apparition de la tâche qui nous occupe, la reconnaissance des entités nommées, et avant d'examiner d'autres événements ayant eux aussi contribué à l'émergence de cette dernière.

1.2.2 Les conférences MUC

1.2.2.1 Les trois « cycles » de conférences

Il est possible de distinguer trois « cycles » au sein des 7 conférences qui se sont succédées, en fonction de la définition et de la difficulté de la tâche d'extraction à mettre en œuvre tout d'abord, de la taille et de la nature des corpus à analyser ensuite, et du degré d'aboutissement du processus d'évaluation enfin. Les deux premières conférences (1987 et 1989) forment un cycle liminaire que l'on peut qualifier d'« exploratoire » . Les corpus sont des messages de la *Navy* de style télégraphique et, après l'absence de toute instruction précise quant aux données à en extraire lors de la conférence de 1987, une premier formulaire simple de structuration de données (en anglais *template*) fait son apparition lors de la suivante en 1989. Sont également adoptées les premières mesures d'évaluation, précision et rappel, issues de la recherche d'information. Ces deux sessions pionnières, si elles n'ont révélé aucune méthode ou système particulier, ont le mérite

[1] http://www-nlpir.nist.gov/related_projects/muc/.

d'avoir rassemblé autour d'une même tâche plusieurs participants, ainsi amenés à discuter de leur travail et des moyens de l'évaluer.

Les conférences MUC-3, MUC-4 et MUC-5 constituent le second cycle, au cours duquel la tâche d'extraction d'information, telle que présentée ci-dessus et initiée par les précédentes conférences, s'est progressivement définie, gagnant en précision mais également en complexité. MUC-3 (1991) et 4 (1992) ont travaillé sur des corpus de nature journalistique, traitant d'événements ou d'actes terroristes en Amérique Centrale et du Sud. Les *templates* comportent alors de plus en plus de champs à remplir, ces derniers pouvant atteindre le nombre de 24. La figure 1.1[1] montre un exemple de formulaire à remplir pour MUC-3 : à partir d'une dépêche sur un acte terroriste, il importait d'en extraire le type d'incident, le lieu, la date, les exécutants, la cible ainsi que les effets sur cette dernière. Si les textes sont mieux écrits (moins de problèmes de casse, rédaction plus soignée et plus homogène), ils sont en revanche plus difficiles à analyser (plus longs, l'information à en extraire est plus difficile à identifier). Les collections de textes d'apprentissage sont distribuées en grand nombre et les premiers systèmes à base d'automates [Appelt *et al.*, 1993] ainsi que d'autres basés sur des méthodes statistiques font leur apparition. MUC-4 introduit également une nouvelle mesure d'évaluation, la F-mesure, qui combine les taux de précision et de rappel et rend ainsi plus facile les comparaisons entre systèmes. MUC-5 suit de près (un an) ces deux conférences et gagne encore en complexité : deux domaines sont proposés (technologique avec la microélectronique et commercial avec la vente d'entreprises) pour deux langues, anglais et japonais. Cette diversification correspond à une volonté d'améliorer la portabilité[2] des systèmes ; néanmoins, les temps de développement de ces derniers sont extrêmement longs (6 mois) et les niveaux de performance ne dépassent pas les précédents. Vue par certains comme un échec, cette dernière conférence de 1993 infléchit néanmoins de manière significative la vision de la tâche d'extraction d'information : devant traiter plusieurs domaines en plusieurs langues, les participants sont amenés à rendre plus génériques leurs architectures et certains modules d'analyse apparaissent comme nettement indépendants. MUC-5 marque ainsi un point d'aboutissement de ce deuxième cycle de conférence, révélant la nécessité de fragmenter en fonctionalités indépendantes une tâche d'extraction d'information devenue trop complexe.

Le dernier cycle est composé des conférences MUC-6 et MUC-7. La première, qui a lieu en 1995, marque un profond tournant pour ces campagnes d'évaluation : trois nouvelles tâches sont introduites et la tâche « traditionnelle » est simplifiée,

[1] Cette figure est issue de [Grishman, 1997].

[2] La portabilité désigne la capacité d'un système à être utilisé sans modification importante pour une autre application que celle pour laquelle il a été conçu.

19 March – A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unoffial sources, the bomb – allegedly detonated by urban guerrilla commandos – blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).

INCIDENT TYPE	bombing
DATE	March 19
LOCATION	El Salvador : San Salvador (city)
PERPETRATOR	urban guerrilla commandos
PHYSICAL TARGET	power tower
HUMAN TARGET	-
EFFECT ON PHYSICAL TARGET	destroyed
EFFECT ON HUMAN TARGET	no injury or death
INSTRUMENT	bomb

FIG. 1.1 – Exemple de formulaire d'extraction d'information pour des actes terroristes (MUC-3).

afin de répondre aux nouveaux objectifs fixés par le comité scientifique. Retenant les leçons du « cycle » précédent, MUC-6 se donne en effet pour programme de démontrer la possibilité de concevoir des systèmes indépendants pouvant facilement être utilisables, d'encourager à la réalisation de systèmes plus portables, et de favoriser des travaux pouvant contribuer à des modules de « compréhension profonde » [Grishman et Sundheim, 1995, Grishman et Sundheim, 1996]. Concernant ce dernier objectif, trois tâches sont envisagées : résolution de coréférence[1], désambiguïsation lexicale et détection de structure prédicat-argument. Baptisé « SemEval », ce pôle regroupant plusieurs tâches ne voit cependant se réaliser que celle de coréférence lors de la conférence. Au regard de la portabilité, un « mini-MUC » est adopté, avec la simplification du formulaire traditionnel d'extraction d'information reflétant les relations entre entités (*Scenario Template*), et un formulaire d'entité est proposé (*Template Element*), comportant 6 champs. Enfin, et surtout, la volonté de transformer certains modules impliqués dans le processus d'extraction d'information en véritables fonctionnalités indépendantes d'analyse de texte se traduit par la création de la tâche de reconnaissance des entités nommées (*Named Entities*). MUC-6 comporte donc au final les tâches suivantes :

- Entités Nommées
- Coréférence
- Formulaire des entités (*Template Element*)
- Formulaire des scénarios (*Scenario Template*, correspondant au » template« tra-

[1]Voir définition en 1.3.1.

ditionnel des MUCs)

Ce programme, certes ambitieux, reste cependant réalisable et correspond bien à l'esprit de la conférence de MUC-6 : en définissant des tâches séparées, le comité d'organisation cherche à encourager des participations pour telle ou telle tâche seulement, chacune d'elles se focalisant sur l'extraction d'un type d'information. Cette organisation de la tâche d'extraction d'information sous la forme de modules constitue la principale innovation de cette sixième conférence.

Le déroulement lui-même de cette conférence est en rupture avec les pratiques précédentes : les corpus, portant sur les changements de positions dans les entreprises, sont de taille inférieure et distribués pour une durée moindre[1] (temps d'apprentissage et temps de test), incitant ainsi les participants à concevoir des systèmes indépendants et nécessairement portables. Partant, cette conférence voit se multiplier les méthodes probabilistes et les systèmes à base d'apprentissage. Les niveaux de performances, pour chacune des tâches définies, sont encourageants, voire très encourageants : pour la tâche sur les entités nommées, des F-mesure supérieures à 0,9 sont atteintes par plusieurs systèmes (nous revenons plus précisément sur ces résultats un peu plus loin) ; le formulaire des entités (*Template Element*) est quant à lui rempli selon des taux de 75 % pour le rappel et de 86 % pour la précision et, enfin, le formulaire de scénario (ou mini-MUC) voit se réaliser des performances de l'ordre de 50 % pour le rappel et de 70 % pour la précision. Ces résultats prometteurs, au sortir de MUC-6, attestent donc de la nécessité (ou pour le moins de l'utilité) de décomposer la tâche d'extraction d'information pour, d'une part, obtenir des résultats probants et, d'autre part, faciliter la conception de systèmes génériques indépendants. Si d'importants progrès sont encore à réaliser pour permettre la portabilité des systèmes, le bilan de MUC-6 est en grande partie conforme aux objectifs fixés préalablement par le comité.

Organisée trois ans plus tard, MUC-7 ne poursuit pourtant pas le renouveau de la tâche d'extraction d'information avec autant de brio et d'enthousiasme que la précédente. Mise à part la généralisation des techniques d'apprentissage, peu de choses nouvelles sont à noter pour cette conférence de 1998, qui marque en réalité la fin des conférences MUC. Parallèlement à MUC-6 et MUC-7, une campagne d'évaluation multilingue en extraction d'information a été organisée : MET ou *Multilingual Entity Task*. Cette conférence a permis, entre autres, le développement de systèmes de reconnaissance d'entités nommées pour l'espagnol, le japonais et le chinois, expatriant avec succès cette tâche du monde anglophone vers d'autres langues.

[1] Se reporter aux figures 11 et 12 de [Hirschman, 1998].

Ce dernier cycle de la série des *Message Understanding*, s'il s'achève brutalement, n'en reste pas moins fondamental au regard de l'évolution de la tâche d'extraction d'information en général, et de la tâche de reconnaissance des entités nommées en particulier : celle-ci fait une apparition remarquée, et celle-là gagne en maturité. Après ce rapide survol des conférences MUC, considérons à présent la tâche sur les entités nommées d'un peu plus près.

1.2.2.2 MUC et la reconnaissance d'entités nommées

C'est ainsi à l'occasion d'une refonte de la tâche d'extraction d'information (MUC-6) qu'apparaît la tâche de reconnaissance des entités nommées. Pourquoi les entités nommées ? L'intérêt pour ce type de noms s'explique par le fait qu'ils sont présents dans tous types de textes, quel que soit le domaine ; ils constituent ainsi un point de passage obligé pour tout système cherchant à rendre compte de l'information contenue dans un texte. En effet, qu'il s'agisse de messages militaires ou de dépêches journalistiques portant sur des actes terroristes, sur des fusions d'entreprises ou encore sur de la microélectronique, l'essentiel est de repérer les actants ainsi que les coordonnées des événements relatés. La tâche d'extraction et de reconnaissance d'entités nommées lors de MUC-6 s'est ainsi focalisée sur les trois types d'entités suivants :

ENAMEX : pour les noms d'entités correspondant à des noms de personnes, d'organisations et de lieux. Les sous-types sont : PERSON, ORGANISATION et LOCATION.

TIMEX : pour les expressions temporelles. Les sous-types sont : DATE et TIME.

NUMEX : pour les expressions numériques, de monnaie et de pourcentage. Les sous-types sont : MONEY et PERCENT.

Relatant la préparation de la sixième conférence, les auteurs de [Grishman et Sundheim, 1995] illustrent la tâche d'extraction d'entités nommées à l'aide du texte de la figure 1.2, où sont annotées des entités de type ENAMEX (personne et organisation) et NUMEX.

Instaurée et définie de la sorte lors de la conférence de 1995, cette tâche est également présente lors de MUC-7 et de MET. Pour chaque campagne d'évaluation, les performances atteintes sont remarquables. Quinze participants (pour 20 systèmes) s'attellent à reconnaître ces entités dans des dépêches portant sur des changements de position dans les entreprises lors de MUC-6. Sur la totalité des systèmes, plus de la moitié affichent des taux combinés de rappel et de précision supérieurs à 0,9, et le premier d'entre eux réalise une F-mesure de 0,96. Ces résultats inattendus sont d'autant plus impressionnants qu'ils ap-

Mr. <ENAMEX TYPE=« PERSON » > Dooner </ENAMEX> met with <ENAMEX TYPE=« PERSON » > Martin Puris </ENAMEX>, president and chief executive officer of <ENAMEX TYPE=« ORGANIZATION » > Ammirati & Puris </ENAMEX>, about <ENAMEX TYPE=« ORGANIZATION » > McCann </ENAMEX>'s acquiring the agency with billings of <NUMEX TYPE=« MONEY » > $400 million </NUMEX>, but nothing has materialized.

FIG. 1.2 – Exemple d'annotation d'entités nommées (MUC-6).

prochent des performances humaines. Néanmoins, comme le font remarquer B. Sundheim [Sundheim, 1995] et R. Grishman [Grishman et Sundheim, 1996], ces performances sont à apprécier en tenant compte du fait que les corpus d'évaluation sont homogènes, de très bonne qualité rédactionnelle, et en nombre relativement restreint (30). La conférence MUC-7 voit se réaliser des performances du même ordre, quoique moins impressionnantes : le meilleur système n'obtient « que » 92 % pour le rappel et 95 % pour la précision. Ceci s'explique par l'ajout des expressions temporelles relatives dans la catégorie TIMEX, et par le fait que le corpus d'entraînement soit d'un autre domaine (accidents d'avions) que celui de test (lancements de satellites).

Force est donc de constater[1] que la tâche de reconnaissance des entités nommées fut un réel succès pour les conférences MUC. Menée à bien avec d'excellents résultats dès son lancement, cette tâche suscite dès lors un engouement certain, de la part et des participants, et des utilisateurs potentiels. En effet, à l'issue de MUC-6, deux systèmes sont commercialisés et d'autres intégrés dans des systèmes gouvernementaux d'analyse de textes, démontrant ainsi la possibilité de concevoir des systèmes génériques rapidement utilisables. En définitive, cette tâche répond positivement aux ambitions des dernières MUC, soucieuses, à juste titre, de définir des fonctionnalités indépendantes pour la tâche d'extraction d'information.

Au final, avec cette série de conférences américaines *Message Understanding*, ce sont dix années d'évolution de la tâche d'extraction d'information qui peuvent ainsi être appréciées. Coupant court aux systèmes de compréhension de textes, l'extraction d'information prend à la fin des années 1980 un parti plus modéré et ne s'intéresse, dans un texte, qu'à certains éléments précis et préalablement définis. Ainsi posée, cette tâche est-elle plus simple ? Plus facilement concevable certainement, mais plus simple, rien n'est moins sûr. La concentration de l'effort sur un type précis d'information n'empêche pas ce dernier d'être complexe et difficile à extraire. Preuve en est, si cette tâche gagne en définition au fur et à me-

[1] Ce constat est à la fois un jugement personnel et la reprise d'un jugement des organisateurs de la tâche (cf. R. Grishman).

sure des conférences, elle gagne également en complexité, les éléments à extraire étant de plus en plus nombreux. C'est ainsi que, compte tenu de cette évolution et de la nécessité de prendre en compte la réalité des applications, cette tâche a progressivement pris un caractère modulaire, se décomposant en plusieurs fonctionnalités autonomes. La tâche de reconnaissance des entités nommées apparaît alors, connaissant un succès immédiat, tant par ses résultats que par l'enthousiasme suscité auprès des participants nombreux. Aussi, avec la caractérisation progressive de la tâche, la réalisation de systèmes relativement performants et la promotion d'un véritable processus d'évaluation, ce cycle global de conférences constitue à coup sûr l'occasion d'un important progrès pour la tâche d'extraction d'information[1] et, au-delà, pour le traitement automatique du langage en général.

Au regard de la tâche de reconnaissance des entités nommées, cette évolution est d'autant plus manifeste que de nombreuses autres campagnes d'évaluation ont succédé à cette première série de conférences, consolidant et généralisant cette tâche nouvellement apparue.

1.2.3 Les autres conférences

Dans la droite ligne des conférences MUC, de nombreuses autres campagnes d'évaluation ont en effet poursuivi l'organisation de compétitions autour des entités nommées. Nous avons déjà évoqué les deux campagnes MET[2], organisées parallèlement à MUC-6 puis MUC-7, et dédiées à la reconnaissance des entités nommées dans d'autres langues que l'anglais [Merchant *et al.*, 1996]. Informelles et anonymes, ces deux campagnes ont motivé diverses expérimentations et favorisé la conception de systèmes indépendants de la langue, pouvant reconnaître de manière satisfaisante des entités nommées dans des corpus de nature journalistique en espagnol, chinois et japonais. Immédiatement après cette première généralisation de la tâche à d'autres langues, a été organisé au Japon le projet d'évaluation IREX[3]. L'objectif des organisateurs [Sekine et Isahara, 1999] était de réaliser quelque chose de similaire à MUC, en évaluant des systèmes d'extraction d'entités nommées sur des bases communes et en partageant données et expériences. Une quinzaine de systèmes ont participé avec des textes extraits de quotidiens nippons, obtenant pour certains des scores honorables avec des F-mesures supérieures à 0,8 [Sekine et Eriguchi, 2000]. En 2002 et 2003, CoNLL[4] a proposé une tâche de reconnaissance d'entités nommées, pour l'espagnol et le hollandais tout

[1] Même si les taux de précision et de rappel n'augmentent pas franchement entre MUC-3 et MUC-6, paradoxe relevé et expliqué par [Hirschman, 1998].

[2] Multilingual Entity Task.

[3] Information Retrieval and Extraction Exercise.

[4] Conference on Natural Language Learning.

d'abord, l'anglais et l'allemand ensuite, réunissant plus d'une dizaine de participants à chaque fois ([TjongKimSang, 2002] et [TjongKimSang et Meulder, 2003]). En France, la campagne ESTER[1], intégrée au projet EVALDA et organisée de 2002 à 2006, a proposé l'évaluation de systèmes de transcription d'émissions radiophoniques en langue française, ces transcriptions devant être enrichies par un ensemble d'informations annexes dont le marquage des entités nommées. Poursuivant encore la diffusion de cette tâche à d'autres langues, la campagne HAREM[2] proposa quant à elle une évaluation pour le portugais (portugais du Portugal, du Brésil, d'Afrique et d'Asie), réunissant près de 10 participants autour de corpus de natures diverses, issus de journaux, de courriers électroniques, de fictions, de rapports techniques ou encore de pages web [Santos *et al.*, 2006]. Enfin, il importe d'évoquer le programme ACE[3], mis en œuvre de 2000 à 2004 : prenant le relais des campagnes MUC pour l'anglais, ce programme de recherche entend poursuivre des travaux dans la même direction, à savoir détection des entités, des événements et des relations entre ces éléments, avec cependant un esprit différent, plus centré sur la mise au point de technologies que sur le traitement d'applications spécifiques[4]. Au regard de l'analyse de texte elle-même, ce programme envisage les choses différemment de MUC, faisant le choix d'une perspective plus « sémantique » que « linguistique » [Maynard *et al.*, 2005]. L'objectif dans ACE n'est effet plus d'extraire des entités nommées mais des *entités* tout court, nommées ou non : l'intérêt n'est plus pour la chaîne de caractères mais pour le *concept* de l'entité elle-même, détectable au travers de ses diverses *mentions*, ces dernières pouvant être des noms propres, des expressions nominales ou encore des pronoms. La chaîne référentielle d'une entité est donc explorée et annotée dans son intégralité ; néanmoins, il n'est pas question de reconnaître tous les types d'entités dans les textes et ces derniers sont limités aux désormais classiques PERSONNE, ORGANISATION et LIEU, légèrement remaniés toutefois, et auxquels sont ajoutés d'autres types. Présentant ainsi des modifications quant à la vision de la tâche avec une dimension plus sémantique — voire ontologique — mise en avant, le programme ACE met en œuvre un travail de recherche somme toute traditionnel autour des entités nommées, s'intéressant à l'extraction et au typage de certains types d'entités dans les textes, à l'instar des premières MUC et des autres conférences évoquées ci-avant.

L'évolution finale des conférences MUC, rappelons-le, fut celle de la valori-

[1] Evaluation des Systèmes de Transcription Enrichie d'Emissions Radiophonique.

[2] Avaliação de sistemas de Reconhecimento de Entidades Mencionadas

[3] Automatic Content Extraction.

[4] « The ACE program is a « technocentric » research effort, meaning that the emphasis is on developing core enabling technologies rather than solving the application needs that motivate the research. », [Doddington *et al.*, 2004].

sation des tâches indépendantes et des systèmes génériques. Atteignant au-delà de toute espérance des niveaux d'autonomie et de performance satisfaisants, la reconnaissance des entités nommées constitua la meilleure illustration de ce mouvement, d'ailleurs confirmé et poursuivi au travers d'autres conférences ou campagnes d'évaluation (un tableau récapitulatif de ces dernières figure en annexe A). Celles-ci se sont en effet succédées régulièrement de l'arrêt de MUC jusqu'à aujourd'hui, « démocratisant » et généralisant cette tâche à d'autres langues, à d'autres domaines et à des corpus de natures différentes. Si la tâche d'extraction d'entités nommées conserve son principe de base (repérage et typage d'éléments de types prédéfinis dans le texte), elle varie cependant quelque peu d'une conférence à l'autre, qu'il s'agisse des types d'entités à reconnaître, des occurrences à annoter ou des métriques d'évaluation utilisées. Ces particularités rendent difficile la comparaison des performances d'une campagne l'autre ; il n'empêche, cette tâche s'est de toute évidence affirmée durant cette dernière décennie, son utilisation dans de nombreuses et diverses applications marquant d'autant plus son succès.

1.3 Applications

Si la reconnaissance des entités nommées a connu un tel succès, offrant aux compétitions MUCs une indirecte postérité et donnant lieu à de nombreuses autres évaluations, c'est non seulement en raison de leur apparente facilité de traitement mais aussi et surtout en raison de l'important profit que peuvent en tirer de nombreuses applications. Il est possible de distinguer deux types d'applications, de natures différentes : la reconnaissance des entités nommées peut, d'une part, faire partie d'un composant TAL qui bénéficie alors de cette information (application que l'on pourrait qualifier d' « indirecte ») et, d'autre part, faire partie d'une chaîne de traitement avec une application directe particulière. Il importe de considérer successivement ces deux types d'applications.

1.3.1 La reconnaissance des entités nommées comme composant interne au TAL

1.3.1.1 L'analyse syntaxique

La reconnaissance des entités nommées peut constituer un module fort bénéfique pour la réalisation d'étapes intermédiaires dans une chaîne de traitement TAL. L'analyse syntaxique peut ainsi bénéficier de ce type de module, comme le montrent [Brun et Hagège, 2004] et [Osenova et Kolkovska, 2002]. C. Brun et C.

Hagège [Brun et Hagège, 2004], examinant les tenants et aboutissants de l'intégration d'un module de reconnaissance d'entités nommées au sein d'un analyseur syntaxique robuste, identifient plusieurs niveaux du processus d'analyse pouvant mettre à profit ce module. En amont tout d'abord, les étapes de prétraitement que sont la segmentation et l'étiquetage morpho-syntaxique peuvent gagner en précision et en rapidité : il peut en effet être utile de savoir que la virgule et le point dans *HyOx, Inc.* ne constituent pas des séparateurs à proprement parler puisqu'ils sont parties intégrantes d'une entité de type organisation, et que *Seat* dans *Seat and Porsche had fewer registration in July 1996* est une organisation et ne peut par conséquent pas être un verbe. Ensuite, l'analyse syntaxique proprement dite peut également éviter des erreurs, notamment pour ce qui est du traitement de la coordination ; sur la base de types similaires, les entités *Egypt* et *Jordan* peuvent être coordonnées dans :

> He will be replaced by Eliahu Ben-Alissar, a former israeli envoy to <LOC> Egypt</LOC> and <LOC>Jordan/LOC>.

En revanche, dans la phrase :

> He will be replaced by Eliahu Ben-Alissar, a former israeli envoy to <LOC> Egypt</LOC> and <ORG>Likud party</ORG> politician.

il importe de coordonner *politician* avec *envoy*, et ce grâce à l'information apportée par les entités nommées, de deux types différents (LOC et ORG) cette fois-ci. Enfin, la construction de dépendances syntaxiques (ou relations grammaticales), peut « gagner en sémantique » grâce aux entités nommées, avec par exemple la construction d'une dépendance LOCALISATION entre *met* et *Baghdad* dans la phrase *They met in Baghdad*, et ce grâce à l'information géographique connue concernant l'entité *Baghdad*. Ainsi, c'est presque à toutes les étapes de traitement d'un analyseur syntaxique qu'un module de reconnaissance d'entités nommées peut être utile, venant soutenir des processus de base de l'analyse ou permettant d'enrichir cette dernière.

1.3.1.2 La coréférence

Une autre application « interne » du TAL bénéficiant de la reconnaissance des entités nommées est la résolution de coréférence, ou le traitement des chaînes anaphoriques. Dans un texte, ou dans tout autre format de communication, les référents ou objets du monde dont il est question sont évoqués le plus souvent *via* diverses expressions référentielles. Ces dernières peuvent être des noms propres (*Lionel Jospin*), des descriptions définies complètes ou incomplètes (*l'ancien premier ministre socialiste*), ou encore de simples pronoms (*il*). Le fait de repérer

et de grouper des expressions référant à une même entité correspond à la résolution de coréférence et les systèmes automatiques réalisant une telle opération comportent très souvent un module de reconnaissance des entités nommées. En effet, il s'avère que la plupart des antécédents des pronoms personnels (80-85 % selon [Bontcheva *et al.*, 2004]) sont des noms de personnes, exprimés peu avant dans le texte. Par exemple, pour résoudre la coréférence du pronom neutre *it* en anglais, il peut être utile de discriminer parmi ses antécédents possibles ceux référant à une personne ou non, afin d'empêcher les rattachements malheureux. Considérons l'exemple suivant :

John bought a new computer. It is able to process XML.

Ici, le fait de savoir que *John* est une entité de type PERSONNE interdit le rattachement du pronom *it* à celle-ci, même si sa position sujet lui attribue l'importance d'un bon candidat. Ayant participé à la tâche de détection de relations entre entités définie par ACE, [GuoDong *et al.*, 2005] rendent compte de l'apport de diverses sources d'information pour cette tâche et montrent que celle du type d'entité nommée permet d'augmenter la F-mesure de leur système de 8.1 points. Ainsi, les entités nommées semblent constituer une source d'information non négligeable pour un système s'attachant à calculer des liens de coréférence.

1.3.1.3 La désambiguïsation lexicale

Il convient par ailleurs d'examiner un autre processus TAL pour lequel les entités nommées sont d'une participation fructueuse : la désambiguïsation lexicale. Cette tâche, connue en anglais sous le nom de *Word Sense Disambiguation*, est primordiale pour tout système de traitement automatique des langues. Par désambiguïsation lexicale on désigne l'opération consistant à déterminer le sens d'un mot en contexte. En effet, un nombre important des mots de toute langue naturelle est susceptible de recevoir plusieurs interprétations ; ce phénomène, loin d'être un handicap pour les locuteurs d'une langue, qui savent bien au contraire se l'approprier en en explorant toutes les possibilités, constitue à l'égard des langues un gage de productivité et d'expressivité. La chose est cependant nettement plus compliquée pour un système automatique : comment un ordinateur peut-il sélectionner le sens à attribuer à un mot en contexte ? De nombreux travaux portent sur ce domaine et, si la tâche est loin d'être résolue, les pistes explorées sont multiples. C'est en tant qu'information sémantique que les entités nommées peuvent prendre part à un processus de désambiguïsation lexicale. Plus exactement, elles peuvent intervenir comme « filtre » au niveau des restrictions de sélection (ou sous-catégorisation sémantique) des sens d'une unité lexicale. En écho à la sous-catégorisation syntaxique, on appelle restriction de sélection le conditionnement

sémantique qu'une unité lexicale peut imposer à ses compléments, ou arguments pour les verbes, pour un sens donné. Pour donner, si l'on peut dire, du sens à tout cela, prenons les exemples suivants :

> Il est difficile de quitter Paris le vendredi soir.
> Certains se posent la question de quitter le Parti Socialiste.

Dans les phrases ci-dessus, le verbe *quitter* apparaît avec deux sens différents : dans la première il s'agit du sens de s'éloigner d'un lieu et dans la seconde de celui de se désengager. Pour les différencier, il peut être utile de prendre en compte le type sémantique des arguments convoqués par ce verbe, d'où l'intérêt de la reconnaissance des entités nommées : la présence d'une entité de type « lieu » en objet du verbe de la première phrase et d'une entité de type « organisation » en objet du verbe de la seconde peut en effet permettre de choisir le bon sens parmi les sens possibles pour le verbe *quitter*. Cette méthode d'exploitation des entités nommées pour discriminer le sens d'une unité lexicale relativement à ses restrictions de sélection fait d'ailleurs partie du cadre méthodologique de désambiguïsation lexicale déterminé par Mark Stevenson, lequel repose sur le principe de la combinaison des sources d'information, qu'elles soient d'ordre morphologique, syntaxique, sémantique, contextuel, etc. [Stevenson, 2003, Ehrmann, 2005].

1.3.1.4 La traduction automatique

Enfin, il importe de dire un mot de la traduction automatique, tâche pour laquelle la reconnaissance des entités nommées constitue également une amorce importante. Un système de traduction automatique permet de traduire un document original en langue source en un document final en langue cible. Pour ce faire, diverses méthodes peuvent être utilisées, dont la plupart se concentrent à un moment ou à un autre sur les entités nommées. Ces dernières font l'objet de deux processus au cours d'une traduction : la translittération et la traduction. La translittération correspond à la mise en correspondance de signes d'entités nommées, sorte de traduction autonymique, entre *Londres* et *London* par exemple. La traduction des entités nommées correspond quand à elle à la mise en correspondance des signifiés, ou plutôt des référents des entités nommées. Si la translittération des entités nommées est un processus relativement facile et sans incidence sur la traduction globale du texte considéré, leur traduction peut en revanche conditionner la justesse du résultat final à l'échelle du texte. À titre d'illustration, considérons la traduction suivante, réalisée par le logiciel Systran[1] :

[1] http ://trans.voila.fr/voila

TEXTE SOURCE : Jack London was an american writer.
TEXTE CIBLE : Jack Londres était un auteur américain.

Ici, le patronyme *London* a été traduit avec le nom de la capitale *Londres*, ce qui n'est pas très adéquat. Cette mauvaise traduction aurait pu être évitée si la séquence *Jack London* avait été reconnue comme une entité de type personne. Un module de reconnaissance des entités nommées peut donc soutenir des systèmes de traduction automatique.

Analyse syntaxique, coréférence, désambiguïsation lexicale, traduction automatique, les entités nommées peuvent de fait venir soutenir divers composants internes au TAL. Leur utilité ne s'arrête cependant pas à cet aspect que nous avons qualifié d'« indirect », mais s'étend bien sûr au delà, avec de nombreuses applications directes.

1.3.2 Les applications directes

1.3.2.1 L'extraction d'information et la veille

L'extraction d'information, application historique s'il en est, a déjà été décrite auparavant (cf. section 1.2.1), point n'est besoin donc de revenir longuement sur la question. Rappelons simplement que, pour remplir des bases de données avec une collection de formulaires comportant des informations ciblées relativement à un objet informationnel prédéfini, la reconnaissance des entités nommées est indispensable, ces dernières permettant de déterminer les actants de l'information à extraire. Une autre application, proche de l'extraction d'information en cela qu'elle cherche à donner une vision rapide et synthétique d'une collection de documents, est la veille documentaire. Selon l'Atalapédie[1], la veille correspond à l'activité qui vise à « surveiller, actualiser et prédire l'évolution des connaissances sur un domaine donné » . Pour ce faire, des analystes ont besoin de prendre connaissance de documents en grande quantité et en un temps relativement limité, ces documents portant le plus souvent sur des sujets économiques ou technologiques. Cette prise de connaissance demande bien sûr à être épaulée par des outils de fouille de texte ou d'extraction d'information, et c'est ici qu'interviennent de nouveau les entités nommées. Comme le développe T. Poibeau [Poibeau, 1999], celles-ci peuvent en effet constituer un véritable « enjeu pour les systèmes de veille », permettant de repérer rapidement les personnes ou entreprises dont il est question dans un do-

[1]Basée sur le principe du « wiki », l'Atalapédie est une encyclopédie en ligne offrant des informations sur le TAL, lancée par l'Association pour le Traitement Automatique des Langues. http ://www.atala.org/AtalaPedie/index.php ?title=Accueil.

cument, et donc de déterminer la pertinence de ce dernier au regard du type de veille mis en œuvre.

1.3.2.2 La tâche de question-réponse

Au-delà de cette recherche documentaire basée pour l'essentiel sur des techniques d'indexation, d'autres modes d'accès à l'information sont apparus (dits de « troisième génération », cf. [Enjalbert et Bilhaut, 2005]), parmi lesquels les sytèmes de question-réponse. Les systèmes de question-réponse (tâche connue sous le nom de *Question-Answering* en anglais) s'inscrivent ainsi dans la continuité des systèmes de recherche d'information, à la différence que ceux-ci renvoient un ensemble de documents en réponse à une requête formulée à l'aide de mots-clés tandis que ceux-là renvoient une réponse précise ou un court extrait de document en réponse à une requête formulée à l'aide d'une question en langue naturelle. Cette tâche, mise en avant lors de la huitième édition de TREC[1] (*Text REtrieval Conference*) en 1999, s'est progressivement complexifiée et étendue à d'autres langues que l'anglais, avec notamment les campagnes NTCIR (*Evaluation of Information Access Technologies*) pour les langues asiatiques et CLEF (*Cross Language Evaluation Forum*) pour les langues européennes. À l'heure actuelle, les systèmes de question-réponse doivent pouvoir répondre à trois types de question[2] :

- des questions factuelles, telles que :
  When did Hawaii become a state ?
- des questions définitionnelles :
  What is Francis Scott Key famous for ?
- et des questions portant sur des listes :
  List musical compositions by Aaron Copland.

Sans entrer plus avant dans les détails de cette tâche, il convient de dire un mot de l'architecture classique des systèmes de question-réponse, afin de mieux évaluer le rôle des entités nommées. De tels systèmes procèdent traditionnellement en trois étapes[3] [Ayari, 2007] : analyse de la question, traitement des documents, puis extraction de la réponse. La première a pour mission d'identifier le *focus* de la question (ou l'élément important de la question) et de typer la réponse attendue (une personne, un lieu, etc.). Pour le premier exemple donné ci-dessus, le *focus* correspond à *Hawaï* et le type de la réponse attendue à une *date*. La deuxième étape a pour objet d'identifier, au sein d'une collection plus ou moins importante, des documents pertinents par rapport à la question puis, à l'intérieur

[1] Pour plus de détails sur ces conférences : http ://trec.nist.gov/

[2] Ces exemples sont repris de G. Marton, [Marton, 2003].

[3] Ou quatre, cela dépend de la complexité des traitements ; moteur de recherche et traitement des documents peuvent constituer deux modules séparés.

même de ces documents des passages comportant des éléments de réponse ; cette étape est le plus souvent réalisée à l'aide d'un moteur de recherche classique, auquel sont ajoutés des modules d'analyse sémantique. Enfin, la dernière étape est consacrée à l'analyse des extraits sélectionnés dans les textes afin d'en dégager la réponse souhaitée. Au cours de cette chaîne de traitement, les entités nommées interviennent à plusieurs reprises : leur reconnaissance est utile pour spécifier le type de la réponse attendue tout d'abord, pour repérer la réponse ensuite. [Ferret *et al.*, 2001] détaille cette utilisation à partir de l'exemple suivant :

QUESTION : How many people live in the Falklands ?
RÉPONSE : Falklands population of 2,100 is concentrated...

où il importe de caractériser la réponse attendue comme étant de type NUMBER, puis de repérer dans les textes des occurrences d'entités de ce même type. La même chose peut être réalisée pour notre premier exemple, où la réponse à chercher dans les documents est de type DATE. Bien sûr cette opération ne suffit pas à elle seule à découvrir la bonne réponse : il est également essentiel de travailler sur le focus et d'exploiter d'autres indices mais il n'empêche, la reconnaissance des entités nommées joue un rôle non négligeable au sein d'un système de question-réponse. À cet égard, l'étude du rôle et de l'importance des différents modules présents dans un système de *question-answering* par [Moldovan *et al.*, 2001], cités par [Marton, 2003], est probante, montrant qu'un système peut perdre plus des deux tiers de performance en l'absence de reconnaissance d'entités nommées (68 % plus exactement pour leur système). [Toral *et al.*, 2005] arrivent eux aussi à des conclusions similaires : faisant l'expérience de l'utilisation ou non d'un système de reconnaissance d'entités nommées en aval du moteur de recherche, ils démontrent que l'utilisation d'un tel module permet de réduire significativement le nombre de textes à considérer pour l'extraction de la réponse. Les systèmes de question-réponse constituent ainsi une des applications majeures de la reconnaissance des entités nommées.

1.3.2.3 L'anonymisation

Autre application directe des entités nommées : l'anonymisation. Ce processus, défini avec précision par [Medlock, 2006], correspond à « l'identification et la neutralisation de références confidentielles dans un document ou un ensemble de documents »[1]. Cette tâche revêt toute son importance au regard de nombreuses activités où le partage de données textuelles comportant des informations confi-

[1] Traduction de [Medlock, 2006] : *« Anonymisation is the task of identifying and neutralising sensitive references within a given document or set of documents »* .

dentielles est indispensable. Au premier rang de ces dernières, figurent celles relevant des domaines juridique et médical. Le besoin d'anonymisation peut être motivé par la nécessité d'échanger ou de travailler réellement sur des données comportant des éléments confidentiels (cours de médecine pouvant s'appuyer sur des cas réels) ou par la nécessité d'appliquer des processus de TAL pour extraire des informations ou connaissances à partir d'une base textuelle (étude des décisions de justice sur une période donnée par exemple). Quoiqu'il en soit, certains éléments doivent être neutralisés et cela peut se faire soit par effacement, soit par remplacement par la catégorie à laquelle ils appartiennent, soit encore par pseudonymisation [Medlock, 2006]. Ces éléments confidentiels, on l'aura deviné, correspondent la plupart du temps à des noms de personnes, de lieux, des dates, etc., soit des entités nommées. C'est ainsi que D. Kokkinakis et A. Thurin exploitent un module de reconnaissance d'entités nommées dans leur systèmes d'anonymisation des lettres de décharge dans les hôpitaux [Kokkinakis et Thurin, 2007], tout comme L. Plamondon qui s'intéresse aux décisions de justice [Plamondon *et al.*, 2004]. Si A. Medlock souligne que la tâche d'anonymisation ne peut se permettre d'exploiter tels quels des modules de reconnaissance des entités nommées[1], ces derniers sont tout de même indispensables à un tel processus.

Enfin, pour clore ce tour (non exhaustif) des applications directes des systèmes de reconnaissance des entités nommées, il est possible d'évoquer, mais seulement d'évoquer, les moteurs de recherche cherchant à prendre davantage en compte la composante sémantique dans leur analyse de textes. Cette application est émergeante et ne correspond pas exactement à la reconnaissance des entités nommées telle qu'elle a été définie jusqu'à maintenant ; il en sera question de manière plus précise dans le reste de l'exposé (chapitre 6 notamment).

Aussi, qu'il intervienne en tant que composant interne au TAL ou bien directement, un module de reconnaissance d'entités nommées peut de toute évidence servir de nombreuses applications. De l'analyse syntaxique aux questions-réponses en passant par la traduction, l'extraction d'information ou encore l'anonymisation, le traitement des entités nommées trouve en effet sa place, plus ou moins importante, au sein des processus mis en œuvre. L'intérêt et le bénéfice de la reconnaissance d'entités nommées ayant ainsi été détaillés, il est temps d'en considérer la mise en œuvre et d'examiner les méthodes et systèmes permettant de reconnaître ces unités dans les textes.

[1]Les éléments à neutraliser peuvent ne pas être des entités nommées, cela dépend du domaine, et certains modules peuvent ne pas fonctionner correctement sur certains types de corpus comme par exemple ceux de courriers électroniques.

1.4 Systèmes et performances

Placées sur « le devant de la scène » TAL à l'occasion des conférences MUC, les entités nommées n'ont depuis cessé de motiver d'autres projets ou campagnes d'évaluation (cf. 1.2.3), favorisant de la sorte de nombreux travaux sur des systèmes permettant leur reconnaissance. Ces derniers ont donné lieu à de multiples publications et, un aperçu des plus importants étant effectué avec précision dans [Sekine et Eriguchi, 2000, Daille et Morin, 2000, Poibeau, 2001, Poibeau, 2003] et [Friburger, 2002], nous ne procéderons ici qu'à une description rapide de leur fonctionnement. Au delà d'un recensement, nous tenterons en effet de considérer certains principes de base de tout système de reconnaissance d'entités nommées, afin d'en permettre l'appréhension et la juste appréciation. Aussi, il sera question dans un premier temps des indices manifestant la présence d'entités nommées à l'écrit ainsi que des différentes méthodes possibles pour leur exploitation ; suivra dans un second temps un rapide descriptif de quelques systèmes et, enfin, un point sur les éléments et aspects essentiels de la reconnaissance d'entités nommées.

1.4.1 Reconnaissance d'entités nommées : indices et méthodes

Comment reconnaître des entités nommées dans des textes ? La question, préalable à toute conception de système de reconnaissance de ces unités, a déjà été largement explorée. Seront ici considérés les indices et moyens de reconnaissance d'entités nommées tout d'abord, leurs différentes méthodes d'exploitation ensuite.

1.4.1.1 Comment identifier une entité nommée ?

D. McDonald distingue, dans un article abondamment cité depuis sa parution [McDonald, 1996], les désormais célèbres « preuve interne » (*internal evidence*) et « preuve externe » (*external evidence*). La première se rapporte à ce qui, à l'intérieur même d'une unité lexicale ou d'un syntagme, peut indiquer qu'il s'agit d'une entité nommée[1]. Ces indices, appelés « marqueurs » ou « mots déclencheurs » (*trigger words*), correspondent à des mots ou abréviations accompagnant régulièrement une entité nommée et permettant, dans la plupart des cas mais pas toujours, de la catégoriser. Le premier indice de cette sorte est bien sûr la majuscule, marque (typo)graphique que portent d'ordinaire les noms de personnes, de lieux ou d'organisation, soit la plupart des entités nommées. Cet indice est cependant à manipuler avec précaution, pour deux raisons principalement : d'une part, le premier mot d'une phrase comporte toujours une majuscule, par conséquent,

[1] « *Internal evidence is derived from within the sequence of words that comprise the name* », [McDonald, 1996].

même s'il s'agit d'un nom, ce n'est pas forcément une entité nommée et, d'autre part, cette marque de la majuscule n'est pas de règle dans toutes les langues (en allemand les noms communs prennent aussi une majuscule). La majuscule manifeste donc la présence potentielle d'une entité nommée mais ne permet pas de la catégoriser. Autres indices, cette fois-ci plus certains : les prénoms et les indicateurs générationnels pour les noms de personnes, des mots ou affixes de type classifiant pour les noms d'organisation et de lieux, ou encore des sigles ou des esperluettes, pour les noms d'organisation seulement. Ces marqueurs, tous catégorisant, sont surlignés dans les exemples suivants :

Lionel *Jospin*
L. *Jospin*
Benoit ***XIII***
la ***Banque*** *Populaire, Crédit Agricole* ***SA***
Microsoft ***Inc.***
*l'****avenue*** *des Champs Elysées*
le ***Mont*** *Granier*

Outre ces indices situés « à l'intérieur » des entités nommées, le contexte d'apparition de ces dernières peut également constituer un moyen de les reconnaître : il s'agit là de la preuve externe[1]. En effet, tout discours, ou autre format de communication, lorsqu'il réfère à une personne, un lieu ou une autre entité nommée le fait *via* son nom, lui adjoignant aussi le plus souvent des informations supplémentaires afin d'en indiquer les propriétés spécifiques. C'est ainsi qu'un nom de personne est souvent accompagné (surtout en première mention) d'un titre ou d'un grade, et un nom d'organisation d'un mot-clé de type classifiant :

Monsieur *Jospin,* ***Mme*** *Denise*
Général *Leclerc*
*l'****entraîneur*** *Aimé Jacquet*
le ***groupe*** *Sanofi-Aventis*
the Coca-Cola ***company***

Ces indices, internes et externes, constituent des éléments importants à considérer pour un système de reconnaissance d'entités nommées. Les premiers, exploités par la plupart des systèmes, peuvent toutefois entrer en conflit avec les seconds ; c'est la preuve externe qui l'emporte dans ce cas, le type d'une entité nommée étant contraint au final par son contexte d'apparition (conflit fréquent entre les types PERSONNE et ORGANISATION, les occurrences de ce dernier portant

[1] *« External evidence is the classificatory criteria provided by the context in which a name appears »*, [McDonald, 1996].

couramment soit le nom de leur fondateur, soit le nom d'une autre personne).

Preuve interne et preuve externe résument à peu près ce qui, dans la matière textuelle, peut aider un système de reconnaissance d'entités nommées. Ceci n'est cependant pas suffisant, et un autre moyen de compléter ces informations pour un système est le recours à des lexiques. Le terme de lexique en TAL renvoie à la notion (lexicographique) de recueil de mots et non à celle (linguistique) de l'ensemble des mots d'une langue. Un lexique, ou base de connaissance lexicale, a pour objet de décrire des mots dans leurs différents sens, leurs relations et leurs emplois et peut prendre différentes formes suivant l'organisation de cette description, dictionnaire, thésaurus ou terminologie (voir [Habert *et al.*, 1997] pour plus de détails). Pour ce qui est de la reconnaissance d'entités nommées, la notion de lexique renvoie à son interprétation la plus simple, à savoir une liste de mots auxquels sont associées des catégories sémantiques indiquant s'il s'agit d'une personne, d'un lieu ou autre. L'utilisation de lexiques a été initiée dès MUC-6 et, bien que controversée (cf. 1.4.3 ci-après), demeure très répandue à l'heure actuelle.

Information contextuelle et information lexicale constituent les deux sources traditionnelles de connaissances exploitées par les systèmes d'annotation d'entités nommées. La plupart des travaux se sont concentrés sur ces informations, à juste titre au demeurant, ces dernières s'avérant très fiables (cf. performances obtenues lors des campagnes d'évaluation). De nouvelles tentatives se font jour cependant pour trouver d'autres indices et il importe à cet égard de considérer les dernières expériences de H. Shinnou et S. Sekine. Cherchant à reconnaître des entités nommées rares pour lesquelles il est difficile d'obtenir les connaissances nécessaires à leur identification, [Shinnou et Sekine, 2004] se proposent d'exploiter la synchronicité d'apparition de mots dans des corpus dits « comparables ». Leur hypothèse est la suivante : étant donnés deux corpus de textes journalistiques alignés sur la base de leur date de parution (articles alignés au jour le jour donc), un mot pour lequel il est possible d'observer un « pic » de fréquence similaire dans les deux corpus a de fortes chances d'être une entité nommée. En effet, contrairement aux autres mots du lexique, une entité nommée conserve une forme identique d'une occurrence à l'autre, et donc d'un article à l'autre, ne pouvant que difficilement faire l'objet de paraphrase. Utilisant deux corpus journalistiques de 1995, *Los Angeles Times* et *Reuters*, les auteurs ont observé la distribution « temporelle » de deux mots : « killed » et « yigal », soit un verbe et le nom de l'assassin du premier ministre israëlien Yitzhak Rabin, décédé le 7 novembre 1995. Le premier apparaît de manière régulière dans de nombreux articles dans les deux corpus tout au long de l'année, tandis que le second voit sa distribution augmenter fortement au même moment, mouvement dont on peut déduire qu'il s'agit d'une entité nommée. H. Shinnou et S. Sekine ont mené à bien d'autres

expériences similaires, raffinant leurs calculs par la prise en compte de divers paramètres (variation du nombre d'articles publiés par jour, délai de publication différents pour un même événement, etc.) et validant leur hypothèse de départ. Partant, la date d'apparition d'une entité nommée peut constituer un bon indice, d'ordre temporel donc, pour son repérage. Cette démarche comporte néanmoins quelques biais : les entités nommées reconnues ne sont pas nombreuses, et surtout ne sont pas catégorisées. Les auteurs insistent à cet endroit sur la complémentarité de leur approche avec les moyens traditionnels de reconnaissance de ces unités. Ainsi, si les preuves internes et externes, tout comme les lexiques, constituent d'excellents indices et moyens de reconnaître des entités nommées dans un texte et sont abondamment utilisés, des travaux s'attèlent encore, avec succès, à la découvertes d'indices supplémentaires.

Les diverses sources de connaissances (contenues dans le texte à analyser ou ailleurs) nécessaires à prendre en compte par un système de reconnaissance d'entités nommées ayant été décrites, il convient d'examiner les différentes manières de les acquérir et de les exploiter.

1.4.1.2 Comment annoter une entité nommée ?

Pour la plupart des processus de TAL, on distingue traditionnellement deux grandes approches : l'approche dite linguistique ou symbolique, et l'approche dite statistique ou à base d'apprentissage. Pour réaliser un système automatique d'analyse linguistique, que cette analyse soit d'ordre morphologique, syntaxique, sémantique, voire pragmatique, il importe de prendre en compte des informations, de les modéliser et de les manipuler *via* un formalisme adéquat quant à l'analyse escomptée. Ce qui distingue les approches citées ci-avant, ce n'est pas tant la nature des informations prises en compte que leur acquisition et leur manipulation.

La première repose sur l'intuition humaine, avec la construction manuelle des modèles d'analyse, sous la forme de règles contextuelles le plus souvent. Ces règles prennent la forme de patrons d'extraction, c'est-à-dire de descriptions d'enchaînements possibles de syntagmes nominaux ou verbaux, attendu qu'ils expriment l'information à repérer. Ces patrons exploitent généralement des informations d'ordre morpho-syntaxique, ainsi que celles contenues dans des ressources (lexiques ou dictionnaires). Pour ce qui est des entités nommées, ce type d'approche linguistique fut largement répandu, voire majoritaire durant les années 1990, au temps des premières conférences MUC. Un système de reconnaissance d'entités nommées basé sur une telle méthode comporte par exemple les règles suivantes (exprimées ici « verbalement ») : si un prénom connu (connaissance issue d'un lexique) précède un mot inconnu commençant par une majuscule, alors le syntagme peut être

étiqueté comme un nom de personne ; ou bien : si un mot inconnu est suivi du mot (ou de la forme) *Inc.*, alors il s'agit d'un nom d'organisation. Les choses ne sont bien sûr pas si simples, il faut savoir par exemple attribuer les bonnes frontières aux entités nommées, mais l'essentiel de cette approche est là.

L'autre type d'approche a pour principe de base la mise au point *automatique* de modèles d'analyse à partir de volumes importants de données. Ces méthodes sont dites statistiques ou à base d'apprentissage car elles *apprennent*, à partir de corpus annotés, des modèles d'analyse de textes, ces derniers pouvant prendre différentes formes, arbres de décision, ensembles de règles logiques, modèles probabilistes ou encore chaînes de Markov cachées. Au regard de la reconnaissance d'entités nommées, un système « observant » plusieurs fois la présence de l'abréviation *Mme* devant un mot annoté comme nom de personne dans le corpus d'apprentissage pourra facilement en déduire un modèle d'analyse. Ces systèmes à base d'apprentissage se sont considérablement multipliés ces dernières années.

Les avantages et inconvénients respectifs de ces deux types d'approches sont connus : les premiers reprochent aux seconds, entre autres, l'indispensable disponibilité de corpus annotés, et les seconds critiquent les premiers pour leur temps de développement ainsi que leur coût. Il est vrai qu'un travail de plusieurs mois d'un linguiste-informaticien est nécessaire pour l'écriture de règles, mais l'inverse est vrai également pour l'annotation de corpus qui peut être tout aussi longue, même si cela peut se faire par des gens moins experts. Hormis ces querelles de conception, l'intérêt se situe véritablement dans ce que chaque type de système est capable de faire et comment il peut fonctionner : si un concepteur de règles ne peut bien sûr pas penser à toutes les exceptions, il peut en revanche prévoir des patrons plus ou moins complexes pour le captage d'éléments difficiles, ce qu'un système probabiliste ne peut faire. La précision est ainsi d'ordinaire plus importante pour les systèmes symboliques tandis que les systèmes à base d'apprentissage présentent l'avantage d'être plus flexibles quant à leur adaptation à une tâche similaire mais portant sur un autre domaine et d'être plus robustes sur des corpus difficiles (ou bruités). Cette partition entre bienfaits et écueils de telle ou telle approche se reproduit bien sûr pour les systèmes de reconnaissance d'entités nommées. A. Borthwick , proposant une méthode de reconnaissance d'entités nommées à partir de calculs d'entropie maximale, ne manque pas de remarquer que, lors de la compétition MUC-7, le second système (IsoQuest, construit à partir de règles) est en de nombreux points similaire au dernier système symbolique (FACILE), à l'exception du temps de développement annoncé par les concepteurs respectifs [Borthwick, 1999]. L'auteur pointe là le coût de développement de tels systèmes ; il poursuit par ailleurs en soulignant leur difficile adaptation à d'autres domaines ou d'autres langues, les règles devant être, selon lui, totalement ré-

écrites. Inversement, [Poibeau, 2003] attire l'attention sur le volume de données annotées nécessaires pour entraîner un système, remarquant que celui de BBN [Miller *et al.*, 1998] exige un corpus annoté de 30 000 mots pour obtenir 0,81 de F-mesure, mais de 1,2 millions de mots pour atteindre 0,91. D'autres observations similaires l'amènent à la conclusion que le coût de l'écriture de règles n'est guère plus élevé que celui de l'annotation de corpus.

Enfin, au-delà de ces deux types d'approches et des désaccords de leurs partisans respectifs, il existe une troisième voie consistant à coupler l'approche symbolique et l'approche statistique en une approche alors qualifiée de *mixte* ou d'*hybride*. Cette dernière, rendue possible grâce à la maturité acquise par les deux autres, est sans doute la plus prometteuse.

Ayant ainsi présenté les divers types de processus utilisables pour la reconnaissance d'entités nommées[1], il est temps de les illustrer par l'analyse de quelques systèmes.

1.4.2 Un échantillon de systèmes

N. Friburger détaille abondamment un grand nombre de systèmes de reconnaissance d'entités nommées, qu'ils soient symboliques, à base d'apprentissage ou mixtes [Friburger, 2002]. Par conséquent, nous ne présenterons ici qu'un système par approche, renvoyant pour le reste au travail cité ci-avant [Sekine et Eriguchi, 2000, Daille et Morin, 2000].

1.4.2.1 Un exemple de système symbolique : LaSIE

Développé à l'Université de Sheffield, le système LaSIE[2] (voir Gaizauskas *et al.*, [Gaizauskas *et al.*, 1995]) fut initialement conçu dans le cadre d'un projet de recherche autour de l'extraction d'information ou, plus généralement, autour de traitement du langage naturel. Ses concepteurs, participant à la compétition

[1]Cette présentation fait état d'une opposition entre des approches symboliques et des approches statistiques, opposition pouvant paraître quelque peu réductrice. En effet, si ce classement binaire permet une vision simple et rapide des types de traitements mis en œuvre en TAL, il n'est peut-être pas le plus approprié et met seulement en valeur l'utilisation ou non du quantitatif. Si l'on considère les choses d'un autre oeil, il est possible de dire que tout système de traitement automatique de données linguistiques nécessite des connaissances et qu'il doit, premièrement, les acquérir et, deuxièmement, les utiliser. Ce sont pour ces deux opérations d'acquisition et d'utilisation de connaissances que divers processus peuvent être mis en œuvre, d'orientation symbolique ou statistique. Acquérir, utiliser, symbolique, statistique, nous avons donc deux opérations et deux processus, ce qui fait au final quatre combinaisons possibles (dont seulement trois sont réellement mises en œuvre). Les différences observables entre ces différentes combinaisons (et qui permettent donc de choisir entre telle ou telle combinaison) ont trait à la granularité et l'échelle des traitements, le tout ayant des conséquences sur la fiabilité des systèmes.

[2]Large Scale Information Extraction.

MUC-6, firent le choix d'un système intégré, réunissant dans une même architecture plusieurs modules capables de répondre aux différentes tâches proposées à l'évaluation. C'est ainsi que, déclinées en plusieurs sous-traitements, trois étapes se succèdent : analyse lexicale, analyse puis interprétation sémantique et enfin interprétation du discours. L'entrée de ce processus est un texte (dont les paragraphes sont marqués en SGML[1]), la sortie une représentation sémantique de ce dernier, permettant d'offrir des réponses à chacune des tâches MUC (entités nommées, coréférence, « template element » et « scenario template »).

La phase de traitement lexical comprend les opérations suivantes : segmentation en mots (transformation du flux de caractères en suite d'unités ou *tokenization*), étiquetage des parties du discours (à l'aide de la méthode d'E. Brill [Brill, 1995]), analyse morphologique (ici lemmatisation) et étiquetage d'entités nommées sur la base de lexiques de noms propres et de lexiques spécialisés comprenant des mots déclencheurs. Durant cette phase, correspondant en quelque sorte à l'acquisition des connaissances nécessaires concernant les éléments de base d'un texte, le seul processus faisant intervenir un apprentissage automatique est l'étiquetage morpho-syntaxique (l'étiqueteur d'E. Brill est une méthode à base de règles avec apprentissage, voir [Paroubek et Rajman, 2000] pour plus de détails) ; l'étiquetage d'entités nommées est réalisé *via* une simple interrogation de lexique (ou *look-up*).

La deuxième étape s'attache pour sa part à la mise en œuvre de ces connaissances, opérant une analyse sémantique par le biais de l'utilisation de grammaires hors contexte : grammaire pour les entités nommées tout d'abord, pour l'analyse de phrases ensuite. Ces dernières sont écrites manuellement en Prolog et exploitent les informations recueillies lors de la phase initiale. La grammaire dédiée aux entités nommées comporte un peu plus de 200 règles au total, dont presque la moitié pour les noms d'organisations. Ci-après un exemple de règle de reconnaissance d'entité nommée utilisée dans LaSIE :

```
ORGAN_NP -> NAMES_NP '&' NAMES_NP
```

Cette règle (cf. [Gaizauskas *et al.*, 1995]) indique qu'un nom non encore classifié ou un nom propre ambigu (`NAMES_NP`, indication recueillie et attachée au token lors de la phase de traitement lexical), suivi d'une esperluette, puis de nouveau suivi d'un `NAMES_NP` est un nom d'organisation (`ORGAN_NP`). Des règles de facture similaire existent également pour l'analyse des phrases, à partir de laquelle est ensuite produite une représentation sémantique.

Enfin, la troisième phase de traitement se livre à une analyse du discours,

[1] *Standard Generalized Markup Language*. SGML est un langage générique de balisage des informations.

transformant la représentation sémantique générée par la phase d'analyse en une représentation d'instances, auxquelles sont associées des classes sémantiques (les auteurs parlent de « *ontological classes* ») ainsi que des propriétés. C'est à partir de cette représentation que la résolution de coréférence est réalisée. Interviennent également à cet endroit diverses heuristiques permettant de compléter l'annotation des entités nommées : si, par exemple, un nom d'organisation est reconnu en tant que tel puis, dans la suite du texte, est repérée une forme abrégée de ce même nom ou une variante, alors cette dernière se voit attribuer le même type que sa forme parente ou coréférente (propagation). Les connaissances utilisées lors de cette étape sont construites manuellement (objets et attributs de l'ontologie peu nombreux), et les mécanismes d'exploitation de ces dernières ne font nullement intervenir de données quantitatives (utilisation de mécanismes d'héritage). On le voit, cette dernière passe du traitement permet de « revenir en arrière » et d'enrichir des analyses précédentes. Cette possibilité correspond à une réelle visée des concepteurs du système, cherchant à exploiter des informations linguistiques de tous niveaux pour mener à bien leurs analyses.

Au final, LaSIE est un système intégré d'analyse de textes entièrement fondé sur une approche linguistique ; les résultats obtenus lors de MUC-6 pour la tâche de reconnaissance des entités nommées furent de 0,91 de F-mesure, et pour MUC-7 de 0,85 (performance réalisée par LaSIE-II, pour plus de précision, se reporter à [Humphreys *et al.*, 1998]). À la suite de cette description d'un système de reconnaissance d'entités nommées de type « symbolique », considérons à présent un exemple de système à base d'apprentissage.

1.4.2.2 Un exemple de système à base d'apprentissage : MENE

Le système MENE[1] a été présenté pour la première fois à MUC-7 (1998) par l'Université de New-York. Conçu par [Borthwick *et al.*, 1998], MENE opère la reconnaissance d'entités nommées dans les textes en exploitant le principe de l'entropie maximale. À l'origine grandeur thermodynamique mesurant le degré de désordre de la matière, le principe de l'entropie a été adaptée à la théorie de l'information par C. Shannon. Ce dernier correspond alors à la mesure de la quantité d'incertitude liée à la distribution d'un événement aléatoire. Au regard du problème de la reconnaissance d'entités nommées, l'entropie maximale peut être vue comme le degré de certitude de l'information dérivable d'un corpus d'apprentissage relativement à une unité lexicale donnée, unité étant reconnue comme entité nommée et dont on cherche à établir le type. La mise en œuvre du calcul d'entropie maximale nécessite, comme tout système probabiliste, la détermina-

[1] Maximum Entropy Named Entity.

tion d'un certain nombre de traits (ou éléments d'information attachés à l'objet étudié) sur lesquels travailler. A. Borthwick *et al.*, décrivant le fonctionnement de leur système, font état de l'utilisation des traits suivants :

- traits binaires : traits dont la valeur est soit positive soit négative, ils correspondent grossièrement à ceux utilisés par le système Nimble de BBN [Bikel *et al.*, 1997] (capitalisation, présence de caractères numériques, etc.).
- traits lexicaux : traits issus du contexte lexical de l'unité considérée. Soit U_n l'unité à catégoriser, on peut alors lui associer le trait U_{n-1} correspondant par exemple à *Mrs.* ou le trait U_{n+1} correspondant par exemple à *to* ; il est ensuite possible d'observer des régularités pour ces traits (le premier indique plutôt que l'unité est de type « personne », et le second qu'elle est de type « lieu »).
- traits « textuels » : traits véhiculant l'information de la structure du texte, *i.e.* localisation de l'unité dans le titre, le résumé, le corps du document, etc.
- traits issus de dictionnaires : traits signalant l'appartenance d'une unité à l'un de ces cinq états possibles, « start, continue, end, unique, other », sur la base d'une récupération des informations contenues dans des dictionnaires de noms propres. Ainsi, à partir de la connaissance de l'unité « British Airways » contenue dans le dictionnaire, si l'expression « on British Airways Flight 962 » est rencontrée, alors elle se verra attribuer les traits « other, start, end, other, other ». MENE exploite de la sorte des dictionnaires de noms de personnes, d'organisations, d'universités, de régions, ainsi que d'affixes de noms d'entreprises. Les auteurs insistent sur l'importance de ce type de traits pour leur système.
- traits externes : traits issus d'autres systèmes de reconnaissance d'entités nommées, soit l'indication de types sur les unités.

Une fois collecté l'ensemble de ces traits, MENE procède au calcul d'entropie maximale pour l'annotation d'entités dans de nouveaux textes[1]. Lors de la compétition MUC-7, ce système afficha 0,92 de F-mesure sur le corpus d'entraînement et 0,84 sur le corpus de test. Rappelons, à l'instar des concepteurs de ce système, que le changement de domaine entre le corpus d'entraînement (sur les accidents d'avion) et le corpus de test (sur les lancements de satellites), ne fut pas annoncé par les organisateurs de la compétition et prit par surprise la quasi totalité des systèmes à base d'apprentissage participant. Nous avons ici l'illustration de l'un des points faibles de ce type de systèmes, nécessitant de nouveaux corpus d'apprentissage pour pouvoir fonctionner sur de nouveaux domaines, *a contrario* des systèmes symboliques qui, si leur règles sont bien structurées et s'ils possèdent

[1]Sur le concept et le calcul d'entropie maximale, voir l'introduction de [Ratnaparkhi, 1997].

une bonne ergonomie (cf. [Poibeau, 2003]), sont facilement adaptables. Quoi qu'il en soit, les résultats obtenus par MENE sont fort bons. Leur niveau varie bien sûr en fonction de la quantité de données disponibles pour l'apprentissage : avec 20 documents, MENE propose 0,8 de F-mesure, mais cette dernière monte à 0,92 avec 425 documents. Toutefois, la meilleure amélioration du système provient de sa combinaison avec d'autres systèmes, de type symbolique, tels que IsoQuest [Krupka et Hausman, 1998], Manitoba [Lin, 1998] ou Proteus [Grishman, 1995] : utilisant les sorties fournies par ces trois systèmes, MENE affiche alors, sur le corpus d'entraînement de MUC-7, 0,97 de F-mesure. Ainsi, A. Borthwick *et al.* plaident à la fin de leur article pour l'utilisation combinée des deux approches, symbolique et à base d'apprentissage, l'une pouvant pallier les défauts ou insuffisances de l'autre. C'est ce postulat que D. Lin s'applique à mettre en œuvre.

1.4.2.3 Un exemple de système mixte : travail de D. Lin.

Pour la compétition MUC-7, D. Lin ([Lin, 1998]) a présenté une « extension » du système symbolique réalisé par l'Université de Manitoba lors de MUC-6, extension consistant en l'addition au système initial d'un mécanisme d'enrichissement et de génération automatique de règles de reconnaissance d'entités nommées sur la base de données quantitatives. En un mot, le cœur du travail repose sur la construction d'une base de données de collocations. Le processus d'annotation des entités nommées opère en deux passes : la première construit, à partir d'un corpus annoté par le système initial et analysé syntaxiquement, une base de données de collocations de mots à partir de laquelle sont ensuite générées de nouvelles règles, et la seconde exploite le fruit de cette acquisition pour annoter, cette fois-ci définitivement, des entités nommées dans un texte.

La construction de la base de collocations consiste plus concrètement en la collection, dans un corpus donné, de l'ensemble des mots reliés deux à deux sur la base de leur relation grammaticale. Ces relations sont identifiés grâce à un parseur (D. Lin utilise MINIPAR), et sont stockées avec leur fréquences d'apparition. Une collocation (baptisée « dependency triple » par D. Lin) prend alors la forme suivante : (`word, relation, relative`), où le champ `word` correspond à un mot du corpus, celui de `relative` à un modifieur de ce mot, et celui de `relation` au type de la relation grammaticale reliant les instances des deux champs précédents, avec l'indication de leurs parties du discours. Chaque mot du corpus peut ainsi être stocké dans la base, en fonction de sa partie du discours, du mot auquel il est relié, et du type de relation opérant ce lien. Afin de se faire une idée plus précise du contenu de cette base, examinons quelques informations relatives à l'entrée `review` présentée par D. Lin :

`[review, V:comp1:N, acquisition = 3]` signifie que le verbe « to review » a eu trois fois le nom « acquisition » en relation objet dans le corpus analysé.

`[review, N:nn:N, admission = 2]` signifie que le nom « review » a eu deux fois le nom « admission » en relation modifieur de nom dans le corpus analysé.

Ce mot apparaît également dans d'autres collocations, au sein d'autres relations grammaticales ; la base de données contient au total 22 millions de mots. Au final, même si D. Lin ne fait pas référence à Z. Harris, il semble bien s'agir ici d'analyse distributionnelle.

Une fois construite, cette base de données de collocations est exploitée selon deux mécanismes. Le premier permet de générer automatiquement des règles d'annotation d'entités nommées et le second d'annoter des entités inconnues. Rappelons avant tout que le système initial utilisé est purement linguistique : il exploite des patrons à base d'automates à états finis, tirant profit de ressources lexicales tout comme des formes de surface du texte à traiter. D. Lin part du principe que le contexte d'apparition d'une entité nommée constitue un bon indice pour sa catégorisation (c'est la preuve externe de MacDonald). Ces contextes, accompagnés de leurs fréquences, sont contenus dans la base de données ; il suffit dès lors d'observer des régularités pour en déduire des schémas d'annotation. L'auteur explique par exemple que, sur 33 occurrences de noms propres apparaissant en tant que pré-modifieurs du syntagme « managing director », 26 sont classées comme relevant de la catégorie organisation. La déduction est relativement simple, et la règle permettant d'annoter les 7 entités restantes peut être produite. Ce mécanisme peut être mis en place pour 3600 contextes environ, pour lesquels la fréquence d'apparition d'un nom propre permet de déduire une règle. Cependant, il est des contextes où la déduction n'est pas si évidente et pour lesquels aucune classification sûre ne peut être proposée. Intervient alors le deuxième mécanisme, cherchant donc pour sa part à classifier des entités inconnues pour lesquelles le contexte d'apparition n'est pas suffisamment discriminant. Le système utilise à cet endroit un classificateur bayésien naïf : cet outil est basé, comme son nom l'indique, sur le théorème de Bayes et permet de calculer des probabilités conditionnelles. L'objectif est donc de prédire la catégorie d'une entité nommée à partir de ses caractéristiques. Ces dernières correspondent ici aux contextes d'apparition d'une entité donnée. Prenons l'exemple suivant (toujours issu de [Lin, 1998]) : étant donné le mot « Xichang », apparaissant à de nombreuses reprises dans le corpus mais inconnu du lexique, il suffit de relever tous ses contextes d'apparition ainsi que leurs fréquences, en autant de traits à partir desquels le classificateur bayésien peut prédire une catégorie. Par ce mécanisme, de nombreux mots inconnus peuvent être ajoutés au lexique. L'inconvénient de ce mécanisme est l'absence de contrôle : un mot peut être mal catégorisé et ajouté

au lexique, provoquant par la suite de nombreuses erreurs. Malgré cela, le système ainsi mis en œuvre par D. Lin reconnaît les entités nommées de MUC-7 avec un taux de F-mesure de 0,86, performance saluée lors de la conférence.

Nous venons de décrire trois systèmes, chacun d'eux participant d'une des approches possibles pour la reconnaissance d'entités nommées. Se reposant sur les mêmes types d'indices mais les exploitant de manières différentes, ces systèmes parviennent tous à passer les 0,85 de F-mesure lors des compétitions MUC. Au-delà du constat d'excellentes performances, est-ce à dire que toutes les approches se valent ? Que retenir de ces différentes méthodes de reconnaissance d'entités nommées ? Le paragraphe suivant tentera de donner des éléments de réponse à ces questions.

1.4.3 Points de controverse et lignes de force

Cela a déjà été dit plus haut et nous venons de l'illustrer en partie : la reconnaissance des entités nommées est une tâche qui, telle que définie et réalisée lors des conférences MUC-6 et MUC-7, recueille de très bonnes performances. En effet, les trois systèmes décrits ci-dessus franchissent tous les 0,85 de F-mesure, comme la plupart d'ailleurs des autres systèmes présentés à la compétition lors de MUC-6 et MUC-7 (cf. [Sundheim, 1995]). À mieux observer ces résultats cependant, une chose peut paraître déconcertante, à savoir le fait qu'aucune approche ou système ne se démarque résolument des autres. Bien plus, systèmes linguistiques, à base d'apprentissage ou bien encore mixtes se partagent régulièrement le podium, interdisant tout arbitrage absolu en faveur de telle ou telle méthode. Les organisateurs de la campagne IREX, homologue japonaise de MUC, S. Sekine et Y. Eriguchi parviennent aux mêmes conclusions : « *Il est intéressant de noter que les trois systèmes les plus performants appartiennent chacun à une catégorie différente. (...) Par conséquent il n'est pas possible de déterminer quel type de système est le meilleur.* »[1] [Sekine et Eriguchi, 2000]. Si l'ensemble des systèmes font usage des mêmes types d'indices et exploitent les mêmes sources de connaissance (preuve interne, preuve externe et lexique), les manières de les exploiter diffèrent cependant fortement. Au-delà de ces performances similaires, n'est-il pas possible d'examiner plus en détails l'apport de tel ou tel mécanisme et de dégager quelques « lignes de force » de la reconnaissance d'entités nommées ? Deux travaux analysant le fonctionnement de systèmes de reconnaissance de ces unités nous semblent

[1] Dans [Sekine et Eriguchi, 2000] : *« It is interesting to see that the top three systems came from each category ; the best system was a hand created pattern-based system, the second was an automatically created pattern-based system and the third system was a fully automatic system. So we believe we can not conclude which type is superior to the others ».* Cité également par T. Poibeau ([Poibeau, 2003]).

instructifs à cet égard. Le premier a suscité ce que nous appelons, peut-être le terme est-il trop fort, la « controverse des lexiques » : il s'agit de l'article « Named Entity Recognition without gazetteers » de A. Mikheev, M. Moens et C. Grover [Mikheev *et al.*, 1999] ; le second est l'« évaluation transparente » d'un système de reconnaissance d'entités nommées réalisée par T. Poibeau.

1.4.3.1 La controverse des lexiques

L'utilisation de lexiques pour la reconnaissance d'entités nommées est fortement répandue, la plupart des systèmes ayant recours à un moment ou à un autre de leur processus à cette source de connaissance. Il s'agit en effet d'un mécanisme sûr, efficace et rapide, fonctionnant par une simple confrontation de la chaîne de surface à analyser avec une liste d'unités lexicales préalablement catégorisées. Si l'interrogation d'un lexique se fait le plus souvent conjointement à d'autres mécanismes, ce processus est de toute évidence facile à mettre en œuvre, séduisant en ce point les concepteurs d'applications commerciales (remarque de [Friburger, 2002]). Leur rôle ainsi que leur utilisation peuvent néanmoins être discutés, et ce sur plusieurs points. Leurs premiers détracteurs sont les partisans des méthodes à base d'apprentissage, incriminant leur coût de construction et leur opposant la possibilité d'acquérir automatiquement des listes de noms sur corpus annoté. Cette acquisition a cependant elle aussi un coût, celui de l'annotation de corpus. D'autres encore mettent en garde sur ce qu'implique, informatiquement parlant, l'exploitation d'une ressource de cette nature ; N. Wacholder *et al.* [Wacholder *et al.*, 1997] refusent ainsi de recourir à des lexiques si leur « chargement » est trop coûteux et qu'aucune méthode rapide d'interrogation n'est implémentable[1].

Pour pallier le premier écueil dénoncé, le coût de construction manuelle de lexique et le coût d'annotation de corpus, une solution peut être d'acquérir automatiquement de telles listes sans corpus annoté. Comment ? La méthode est décrite par M. Collins et Y. Singer dans [Collins et Singer, 1999]. Ces derniers proposent de se débarrasser du « fardeau » de l'annotation préalable en utilisant quelques « points d'amorce » ou *seed rules* à partir desquels le système peut ensuite acquérir d'autres amorces, et ainsi apprendre récursivement des listes de noms. M. Collins et Y. Singer, à partir de 7 amorces[2] et d'un corpus d'appren-

[1] Dans [Wacholder *et al.*, 1997] : « *A reliable database provides both accuracy and efficiency, if fast look-up methods are incorporated.* »

[2] Les *seed rules* du système de [Collins et Singer, 1999] sont les suivantes :
- *New York, California* et *U.S.* sont des noms de lieux.
- un nom contenant *Mr.* est un nom de personne.
- *I.B.M.* et *Microsoft* sont des noms d'organisations.

tissage de 90 000 exemples non annotés, parviennent ainsi à annoter des entités nommées avec 91 % de précision. Le facteur principal de réussite de cette méthode est la régularité et la redondance du corpus d'apprentissage. À l'instar de T. Poibeau, saluant ce travail, on peut néanmoins se demander quelles seraient les performances avec un corpus d'apprentissage moins homogène et surtout s'interroger sur les possibilités de correction des connaissances acquises (aspect interactif). Corpus annoté ou corpus homogène, la contrainte semble ainsi être du même ordre et, si la solution proposée par M. Collins et Y. Singer n'est bien sûr pas à négliger, il importe d'en prévoir l'utilisation en combinaison avec des ressources minimales. Au final, parmi ces différentes possibilités d'acquisition de lexiques, constitués à la main, acquis sur corpus annotés ou non annotés, l'essentiel étant de privilégier une construction peu coûteuse mais apportant une bonne précision ainsi qu'une couverture suffisante, la meilleure solution paraît être d'utiliser au départ une liste « manuelle » des noms les plus courants (de nombreuses listes sont disponibles sur Internet), puis de la compléter par un processus d'apprentissage sur corpus non annoté (en prenant soin de choisir ce dernier en fonction de l'application visée).

Examinons maintenant la deuxième critique énoncée ci-avant, à savoir le coût d'utilisation de tels lexiques. Implicitement, il est question ici de l'utilité de ces listes (mise en balance avec le temps de traitement) et de leur taille. L'article de A. Mikheev *et al.*, au titre somme toute percutant, propose une recherche en ce sens. Leur propos vise en effet à répondre à trois questions initiales : Au sein d'un processus de reconnaissance d'entités nommées, quelle est l'importance des lexiques ? Quelle doit-être leur taille ? et Quels sont les critères de construction de lexiques ? Afin de répondre à ces questions, les auteurs ont conduit diverses expériences, dont le principe consiste à faire varier la taille et la complétude des lexiques utilisés par un système d'annotation d'entités nommées. Quatre cas de figures ont été testés. Le premier fait usage des lexiques complets utilisés par les auteurs lors de MUC-7, soit près de 5 000 noms de pays, 30 000 d'organisations et 10 000 de personnes, lexiques constitués pour une partie automatiquement à partir des corpus d'entraînement, et pour l'autre à partir de listes disponibles sur internet. Ce mode d'utilisation de leur système permet une bonne annotation des entités nommées, chaque type considéré (ici personne, lieu et organisation) ayant une précision et un rappel supérieurs à 90 % (cf. résultats dans le tableau 1.1). La deuxième expérience fut de faire tourner le système sans lexique aucun : pour les noms de personnes, précision et rappel sont encore supérieurs à 90 %, aux alentours de 85 % pour les noms d'organisation et totalement catastrophiques pour les noms de lieux. Ces résultats indiquent que les lexiques ne sont pas d'une grande utilité pour la reconnaissance des noms de personne et d'organisation, ces

	Full gazetteer		Ltd gazetteer		Some locations		No gazetteers	
	recall	prec	recall	prec	recall	prec	recall	prec
organisation	90	93	87	90	87	89	86	85
person	96	98	92	97	90	97	90	95
location	95	94	91	92	85	90	46	59

TAB. 1.1 – Résultats obtenus par A. Mikheev *et al.* pour leur système testé sur les corpus de MUC-7, avec des lexiques de tailles et de contenus différents.

derniers ayant par ailleurs de bonnes preuves internes et externes, mais qu'ils paraissent en revanche primordiaux pour les noms de lieux. Le troisième essai consista en l'utilisation d'un lexique de taille réduite (200 noms) et limité aux lieux : l'annotation des entités de ce type retrouve alors de bonnes performances. Enfin, la dernière expérience fut l'utilisation de lexiques « limités » (cf. colonne *Ltd* dans le tableau 1.1), construits à partir d'un tiers du corpus d'entraînement de MUC-7. Acquis à peu de frais, ces derniers permettent au système de retrouver des performances fort honorables pour chacun des types annotés.

Ces quatre expériences d'annotation d'entités nommées avec des lexiques plus ou moins gros et plus ou moins complets apportent quelques éléments de réponses aux questions liminaires d'A. Mikheev. À la question de l'importance des lexiques lors d'un processus de reconnaissance d'entités nommées, il convient de répondre que, si ces derniers ne sont pas primordiaux, leur rôle est tout de même non négligeable : l'absence totale de lexiques fait baisser de manière significative les performances. Une fois démontrée et admise la nécessité de lexiques, la deuxième interrogation porte à considérer la taille de ces derniers. Les expériences décrites ci-avant, faisant intervenir des lexiques limités (généraux ou de lieux seulement), plaident en la faveur de lexiques de taille minimale, ne comportant que les noms les plus courants. A. Mikheev *et al.* remarquent en effet que ces derniers, étant les plus connus des locuteurs, sont introduits « sans ménagement » dans le discours, tandis que les noms les moins connus ont de plus grandes chances d'être accompagnés de preuves internes et/ou externes[1]. T. Poibeau , menant à bien une expérience similaire de variation de taille de lexiques pour un système travaillant sur les corpus MUC-6 [Poibeau, 2003], voit décroître les résultats à mesure de la réduction de la taille des listes, mais constate également que la courbe des performances atteint un certain plafond autour d'un seuil d'environ 25 000 mots par lexique. Ceci confirme le fait que, s'il n'existe pas bien sûr de taille « idéale », pour les lexiques en reconnaissance d'entité nommées, rien ne sert d'utiliser 200 000 ou 100 mots, le raisonnable semblant se situer entre 15 000 et 30 000 mots. Enfin,

[1]Dans [Mikheev *et al.*, 1999] : « *When collecting these gazetteers, one can concentrate on the obvious examples of location and organisation, since these are exactly the ones that will be introduced in text without much helpful context* » (nous soulignons).

sur quels critères se baser pour constituer des lexiques ? Le propos précédent y a déjà répondu en partie : il semble judicieux de s'intéresser aux mots les plus communs car les moins spécifiés en discours. Par ailleurs, le type d'entité nommée paraît être digne de considération, les noms de lieux étant de toute évidence plus sensibles à la présence ou non de lexiques. En dernier lieu, le domaine d'application constitue sans conteste un critère de constitution de lexiques, un corpus de nature journalistique ne mobilisant pas les mêmes unités lexicales qu'un corpus scientifique.

Qu'il soit question de leur acquisition ou bien de leur exploitation, les lexiques n'ont de toute évidence pas manqué de susciter de nombreux débats. Il s'agit d'une composante essentielle de tout système de reconnaissance d'entités nommées et cela est donc tout naturel. De tout cela il se dégage que, manifestement indispensables, ces lexiques sont tout de même à constituer avec précaution, afin que leur utilisation soit judicieuse et pleinement complémentaire des autres composants ou mécanismes impliqués dans la reconnaissance des entités nommées. Ces derniers méritent également attention, T. Poibeau en a proposé une étude qu'il convient dès lors d'examiner.

1.4.3.2 L'évaluation transparente de T. Poibeau

« Deconstructing Harry » [Poibeau, 2001], l'article, entreprend « *d'évaluer sur une base objective les différentes composantes d'un système de reconnaissance d'entités nommées* ». Menant à bien une sorte d'évaluation transparente (évaluation de type *glass box* et non plus de type *black box*), T. Poibeau propose ainsi un protocole de déconstruction d'un système générique de reconnaissance d'entités nommées afin d'en évaluer les différentes composantes ; il entreprend, au cours de ce protocole, d'examiner quatre points ou éléments principaux que sont : les lexiques, les grammaires, les techniques d'apprentissage et les mécanismes de révision. Les lexiques ayant été abondamment discutés, nous ne considérerons ici que les trois derniers éléments. Il convient avant tout de signaler que les expériences réalisées dans [Poibeau, 2001] l'ont été sur trois types de corpus : le corpus MUC-6 (dépêches du *Wall Street Journal*), un corpus Reuters (dépêches financières) et un corpus de courriers électroniques. Cet assortiment de corpus de natures différentes vise à nuancer l'appréciation des composantes évaluées en fonction de la nature des textes traités.

Une grammaire comporte des règles qui, sur la base de connaissances lexicales issues d'une analyse préalable ou de lexiques, permettent de prédire le type d'une entité nommée à partir de l'enchaînement des éléments lexicaux qui la composent et/ou l'encadrent. Sans revenir sur le fonctionnement précis de cette composante,

il importe de s'interroger sur son rôle et son importance au sein d'un système. T. Poibeau est parvenu à décomposer une grammaire et à chiffrer le taux d'utilisation des diverses règles. Il ressort de cette expérience que l'activation des règles suit la loi de Pareto, dite des 80/20 : un petit ensemble de règles permet d'annoter une grande partie des entités, tandis que de nombreuses règles supplémentaires sont nécessaires pour couvrir les entités restantes. Si l'on observe ce qu'il se passe d'un type de corpus à un autre, il apparaît que moins le corpus est régulier[1], plus l'activation des différentes règles est d'un ordre comparable.

Autre point évalué par T. Poibeau : les mécanismes d'inférence, de généralisation et de révision. Ces mécanismes, implantés dans la plupart des systèmes, offrent la possibilité d'annoter des occurrences d'entités inconnues, ou pour lesquelles le contexte n'est pas suffisamment discriminant, en se servant de leur éventuelle annotation déjà effectuée précédemment dans le texte. Un système peut par exemple rencontrer l'entité suivante, *Kosciusko-Morizet*, et ne pas pouvoir l'annoter faute d'indices et d'information suffisants ; un mécanisme d'induction et de généralisation peut toutefois l'aider à le faire, en lui rappelant son annotation précédente de *Mme Kosciusko-Morizet* en tant que personne. Cette annotation dynamique peut encore se transformer en véritable connaissance, avec le stockage de cette entité dans un lexique intermédiaire puis, après validation humaine, dans un lexique proprement dit. Cette induction-généralisation ne présente, d'après les résultats des expériences, que peu d'intérêt pour les corpus réguliers et homogènes. Les performances d'annotation sur le corpus MUC-6 ne varient en effet qu'imperceptiblement avec ou sans ce mécanisme. Celui-ci est en revanche beaucoup plus productif pour les deux autres corpus : les performances augmentent presque de 20 % pour l'annotation des courriers électroniques, montrant ainsi que plus la qualité rédactionnelle diminue, plus il importe d'identifier des mots inconnus sur la base d'annotations précédentes. Le mécanisme de révision consiste quant à lui à « corriger » une annotation en fonction d'indices contraires. L'exemple le plus original en est bien sûr l'entité *Washington* qui, annotée le plus souvent par défaut comme un nom de lieu, peut se voir attribuer le type personne si rencontrée dans le contexte *Mrs. Washington*. Il est alors possible de réviser en conséquence les autres occurrences de cette entité dans le reste du document. Relativement aux performances, ce mécanisme ne permet qu'un gain minime.

Enfin, le dernier élément proposé à l'évaluation correspond aux techniques d'apprentissage. En effet, pour se dégager quelque peu des discussions apparem-

[1]Benoît Habert définit la notion de corpus de la manière suivante : « *un corpus est une sélection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extra-linguistiques explicites pour servir d'échantillon d'emplois déterminées d'une langue* » [Habert, 2000]. La régularité à laquelle nous faisons référence ici concerne plus particulièrement l'aspect linguistique des corpus, c'est-à-dire le niveau de langue et la manière dont sont introduites en discours les entités nommées.

ment sans fin autour des avantages et inconvénients des méthodes symboliques *vs.* les méthodes à base d'apprentissage, le plus profitable est de déterminer l'importance de chacune des approches en situation réelle. Pour ce faire, T. Poibeau a imaginé une mesure de prédiction des gains possibles dûs aux mécanismes d'apprentissage. La mesure prend en compte les éléments suivants :

1. Proportion d'entités nommées déjà répertoriée dans le dictionnaire relativement à la taille du corpus.
2. Proportion de séquences constituées d'une amorce suivie d'un mot inconnu relativement à la taille du corpus.
3. Proportion de mots inconnus commençant par une majuscule relativement à la taille du corpus.

Nous ne reproduisons pas ici la formule de calcul utilisée par l'auteur. Cette dernière permet néanmoins de chiffrer l'apport possible des techniques d'apprentissage pour les trois corpus, ce dernier étant bien sûr le plus élévé pour les courriers électroniques. En effet, moins un corpus est régulier et homogène, moins son traitement est assuré par les grammaires et les lexiques.

De ce panel d'évaluations, il ressort qu'aucune méthode et qu'aucun mécanisme n'est prépondérant ou préférable aux autres dans un système de reconnaissance d'entités nommées. Lexiques, grammaires, inférences, apprentissage, tout est efficace et nécessaire, en un panachage prenant en compte les spécificités du corpus avant tout.

Ainsi, si la reconnaissance d'entités nommées peut s'appuyer sur de nombreux indices, elle peut également être mise en œuvre selon différentes méthodes. Les multiples combinaisons possibles de divers composants et mécanismes ont conduit à la réalisation de nombreux systèmes et alimenté moult discussions. Au-delà de ces (presque) polémiques, le plus probant s'avère de combiner les différentes approches, en tenant compte des particularités du contexte de réalisation du système. Arrivés au terme de la description des systèmes de reconnaissances d'entités nommées, dernier point de ce chapitre, il est temps de rassembler l'ensemble du propos de celui-ci en un bilan récapitulatif.

1.5 Bilan

De sa définition à la description des différents types de systèmes s'attelant à sa réalisation en passant par ses applications, ce chapitre a ainsi présenté la tâche de reconnaissance d'entités nommées, en un état des lieux se voulant le plus complet possible. Le propos, retraçant les évolutions de cette tâche depuis son apparition jusqu'à ses performances actuelles, a ainsi permis de dégager les points suivants.

La tâche de reconnaissance des entités nommées, initialement définie pour améliorer la modularité et la portabilité des systèmes d'extraction d'information, a fait une apparition remarquée, voire inespérée, au milieu des années 1990 dans la communauté TAL. Remarquée, cela est certain : en un espace de temps assez bref (une décennie à peine) à l'échelle de l'histoire de la discipline, les entités nommées ont su occuper le devant de la scène des compétitions MUC et consœurs, suscitant un intérêt toujours plus soutenu. Inespérée, il faut en convenir également : qu'il s'agisse du renoncement momentané à la compréhension globale de textes ou des difficultés des premières campagnes d'évaluation autour de l'extraction d'information, la tâche de reconnaissance des entités nommées a su, par ses performances pour le moins remarquables, éclairer quelque peu l'horizon des recherches en TAL et redonner confiance et enthousiasme, si tant est que ces derniers eussent alors fléchi, à de nombreuses équipes de recherche. C'est d'ailleurs ce qu'affirme en substance B. Sundheim lorsqu'elle conclut son rapport sur MUC-6 : « *The introduction of two new tasks into the MUC evaluations and the restructuring of information extraction into two separate tasks have infused new life into the evaluations.* ». Résultats prometteurs et engouement furent ainsi les ingrédients d'un réel succès pour cette tâche, et ce dès son introduction.

Ce succès fut d'autant plus avéré qu'après s'être ainsi progressivement affirmée, la tâche de reconnaissance des entités nommées se révéla être de toutes les applications : en tant que module à l'intérieur d'un système d'analyse, offrant son concours tant pour l'analyse syntaxique que pour la résolution de coréférence ou encore la désambiguïsation lexicale, ou tout aussi bien en tant qu'application visant une tâche donnée, à l'exemple de l'extraction d'information naturellement, de la tâche de question-réponse ou de l'anonymisation. Ce faisant, de nombreux systèmes furent réfléchis pour sa mise en œuvre, conduisant à l'apparition de modules variés, purement linguistiques, à base d'apprentissage ou mixtes. Cette recherche ne fut pas, nous l'avons vu, sans de nombreux débats autour de l'importance de telle ou telle composante. Profitant de cet irremplaçable moteur de recherche (au sens propre) que sont les campagnes d'évaluation faisant se réunir autour d'une même tâche et pour un temps limité différentes équipes alors assignées à un but précis, les entités nommées ont ainsi vu se diversifier leurs moyens de reconnaissance, ceux-ci gagnant en maturité au fur et à mesure des années et des controverses, s'avérant au final d'autant plus efficaces que mêlant plusieurs approches.

Performances remarquables, applications nombreuses, exploration de diverses méthodes, force est de constater que, depuis le vide, ou l'absence de recherche, dénoncé en 1992 par S. Coates-Stephens [Coates-Stephens, 1992] à l'endroit de ce type d'unité, un véritable acquis s'est constitué, faisant désormais apparaître

cette tâche de reconnaissance des entités nommées comme un incontournable du TAL.

Ainsi, tâche « performante » et utile s'il en est, la reconnaissance d'entités nommées, telle que définie par les conférences MUC, ne semble accuser aucun bémol. Cependant, que sait-on au juste de cet ensemble d'unités ? Sait-on les définir ? Est-il si « facile » de savoir ce qu'il y a à reconnaître ? À y regarder de plus près, et c'est l'objet du chapitre suivant, nous verrons que la situation n'est pas aussi limpide que ce qui est, à l'instar de l'aperçu présenté ci-avant, communément admis et qu'il convient de considérer d'un œil attentif ce que recouvre précisément cette tâche avant d'en imaginer de nouvelles perspectives de recherches.