Annotation sémantique pour des domaines spécialisés et des ontologies riches

Yue Ma[1] François Lévy[2] Adeline Nazarenko[2]
(1) TU-Dresden, Germany
(2) LIPN, Université Paris 13-CNRS, France
mayue@tcs.inf.tu-dresden.de,
{francois.levy,adeline.nazarenko}@lipn.univ-paris13.fr

RÉSUMÉ

Explorer et maintenir une documentation technique est une tâche difficile pour laquelle on pourrait bénéficier d'un outillage efficace, à condition que les documents soient annotés sémantiquement. Les annotations doivent être riches, cohérentes, suffisamment spécialisées et s'appuyer sur un modèle sémantique explicite – habituellement une ontologie – qui modélise la sémantique du domaine cible. Il s'avère que les approches d'annotation traditionnelles donnent pour cette tâche des résultats limités. Nous proposons donc une nouvelle approche, l'annotation sémantique statistique basée sur les syntagmes, qui prédit les annotations sémantiques à partir d'un ensemble d'apprentissage réduit. Cette modélisation facilite l'annotation sémantique spécialisée au regard de modèles sémantiques de domaine arbitrairement riches. Nous l'évaluons à l'aide de plusieurs métriques et sur deux textes décrivant des réglementations métier. Notre approche obtient de bons résultats. En particulier, la F-mesure est de l'ordre de 91,9% et 97,6% pour la prédiction de l'étiquette et de la position avec différents paramètres. Cela suggère que les annotateurs humains peuvent être fortement aidés pour l'annotation sémantique dans des domaines spécifiques.

ABSTRACT

Semantic Annotation in Specific Domains with rich Ontologies

Technical documentations are generally difficult to explore and maintain. Powerful tools can help, but they require that the documents have been semantically annotated. The annotations must be sufficiently specialized, rich and consistent. They must rely on some explicit semantic model – usually an ontology – that represents the semantics of the target domain. We observed that traditional approaches have limited success on this task and we propose a novel approach, phrase-based statistical semantic annotation, for predicting semantic annotations from a limited training data set. Such a modeling makes the challenging problem, domain specific semantic annotation regarding arbitrarily rich semantic models, easily handled. Our approach achieved a good performance, with several evaluation metrics and on two different business regulatory texts. In particular, it obtained 91.9% and 97.65% F-measure in the label and position predictions with different settings. This suggests that human annotators can be highly supported in domain specific semantic annotation tasks.

MOTS-CLÉS : Annotation sémantique, Ontologie de domaine, Annotation automatique, Analyse sémantique des textes, Méthodes statistiques.

KEYWORDS: Semantic Annotation, Domain Ontology, Automatic annotation, Semantic Text Analysis, Statistical methods.

1 Introduction

Les documents techniques sont souvent complexes à lire et à maintenir mais ce sont des ressources critiques pour de nombreuses organisations. Les textes réglementaires décrivent les procédures, les règles et les politiques auxquels les organisations doivent se conformer ; ce sont des sources importantes, qui guide souvent la prise de décision dans ces organisations. Les instructions d'utilisation indiquent comment utiliser et maintenir des objets techniques qui sont parfois extrêmement complexes. Les experts ont besoin d'outils pour les aider à maîtriser et à valider ces documents autant que pour les maintenir à jour quand des évolutions techniques se produisent. Les textes sont de longueur variable (de quelques dizaines à plusieurs centaines de pages), mais souvent trop longs pour être faciles à lire, en particulier quand les informations importantes sont dispersées dans différentes parties. Ils contiennent des descriptions génériques plutôt que des exemples et reposent sur des vocabulaires spécialisés, qui sont souvent définis de façon plus ou moins formelle et précise dans des thesaurus ou des ontologies.

Il est possible d'aider les experts qui consultent ces textes en leur fournissant des outils. Le bénéfice est plus important si les documents sources sont enrichis par des informations sémantiques (ontologiques), qui assurent une certaine interopérabilité et qui permettent de faire des recherches sémantiques plutôt que de simples recherches de chaînes de caractères (Welty et Ide, 1999; Uren *et al.*, 2006; Nazarenko *et al.*, 2011). *L'annotation sémantique* aide à visualiser et à rassembler l'information importante, mais aussi à contrôler la documentation technique (vérification de cohérence, aide à la décision et traçabilité, mise à jour, etc.).

Des outils ont été développés, tels que GATE (Cunningham *et al.*, 2011) ou SemEx (Nazarenko *et al.*, 2011), pour explorer des textes dont certaines portions sont liées par des annotations à divers éléments d'un modèle sémantique de domaine. L'annotation sémantique automatique de la documentation technique spécialisée présente cependant deux caractéristiques importantes.

En premier lieu, les annotations sémantiques intéressantes étiquettent souvent des notions génériques ou des concepts plutôt que des mentions d'entités modélisées comme des instances de concepts. Ceci diffère de la Reconnaissance des Entités Nommées (REN) qui vise à repérer les instances de certains types sémantiques [1]. Par exemple, dans le texte de la figure 1, le fragment "Service conducting approval tests" est annoté par le concept *TestConductingService* et pas par l'une ses instances. Les notions génériques susceptibles d'être annotées sont plus nombreuses que les types canoniques des entités nommées, et les approches d'annotation sémantique traditionnelles sont handicapées dans ce cas par des caractéristiques moins régulières et des ressources plus rares. On observe que les méthodes d'annotation au regard d'une ontologie se concentrent généralement sur les instances de concepts dans une perspective de peuplement d'ontologies (Kiryakov *et al.*, 2004; Amardeilh *et al.*, 2005; Uren *et al.*, 2006).

The seats of the vehicle shall be fitted and shall be placed in the position for driving use chosen by the Technical Service conducting approval tests to give the most adverse conditions with respect to strength, compatible with installing the manikin in the vehicle. The positions of the seats shall be stated in the report.

FIGURE 1 – Exemple : texte réglementaire avec annotations sémantiques

1. Typiquement : Personne, Organisation, Lieu, Temps.

En second lieu, les ontologies génériques (par ex. DBpedia) utilisées par de nombreux services d'annotation sémantique ouverts sont peu utiles pour les documents techniques. Nous avons testé plusieurs d'entre elles sur un corpus traitant de la réglementation dans l'industrie automobile. Quatre produisent très peu d'annotations : OpenCalais [2], Zemanta [3] et DBpedia Spotlight (Mendes *et al.*, 2011) ont des rappels de 3,3 %, 0,8 % et 0 % ; AlchemyAPI [4] reconnaît la mention de deux organisations [5] et d'une ville [6], mais deux de ces annotations sont manifestement erronées dans le domaine considéré. A l'inverse, la Wiki Machine (LiveMemories, 2010) annote surabondamment le règlement : dans le fragment "In the case of an assembly incorporating a retractor", "case" est annoté par *Law, Justice* et "assembly" est lié à *Parliamentary procedure* et *Meetings*, mais ce n'est pas le sens qu'ont ces termes dans nos données. Ces annotateurs du Web basés sur des ontologies publiques renvoient souvent une interprétation trompeuse des textes spécialisés.

Nous en concluons qu'un système d'annotation sémantique des documents techniques devrait avoir les propriétés suivantes : (1) pouvoir noter un concept et pas seulement des instances de types généraux comme signification d'un terme ; (2) fournir une interprétation précise et fiable, en tenant compte des modèles sémantiques du domaine traité ; (3) avoir une bonne couverture du texte, de sorte que les fragments textuels intéressants puissent être facilement détectés et reliés. Notre approche repose sur le constat qu'un expert métier peut fournir un petit nombre d'exemples annotés manuellement, mais ne peut pas annoter des documents volumineux. Nous avons vu que les approches de l'état de l'art répondent mal à ces spécifications.

Nous proposons donc une nouvelle approche d'annotation, à la fois simple et naturelle, qui s'inspire de la traduction automatique basée sur les syntagmes et qui est adaptée à l'annotation spécialisée requise par les textes techniques.

Nous transposons le modèle de la traduction automatique statistique (TAS) basée sur les syntagmes à notre problème et nous montrons expérimentalement, avec différentes métriques d'évaluation, que l'annotateur ainsi construit obtient des résultats significatifs à partir d'un corpus réduit annoté manuellement. Par effet de bord, il peut intégrer dans un modèle unique les interprétations que différents experts auraient données du même texte. Les expériences rapportées ici portent sur un règlement international sur le contrôle des ceintures de sécurité (par la suite « Règlement des ceintures de sécurité »), auquel les constructeurs d'automobiles doivent se conformer.

Le reste de l'article est structuré ainsi : nous discutons l'état de l'art dans la section qui suit et définissons la tâche dans la section 3. Puis notre méthode est décrite dans la section 4. Les expériences et leur évaluation sont présentées dans les sections 5 et 6.

2. `http://www.opencalais.com`
3. `http://www.zemanta.com`
4. `http://www.alchemyapi.com`
5. "cabinet" dans "... shall be placed in a refrigerated cabinet at -10 C + 1 C for two hours" et"Technical Service" dans "One of these axes shall be in the direction chosen by the Technical Service conducting the approval test".
6. "anchorage" dans la phrase "except in the case of retractors having a pulley or strap guide at the upper belt anchorage".

2 Etat de l'art

Les deux facettes de notre problème, prédire les labels sémantiques et les frontières de ces étiquettes, se retrouvent dans la REN (Nadeau et Sekine, 2007) et les annotateurs du Web sémantique (Uren *et al.*, 2006). Dans la REN, les étiquetages sont souvent limités à quelques grandes catégories génériques comme *Personne, Endroit, Organization, Produit, et Date*. Quant aux annotateurs du Web, dont les étiquetages proviennent généralement d'ontologies générales, comme TAP (Dill *et al.*, 2003), DBpedia Lexicalization Dataset [7]), ils ne sont pas efficaces pour les textes spécialisés et des domaines différents. De plus, ils privilégient souvent la précision au détriment du rappel, produisant moins de deux annotations par page en moyenne (Dill *et al.*, 2003; Mihalcea et Csomai, 2007; Cucerzan, 2007).

Un premier type d'approches de l'annotation sémantique consiste à appliquer des règles sur des segments sélectionnés par des *wrappers* (Ciravegna, 2003; Etzioni *et al.*, 2004; Cimiano *et al.*, 2004). S'agissant d'une annotation sémantique précise et spécialisée, il est difficile d'apprendre des règles pour chaque type d'annotation, à cause du grand nombre de catégories sémantiques. De plus, les règles sont souvent plus complexes que pour la REN, où les entités cibles ont généralement une forme particulière (par ex. débutant par une majuscule) ou sont associées à des déclencheurs comme un titre (par ex. « M. », « Le président »). Dans l'annotation spécialisée, les fragments de texte à annoter sont très variés et leurs frontières sont difficiles à identifier. Par exemple, dans le règlement des ceintures de sécurité, "tested according to paragraph 7.6.4.2." a été étiqueté manuellement par $Method$ (voir section 5).

Une seconde famille d'approches d'annotation sémantique repose sur des modèles statistiques ou l'apprentissage automatique (par ex. HMM (Zhou et Su, 2002; Ratinov et Roth, 2009), CRF (Finkel et Manning, 2009), et Perceptron ou Winnow (Collins, 2002)). Ces approches exploitent la richesse des ressources textuelles du Web (Dill *et al.*, 2003; LiveMemories, 2010; Mendes *et al.*, 2011) ou de journaux (Nadeau et Sekine, 2007; Ratinov et Roth, 2009) comme données d'entraînement pour la désambiguïsation. Le traitement de l'ambiguïté est important quand on considère différents niveaux de granularité ontologique : selon le contexte, un terme comme "test" peut faire référence au concept génénral $Test$, à une catégorie précise de tests où à une instance de test particulière. Cependant, dans les domaines spécialisés, on a rarement de gros volumes de données. Notre approche repose sur un modèle statistique différent, qui prend en compte la forme brute des textes (sans traitement linguistique préalable) et montre de meilleures performances que les champs aléatoires conditionnels en chaînes linéaires (CRF) sur un petit volume de données.

Les recherches sur la REN dans des corpus spécialisés (Wang, 2009; Liu *et al.*, 2011) indiquent qu'il faudrait entraîner des systèmes d'annotation spécifiques même dans le cas où le jeu d'étiquettes est le même que pour la REN classique (Wang, 2009; Liu *et al.*, 2011) quand le corpus est spécialisé (ex. Tweet, notes cliniques). Le présent travail s'intéresse aux cas où le corpus et les jeux d'étiquettes sont spécialisés, comme dans (Aronson et Lang, 2010; Müller *et al.*, 2004) qui proposent un entraînement spécialisé pour la biomédicine. A la différence de cette approche qui est difficile à adapter à un autre domaine, notre méthode, fondée sur TAS, peut être facilement appliquée sur un autre domaine spécialisé à condition qu'un petit volume de données d'entrainement annotées soit disponible.

7. `http://dbpedia.org/Lexicalizations`

Les modèles de TAS ont été appliqués à d'autres questions que la traduction, en particulier à la normalisation de textes et de SMS (Aw *et al.*, 2006; Beaufort *et al.*, 2010) et à l'analyse sémantique (Wong et Mooney, 2006). Selon ces auteurs, leurs résultats, mesurés par les métriques de traduction automatique, sont bons. S'agissant de l'annotation sémantique de documents spécialisés, nous adoptons nous aussi un modèle de TAS basée sur les syntagmes, mais nous l'évaluons différemment parce que les métriques de traduction automatique s'avèrent limitées pour notre tâche.

3 Définition de la tâche

On dispose d'une ontologie pour un domaine spécialisé dont le volet lexical est utilisé pour annoter un petit corpus d'entraînement. La tâche consiste à identifier à la fois les frontières et la catégorie ontologique des éléments sémantiques majeurs de chaque phrase du corpus à annoter.

En plus des termes spécialisés qu'il est utile de détecter et d'annoter, un autre problème fréquent pour l'annotation sémantique de documents techniques au regard d'une ontologie riche est qu'un grand nombre de mots identiques en surface peuvent être annotés avec plusieurs étiquettes ontologiques qui ne sont pas logiquement disjointes comme c'est le cas dans l'homonymie, mais qui reflètent simplement une granularité de sens variable en contexte. Par exemple, dans les quatre phrases ci-dessous, « test » a été annoté par l'expert comme $BuckleTest$ à trois reprises (S1, S2 et S3) et $Method$ une fois (S4). La résolution de l'ambiguïté est importante pour le succès de cette tâche.

S1. *The force required to open the buckle in the test as prescribed in paragraph 7.8. below shall not exceed 6 daN.*

S2. *In the case of harness belt buckles, this test may be carried out without all the tongues being introduced.*

S3. *In the case of buckles which incorporate a component common to two assemblies, the strength and release tests of paragraphs 7.7. and 7.8. shall also be carried out with the part of the buckle pertaining to one assembly being engaged in the mating part pertaining to the other, if it is possible for the buckle to be so assembled in use.*

S4. *Retractors shall be subjected to tests and shall fulfill the requirements specified below, including the tests for strength prescribed in paragraphs 7.5.1. and 7.5.2.*

4 Annotation sémantique statistique basée sur les syntagmes

Nous modélisons l'annotation sémantique des documents spécialisés comme une tâche de traduction automatique ayant les caractéristiques suivantes : (1) les unités textuelles pertinentes pour traduire ou annoter sont des syntagmes plutôt que de simples mots ; (2) de même qu'un mot peut être traduit de différentes façons, on peut annoter un fragment de texte de plusieurs manières, des éléments ontologiques différents pouvant avoir des lexicalisations communes.

4.1 L'annotation sémantique en tant que traduction automatique

Dans cette vision d'une annotation sémantique comme traduction, le texte initial non annoté est considéré comme le texte à « traduire » et le texte annoté comme le texte cible « traduit ».

Formellement, on a deux phrases $\langle s_1, s_2 \rangle$ dans deux « langages » L_1 and L_2 : L_1 est ici l'anglais et $L_2 = L_1 \cup Voc(O)$ est l'union de l'anglais et du vocabulaire de l'ontologie, $Voc(O)$, utilisé comme ensemble d'étiquettes [8]. Nous disons que s_2 est une version annotée de s_1 s'il est obtenu en remplaçant certains groupes de mots anglais de s_1 par des éléments de $Voc(O)$ comme illustré dans la figure 2.

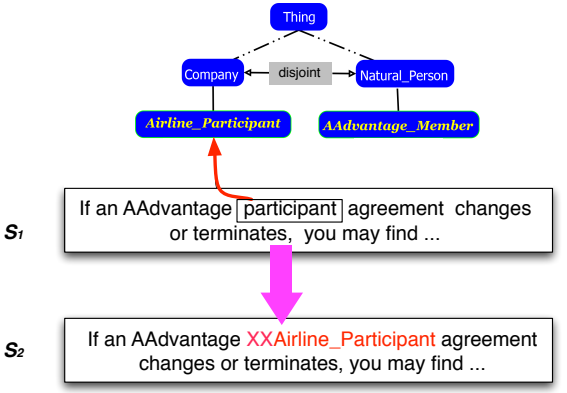

FIGURE 2 – L'annotation sémantique en tant que traduction

D'après (Tomeh, 2012), la TAS a conceptuellement trois étapes [9] : 1) les phrases appariées sont alignées sur les mots – ou les syntagmes – pour constituer la relation de traduction qui spécifie quel élément de s_2 est la traduction de quel élément de s_1 ; 2) des règles de traduction sont apprises sur ces données, en général en s'appuyant sur une table de traduction ; 3) chaque phrase à traduire est segmentée en syntagmes qui sont traduits séparément puis réordonnés pour adapter le résultat au langage cible. Quand il s'agit d'annotation sémantique, la relation de traduction est monotone (sans réarrangement). C'est même l'identité pour tous les éléments qui restent non-annotés. Les données en entrée de l'algorithme d'apprentissage sont donc moins bruitées que dans le cas d'un alignement bilingue. L'obtention d'annotations correctes quand l'information lexicale est ambigüe repose sur l'algorithme d'apprentissage et la projection de ses résultats sur le texte, dans la mesure où cet algorithme prend en compte le contexte pour apprendre les règles. A noter que le modèle tient compte dans ses calculs des éléments qui ne doivent pas être annotés : il apprend aussi à traduire à l'identique.

8. Pour différentier les éléments de $Voc(O)$ du vocabulaire anglais, les noms de O sont préfixés par 'XX' dans L_2.
9. Même si elles peuvent être entrelacées dans le calcul.

4.2 Le modèle

Notre approche repose sur le modèle du canal bruité, qui considère que les phrases annotées constituent l'information visée (en entrée du canal) mais qu'elles ont été brouillées, produisant ainsi le texte brut (reçu en sortie). Il s'agit donc de reconstituer l'entrée. On attribue une étiquette sémantique à une phrase vue pour la première fois $s_1 \in L_1$ en recherchant la phrase $s_2 \in L_2$ qui a la plus grande valeur pour $P(s_2|s_1)$. Par la règle de Bayes et puisque $P(s_1)$ est fixée, il s'agit de calculer

$$s^* = \arg\max_{s_2} P(s_2|s_1) = \arg\max_{s_2} \{P(s_2)P(s_1|s_2)\}.$$

Suivant le modèle de traduction basé sur les syntagmes, la phrase d'entrée non annotée s_1 est segmentée pendant le décodage en une suite de m syntagmes, notée $\{s_1^i\}_{i=1}^m$. Chaque segment s_1^i est associé à sa version annotée s_2^i de sorte que $P(s_1|s_2) = \Pi_{i=1}^m P(s_1^i|s_2^i)$. On suppose que la distribution de probabilité sur toutes les segmentations possibles est uniforme et on a

$$s^* = \arg\max_{s_2} \{P(s_2) \times \Pi_{i=1}^m P(s_1^i|s_2^i)\}.$$

Il y a deux paramètres à calculer dans le modèle ci-dessus : le modèle de langage $P(s_2)$ et la table de traduction des syntagmes $P(s_1^i|s_2^i)$. Le modèle de langage sélectionne la phrase annotée la plus probable parmi toutes celles qui sont possibles et la table de traduction des syntagmes joue le rôle d'un dictionnaire sophistiqué entre les langages source et cible. Nous ne pouvons entrer ici dans les détails, mais, pour nos expérimentations, nous utilisons SRILM (Stolcke, 2002), la boite à outils du SRI servant à construire et exploiter des modèles de langage, pour apprendre un modèle de trigrammes. Parmi les nombreuses solutions proposées pour l'apprentissage d'une table de traduction (Marcu et Wong, 2002; Koehn *et al.*, 2003; Och et Ney, 2003; Chiang, 2007), nous utilisons la méthode relativement simple mais efficace définie dans (Koehn *et al.*, 2003). A cause de la proximité des langages source et cible, les données fournies à cet algorithme sont peu bruitées. Le décodage est réalisé par une recherche en faisceau telle qu'implémentée par Moses (Koehn *et al.*, 2007).

4.3 Repérage des annotations sémantiques

Pour identifier la position précise des annotations sémantiques prédites par l'annotation sémantique statistique basée sur les syntagmes (ASSS), nous utilisons l'alignement des traductions au niveau du mot. Par exemple, dans un tel alignement, la suite "15-14 16-14" indique que les 15ème et 16ème mots de la phrase originale ont été remplacés par le 14ème mot de la traduction. Si le 14ème mot appartient à $Voc(O)$ (par exemple $XXMethod$), c'est que le concept qui la compose (dans notre exemple, le concept $Method$) est l'étiquette sémantique associée au 15ème et au 16ème mots de la phrase originale.

5 Expérimentation

Cette section décrit les données d'évaluation et les métriques utilisées.

L'approche ASSS a été testée sur deux textes annotés. L'un est complètement annoté, c'est-à-dire annoté par l'ensemble des étiquettes sémantiques provenant d'une ontologie construite pour le domaine en question. On trouve dans le texte des mentions de chaque concept, mais en nombre limité : ce corpus permet de tester la tolérance de notre approche à la dispersion des données d'annotation sémantique. L'autre texte est plus volumineux mais il n'est annoté que par une partie de l'ontologie, par les 17 concepts identifiés considérés comme ambigus, car étant associés à des termes ambigus. Ce second corpus permet de tester la capacité de notre approche à résoudre les ambiguïtés, qui sont fréquentes en domaine de spécialité, ne serait-ce parce qu'on peut choisir de rattacher un terme à différents niveaux de l'ontologie.

Deux méthodes de référence ont été définies et sont utilisées pour les expériences. La première est une approche à base de dictionnaire de fréquence qui est traditionnelle pour les tâches de désambiguïsation lexicale et qui peut s'étendre à notre scénario d'annotation sémantique. L'autre repose sur un modèle basé sur l'étiquetage de séquences, plus particulièrement sur les champs aléatoires conditionnels en chaînes linéaires (CRF) (Lafferty, 2001; Sutton et Mccallum, 2006)) : l'annotation sémantique est souvent vue comme une tâche d'étiquetage de séquences et les champs aléatoires conditionnels permettent de tenir compte des noeuds voisins dans un graphe. L'évaluation expérimentale montre que notre méthode dépasse significativement des approches standards sur les deux corpus utilisés.

5.1 Données d'évaluation

Matériel [10] Les corpus choisis sont deux textes extraits d'un règlement international décrivant les tests auxquels les fabricants d'automobiles doivent se plier dans la fabrication des ceintures de sécurité. Après segmentation par Treetagger (Schmid, 1995), le corpus 1 comporte 133 phrases et le corpus 2 en a 1821, dont beaucoup sont longues. L'ontologie [11] qui forme le modèle sémantique contient 154 *entités sémantiques* (73 concepts, 58 individus, 23 propriétés).

Annotation sémantique le corpus 1 a été complètement annoté par un expert du domaine (un des auteurs), soit 364 annotations (2,78 annotations par phrase). Pour la validation croisée, 90 % des données sont utilisées comme données d'entraînement (80 % servent à entrainer le modèle, et 10 % au tunning) et les 10 % restant sont utilisées comme données de test. Les données d'entraînement de chaque expérience comportent plus de 50 entrées sémantiques distinctes.

Deux facteurs principaux ont été pris en compte dans la constitution du corpus 2 : le degré d'ambiguïté (une même forme lexicale peut être annotée différemment dans des contextes différents – voir la section 3 pour un exemple) et la taille du corpus, de façon que notre seconde méthode de référence puisse être calculée en un temps raisonnable eu égard à nos ressources de calcul (Mac OS X 10.6.8, g++ 4.2.1, avec 2Go de mémoire et un CPU Intel Core 2 Duo 2.26GHz). Nous avons sélectionné 17 entités sémantiques ambiguës de l'ontologie, nous nous en sommes servis pour annoter le document entier et nous avons sélectionné les 313 phrases étiquetées au moins une fois.

Pour le corpus 2, la table 1 liste les graphies choisies, le nombre de leurs occurrences annotées et les 17 étiquettes sémantiques qui leur sont associées. Il y a aussi 14 occurrences supplémentaires

10. Ce matériel vient du projet européen OntoRule.
11. A noter que, une fois que les exemples annotés sont disponibles, notre méthode n'a plus besoin de l'ontologie.

Graphie	#Occ	Etiquettes possibles dans la référence
"type"	123	$TypeReactor, TypeRetractor, TypeBelt$
"tested"	29	$RetractorLockingTest$, $BreakingStrengthOfStrapTest$, $DynamicTest$, $Method$, $ColdImpactTest$, NULL
"Test(s)"	19	$DynamicTest, Method$
"tests"	65	$DynamicTest$, $RetractorDurabilityTest$, $Method$, $BreakingStrengthOfStrapTest$, $AccelerationTest$, $DecelerationTest, RetractorLockingTest, BuckleTest$
"test"	190	$ColdImpactTest$, $RetractorLockingTest$, $Method$, $MicroSlipTest$, $BuckleOpeningTest$, $BreakingStrengthOfStrapTest$, $DynamicTest$, $BuckleTest$, $CorrosionTest, RetractorUnlockingTest, FrontalImpactTest$

TABLE 1 – Description des ambiguïtés du corpus2

de « tested » non annotées (notées NULL en ligne 2, colonne 3), ce qui constitue une forme particulière d'ambiguïté.

5.2 Annotations de référence

Nous comparons l'approche proposée avec deux méthodes de référence : l'Annotation Sémantique à base de Dictionnaire et de Fréquence (ASDF) et l'Annotation Sémantique par CRF (ASCRF).

L'approche ASDF est une extension de la désambiguïsation lexicale classique parce qu'elle intègre le fait qu'une annotation peut couvrir plusieurs mots. L'ASDF repose essentiellement sur la construction et la consultation d'un dictionnaire d'annotation. Celui-ci a comme entrées des mots ou groupes de mots associés à des labels sémantiques. Ces groupes sont extraits des textes annotés d'entraînement, et pour chaque mot ou groupe de mots, les labels sémantiques qui annotent ses occurrences sont enregistrés dans le dictionnaire. L'entrée est lemmatisée pour s'affranchir des variations morphologiques. L'algorithme d'annotation cherche d'abord dans le texte lemmatisé les entrées du dictionnaire. Une forme de surface reconnue pouvant être incluse dans une autre, seules les entités sémantiques attachées à la plus longue sont conservées. Pour désambiguïser une entrée donnée, on choisit le label le plus fréquent. ASDF est implémentée en Python.

ASCRF segmente et annote les séquences de mots grâce au modèle discriminant suivant :

$$p_\theta(y \mid x) = \frac{1}{Z_\theta(x)} exp\{\sum_{k=1}^{K} \theta_k F_k(x,y)\},$$

où $x = (x_1, ...x_T)$ et $y = (y_1, ..., y_T)$ sont les séquences d'entrée et de sortie ; $F_k(x,y)$ est défini par $\sum_{t=1}^{T} f_k(x_{t-1}, y_t)$, $\{f_k\}_{1 \leq k \leq K}$ étant un ensemble arbitraire de fonctions de traits ; les $\{\theta_k\}_{k \leq k \leq K}$ sont les valeurs paramétriques associées.

Pour être comparables avec le modèle ASSS proposé, les patrons extraits par ASCRF sont des traits orthographiques et lexicaux des unigrammes et des bigrammes figurant dans une fenêtre

de 3 mots avant et après chaque position observée. Comme les annotations s'étendent éventuellement sur plusieurs mots, elles sont représentées selon le schème D.I.E. (le Début, l'Intérieur et l'Extérieur du segment de texte. Enfin, ASCRF est mis en œuvre grâce à l'implémentation hautement optimisée de la boite à outils Wapiti (Lavergne *et al.*, 2010).

5.3 Métriques d'évaluation

Bien que nous utilisions un modèle de traduction automatique, le système est évalué en calculant la précision, le rappel et la F-mesure, qui sont plus souvent utilisés dans le domaine de l'extraction d'information. Nous considérons en outre deux critères différents : la correction des étiquettes sémantiques (*label*) et celle de leurs frontières (*position*). Bien que seul le critère d'étiquette importe dans certaines applications, comme en REN, la position peut être significative dans d'autres cas, comme par exemple pour l'extraction de relations sémantiques. On peut former d'autres mesures par combinaison des précédentes, comme *label et position considérés indépendamment* (le score cumule l'évaluation des labels et des positions) et *label et position considérés groupés*. Dans ce dernier cas, c'est le couple (label, position) qui est considéré globalement comme correct ou incorrect.

Pour chaque métrique μ parmi {Précision, Rappel, F-mesure}, nous écrivons $\mu\text{-}label$, $\mu\text{-}position$, $\mu\text{-}indep$, et $\mu\text{-}couple$ pour désigner les quatre critères d'évaluation ci-dessus [12]. Pour $\mu\text{-}position$, le critère est l'identité des positions de l'annotation dans le candidat et la référence, même si on pourrait aussi tenir compte du recouvrement partiels des positions.

6 Evaluation

Dans cette section, nous comparons d'abord la méthode ASSS proposée et le système ASDF. Ensuite, nous comparons ASSS et ASCRF sur les même corpus sous des réglages différents. Pour ces deux comparaisons, les expériences ont été effectuées sur les deux corpus en utilisant une validation croisée par 10ème. Pour ASCRF, nous avons partiellement réutilisé la mise en œuvre de MOSES (Koehn *et al.*, 2007) en inactivant son modèle de distorsion.

6.1 Comparaison de ASSS et ASDF sur le corpus 1

Le tableau 2 compare les performances moyennes de ASDF et ASSS sur le corpus 1 et les confronte à ceux de l'approche hybride définie ci-après.

ASSS a été légèrement meilleur pour la prédiction des étiquettes que ASDF (0,26 % d'amélioration de la F-mesure), mais ASDF a fonctionné mieux qu'ASSS dans la prédiction des positions (+5,2 % sur la F-mesure). Les deux systèmes ont réalisé des performances comparables sur le corpus 1.

Cela signifie que si l'on ne considère que les labels d'annotations, la méthode ASSS est un meilleur choix : contrairement à la consultation de dictionnaires, ASSS permet une correspondance approchée. Cependant, ASSS manque plus souvent l'emplacement exact de l'étiquette. Par

12. Dans la section Expérimentation, $\mu\text{-}indep$ et $\mu\text{-}couple$ ne figurent qu'à titre d'explication ; en fait ces mesures sont des combinaisons des deux autres.

Métrique	ASDF	ASSS	Hybride
F-mesure d'étiquette	0,9885	**0,9911**	**0,9911**
F-mesure de position	**0,9858**	0,9369	0,9797

TABLE 2 – Evaluation de la ASSS et ASDF sur Corpus1

exemple, alors que la phrase "The test has to be performed separately from the tensile test" a été annotée avec *Tensiletest* pour la position | 9 - 10 | par l'expert, ASSS n'a associé l'étiquetage *Tensiletest* qu'à la position | 9 - 9 |[13]. Pour remédier à cela, nous faisons une combinaison de ASSS et ASDF pour avoir un système hybride (la 4ème colonne du tableau 2) défini comme suit :

Définition. *(*Hybride de ASSS et ASDF*) Pour une phrase donnée, soit ANNO$_{ASSS}$ et ANNO$_{ASDF}$ les annotations sémantiques générées respectivement par ASSS et ASDF. Nous disons que deux annotations provenant de ANNO$_{ASSS}$ et ANNO$_{ASDF}$ sont* unifiables *si leurs positions se chevauchent.*

Dans le tableau 2, nous pouvons voir que le système hybride a la même F-mesure de label et a amélioré la F-mesure de position d'ASSS de 4,28 %, même si cette dernière est encore inférieure de 0,61 % à celle d'ASDF.

6.2 Comparaison d'ASSS et ASDF sur le corpus 2

Le tableau 3 montre l'intérêt de l'approche ASSS en cas d'ambiguïté. ASSS$_{all}$ signifie que l'expérience a été réalisée sur les 313 phrases du Corpus 2 (sélectionnées pour la présence de syntagmes ambigus) mais qu'elles sont annotées avec *toutes* les entrées sémantiques possibles de l'ontologie. Nous pouvons voir qu'ASSS est robuste à l'ambiguïté, comme en témoigne la F-mesure de label à 92,95 %.

Une autre observation est que, sauf pour la perte de 1,04 % de rappel de position, ASSS a de meilleures performances qu'ASDF. En effet, les différences entre ASSS et ASDF sont importantes pour la prédiction des étiquettes (par exemple +21,17 % pour la F-mesure de label), mais assez faibles pour la prédiction des positions (par exemple +2,21 % pour la F-mesure de position). L'explication est que le choix des annotations appropriées est plus difficile que la localisation de ces annotations dans le corpus 2, en raison d'une plus grande proportion d'ambiguïtés dans le corpus 2 que dans le corpus 1.

Enfin, le tableau 3 montre que même si ASSS a obtenu des scores élevés dans la prédiction de label sur le corpus 2, les scores sont encore inférieurs à ceux de la position (par exemple une F-mesure de 92,95 % en prédiction de label *vs.* 97,65 % en prédiction de position), ce qui contredit le résultat du corpus 1. C'est encore parce que dans le corpus 2, la désambiguïsation d'étiquettes est plus difficile à réaliser que détection de la position.

Il convient enfin de noter que l'approche ASSS fonctionnant mieux en prédiction de position qu'ASDF (97,65 % contre 95,44% pour la F-mesure de position) pour le corpus 2, l'approche hybride considérée pour le corpus 1 est inutile pour le corpus 2.

13. On rappelle que, dans nos définitions, seules les positions exactes (mêmes début et fin) sont comptées correctes dans la F-mesure.

Métriques	ASDF	$ASSS_{all}$	$ASSS_{all}$-ASDF
Précision-groupe	0,7288	0,9369	0,2081
Précision-label	0,7525	0,9369	**0,1844**
Précision-indep	0,8613	0,9598	0,0985
Précision-position	0,9293	0,9826	**0,0533**
Rappel-groupe	0,7699	0,9222	0,1523
Rappel-label	0,6861	0,9222	**0,2361**
Rappel-indep	0,9029	0,9464	0,0435
Rappel-position	0,9809	0,9705	**-0,0104**
F-mesure-label	0,7178	**0,9295**	**0,2117**
F-mesure-position	0,9544	**0,9765**	**0,0221**

TABLE 3 – Evaluation d'ASSS et ASDF sur le corpus 2

6.3 Comparaison d'ASSS et ASCRF

Le tableau 4 compare les performances moyennes d'ASCRF et de'ASSS à la fois sur le corpus 1 et sur le corpus 2. A la différence d'$ASSS_{tous}$ dans le tableau 3, $ASSS_{17}$ et $ASCRF_{17}$ correspondent au cas où les 313 phrases sélectionnées dans le corpus 2 ne sont annotées que par les 17 entrées sémantiques ambigües, ceci pour réduire le temps d'exécution d'ASCRF [14]. Les résultats montrent que :

– Sur le corpus 1, ASSS a supplanté ASCRF pour toutes les mesures. C'est parce que la taille des données d'entraînement dans le corpus 1 n'est pas suffisante pour qu'ASCRF parvienne à une prédiction précise. La comparaison avec la tableau 2 montre qu'ASCRF a fonctionné bien plus mal qu'ASDF sur le corpus 1. Cela signifie qu'ASSS est plus robuste qu'ASCRF lorsque la taille des données d'entraînement est limitée.
– Sur le corpus 2, $ASSS_{17}$ a surpassé $ASCRF_{17}$ de plus de 8 % pour la prédiction des étiquettes, à la fois en précision et en rappel, mais a été surpassé de 1,71 % en précision dans la prédiction de position. Cela montre qu'ASSS a une plus forte capacité de désambiguïsation qu'ASCRF, mais est moins bon qu'ASCRF pour placer les annotations parce que le modèle des positions d'étiquettes est implicite pour ASSS. De plus, il est intéressant de noter que les deux approches ASSS et ASCRF ont obtenu des scores relativement élevés en prédiction de position pour le corpus 2 (> 94 % en précision et en rappel).
– $ASSS_{all}$ a une meilleure performance que $ASCRF_{17}$ et $ASSS_{17}$ sur le corpus 2. Cela montre que le pourcentage plus élevé d'ambiguïtés dans les corpus $ASCRF_{17}$ et $ASSS_{17}$ augmente la difficulté de la tâche.

7 Conclusion et perspectives

Cet article propose une approche statistique basée sur les syntagmes, nouvelle et flexible, qui permet d'annoter les entités sémantiques dans des documents spécialisés en utilisant des onto-

14. Pour ASCRF, l'exécution de la validation croisée au $10^{ème}$ a duré 30 heures en se limitant aux 17 entrées sémantiques ambigües. Traiter toutes les entrées comme pour $ASSS_{tous}$ aurait nécessité beaucoup plus de temps parce que l'entraînement d'un modèle de CRF est quadratique en le nombre d'étiquettes (Lavergne *et al.*, 2011).

	l Corpus 1		Corpus 2		
Métrique	ASCRF	ASSS	$ASCRF_{17}$	$ASSS_{17}$	$ASSS_{all}$
Précision-label	0,8239	**0,9889**	0,8299	**0,9142**	*0,9369*
Précision-position	0,8975	**0,9389**	**0,9577**	0,9406	*0,9826*
Rappel-label	0,8239	**0,9889**	0,8308	**0,9235**	*0,9222*
Rappel-position	0,8975	**0,9349**	**0,9588**	0,9518	*0,9705*

TABLE 4 – Evaluation de ASSS et ASCRF

logies de domaine riches. La méthode a été conçue pour des documents techniques, tels que des textes réglementaires, pour lesquels les approches traditionnelles d'étiquetage sémantique (étiquetage des entités nommées et annotation sémantique générique) présentent des limitations importantes. En utilisant plusieurs métriques d'évaluation, nous avons montré que la méthode proposée donne de meilleurs résultats qu'une approche classique à base de dictionnaire de fréquence ou qu'une approche discriminante, avec un ensemble réduit d'exemples annotés. Elle obtient des scores élevés sur le corpus ambigu : une F-mesure de 92,95 % (resp. 97,65 %) pour la prédiction de label (resp. de position) pour $ASSS_{all}$, et une F-mesure de 91.88% (resp. 94,62 %) pour la prédiction de label (resp. de position) pour $ASSS_{17}$

Nous projetons maintenant d'améliorer notre approche en étendant ASSS pour utiliser des informations linguistiques rendues accessibles en pré-traitant les documents source. Nous envisageons aussi de concevoir des campagnes d'annotation ontologique dans des domaines spécialisés, en exploitant cette méthode qui permet d'entraîner un système d'annotation sur un petit ensemble d'annotations manuelles. En effet, il semble qu'il soit plus facile pour les annotateurs humains de corriger une annotation initiale, pourvu qu'elle soit suffisamment bonne, que d'annoter à partir de rien (Fort et Sagot, 2010).

Remerciements

Ce travail a été partiellement financé par OSEO dans le cadre du programme Quæro. Il s'inscrit également dans l'axe 5 du labex EFL (ANR/CGI).

Références

AMARDEILH, F., LAUBLET, P. et MINEL, J.-L. (2005). Document annotation and ontology population from linguistic extractions. *In Proceedings of the 3rd international conference on Knowledge capture (K-CAP '05)*, pages 161–168, New York, NY, USA. ACM.

ARONSON, A. R. et LANG, F.-M. (2010). An overview of metamap : historical perspective and recent advances. *JAMIA*, 17(3):229–236.

AW, A., ZHANG, M., XIAO, J. et SU, J. (2006). A phrase-based statistical model for sms text normalization. *In Proceedings of COLING-ACL '06 poster sessions*, pages 33–40.

BEAUFORT, R., ROEKHAUT, S., COUGNON, L.-A. et FAIRON, C. (2010). A hybrid rule/model-based finite-state framework for normalizing sms messages. *In ACL*, pages 770–779.

CHIANG, D. (2007). Hierarchical phrase-based translation. *Comput. Linguist.*, 33:201–228.

CIMIANO, P., HANDSCHUH, S. et STAAB, S. (2004). Towards the self-annotating web. *In Proceedings of WWW'04*, pages 462–471.

CIRAVEGNA, F. (2003). (lp) : Rule induction for information extraction using linguistic constraints. Rapport technique, Sheffield university.

COLLINS, M. (2002). Discriminative training methods for hidden markov models : theory and experiments with perceptron algorithms. *In Proceedings of EMNLP'02*, pages 1–8.

CUCERZAN, S. (2007). Large-scale named entity disambiguation based on wikipedia data. *In Proceedings of EMNLP-CoNLL'07*, pages 708–716.

CUNNINGHAM, H., MAYNARD, D., BONTCHEVA, K., TABLAN, V., ASWANI, N., ROBERTS, I., GORRELL, G., FUNK, A., ROBERTS, A., DAMLJANOVIC, D., HEITZ, T., GREENWOOD, M. A., SAGGION, H., PETRAK, J., LI, Y. et PETERS, W. (2011). *Text Processing with GATE (Version 6)*.

DILL, S., EIRON, N., GIBSON, D., GRUHL, D., GUHA, R., JHINGRAN, A., KANUNGO, T., RAJAGOPALAN, S., TOMKINS, A., TOMLIN, J. A. et ZIEN, J. Y. (2003). Semtag and seeker : bootstrapping the semantic web via automated semantic annotation. *In Proceedings of WWW '03*, pages 178–186.

ETZIONI, O., CAFARELLA, M., DOWNEY, D., KOK, S., POPESCU, A.-M., SHAKED, T., SODERLAND, S., WELD, D. S. et YATES, A. (2004). Web-scale information extraction in knowitall (preliminary results). *In Proceedings of WWW'04*, pages 100–110.

FINKEL, J. R. et MANNING, C. D. (2009). Nested named entity recognition. *In EMNLP '09*, pages 141–150.

FORT, K. et SAGOT, B. (2010). Influence of Pre-annotation on POS-tagged Corpus Development. *In ACL 4th Linguistic Annotation Workshop (LAW 2010)*, pages 56–63, Uppsala, Suède. Quaero (en partie).

KIRYAKOV, A., POPOV, B., TERZIEV, I., MANOV, D. et OGNYANOFF, D. (2004). Semantic annotation, indexing, and retrieval. *Journal of Web Semantics*, 2:49–79.

KOEHN, P., HOANG, H., BIRCH, A., CALLISON-BURCH, C., FEDERICO, M., BERTOLDI, N., COWAN, B., SHEN, W., MORAN, C., ZENS, R., DYER, C., BOJAR, O., CONSTANTIN, A. et HERBST, E. (2007). Moses : Open source toolkit for statistical machine translation. *In Proceedings of ACL'07*, pages 177–180.

KOEHN, P., OCH, F. J. et MARCU, D. (2003). Statistical phrase-based translation. *In HLT-NAACL*, pages 127–133.

LAFFERTY, J. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. *In Proceedings of the International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann.

LAVERGNE, T., ALLAUZEN, A., CREGO, J. M. et YVON, F. (2011). From n-gram-based to crf-based translation models. *In Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 542–553, Edinburgh, Scotland. Association for Computational Linguistics.

LAVERGNE, T., CAPPÉ, O. et YVON, F. (2010). Practical very large scale CRFs. *In Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics.

LIU, X., ZHANG, S., WEI, F. et ZHOU, M. (2011). Recognizing named entities in tweets. *In Proceedings of HLT '11*, pages 359–367.

LIVEMEMORIES (2010). Livememories : Second year scientific report. Rapport technique, LiveMemories.

MARCU, D. et WONG, W. (2002). A phrase-based, joint probability model for statistical machine translation. *In Proceedings of EMNLP'02*, pages 133–139.

MENDES, P. N., JAKOB, M., GARCÍA-SILVA, A. et BIZER, C. (2011). DBpedia Spotlight : Shedding light on the web of documents. *In Proceedings of I-Semantics'11*.

MIHALCEA, R. et CSOMAI, A. (2007). Wikify ! : linking documents to encyclopedic knowledge. *In Proceedings of CIKM'07*, pages 233–242.

MÜLLER, H., KENNY, E. E. et STERNBERG, P. W. (2004). Textpresso : An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2:309.

NADEAU, D. et SEKINE, S. (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26. Publisher : John Benjamins Publishing Company.

NAZARENKO, A., GUISSÉ, A., LÉVY, F., OMRANE, N. et SZULMAN, S. (2011). Integrating written policies in business rule management systems. *In Proceedings of RuleML'11*.

OCH, F. J. et NEY, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, pages 19–51.

RATINOV, L. et ROTH, D. (2009). Design challenges and misconceptions in named entity recognition. *In Proceedings of CoNLL'09*, pages 147–155.

SCHMID, H. (1995). Improvements in part-of-speech tagging with an application to german. *In Proceedings of the ACL SIGDAT'95-Workshop*.

STOLCKE, A. (2002). Srilm — an extensible language modeling toolkit. *In In Proceedings of ICSLP'02*, pages 901–904.

SUTTON, C. et MCCALLUM, A. (2006). *Introduction to Conditional Random Fields for Relational Learning*, chapitre 4, pages 93–128. MIT Press.

TOMEH, N. (2012). *Discriminative Alignment Models For Statistical Machine Translation*. Thèse de doctorat, University of Paris 11, Orsay.

UREN, V. S., CIMIANO, P., IRIA, J., HANDSCHUH, S., VARGAS-VERA, M., MOTTA, E. et CIRAVEGNA, F. (2006). Semantic annotation for knowledge management : Requirements and a survey of the state of the art. *J. Web Sem.*, 4(1):14–28.

WANG, Y. (2009). Annotating and recognising named entities in clinical notes. *In ACL/AFNLP (Student Workshop)*, pages 18–26.

WELTY, C. et IDE, N. (1999). Using the right tools : Enhancing retrieval from marked-up documents. *In Journal Computers and the Humanities*, pages 33–10.

WONG, Y. W. et MOONEY, R. J. (2006). Learning for semantic parsing with statistical machine translation. *In Proceedings of HLT-NAACL'06*, pages 439–446.

ZHOU, G. et SU, J. (2002). Named entity recognition using an hmm-based chunk tagger. *In Proceedings of ACL'02*, pages 473–480.