

Chaîne de traitement linguistique : du repérage d'expressions temporelles au peuplement d'une ontologie de tourisme

Stéphanie Weiser (1), Martin Coste (2) , Florence Amardeilh (1-2)

(1) MoDyCo – CNRS, Université Paris Ouest Nanterre La Défense –
200, avenue de la République, 92001 Nanterre.

sweiser@u-paris10.fr

(2) Mondeca – 3, cité Nollez, 75018 Paris.

martin.coste@mondeca.com

florence.amardeilh@mondeca.com

Résumé Cet article présente la chaîne de traitement linguistique réalisée pour la mise en place d'une plateforme touristique sur Internet. Les premières étapes de cette chaîne sont le repérage et l'annotation des expressions temporelles présentes dans des pages Web. Ces deux tâches sont effectuées à l'aide de patrons linguistiques. Elles soulèvent de nombreux questionnements auxquels nous tentons de répondre, notamment au sujet de la définition des informations à extraire, du format d'annotation et des contraintes. L'étape suivante consiste en l'exploitation des données annotées pour le peuplement d'une ontologie du tourisme. Nous présentons les règles d'acquisition nécessaires pour alimenter la base de connaissance du projet. Enfin, nous exposons une évaluation du système d'annotation. Cette évaluation permet de juger aussi bien le repérage des expressions temporelles que leur annotation.

Abstract This paper presents the linguistic data processing sequence built for a tourism web portal. The first steps of this sequence are the detection and the annotation of the temporal expressions found in the web pages. These tasks are performed using linguistic patterns. They lead to many questions which we try to answer, such as the definition of information to detect, annotation format and constraints. In the next step this annotated data is used to populate a tourism ontology. We present the acquisition rules which are necessary to enrich the portal knowledge base. Then we present an evaluation of our annotation system. This evaluation is able to judge the detection of the temporal expressions and their annotation.

Mots-clés : Annotation, expressions temporelles, ontologies, base de connaissance, tourisme

Keywords: Annotation, temporal expressions, ontologies, knowledge base, tourism

1 Introduction

Les travaux décrits dans cet article sont réalisés dans le cadre du projet RNTL EIFFEL. L'objectif global de ce projet est la mise en œuvre d'une plateforme logicielle permettant, autour d'une ontologie tourisme dédiée à un territoire, de sélectionner, classer et qualifier des contenus distribués sur le Web. Elle permet aussi de peupler automatiquement l'ontologie par de nouvelles instances et relations issus des contenus en s'assurant de la cohérence de la base de connaissance. Ceci permet de traiter des requêtes mettant en œuvre les instances de l'ontologie du territoire et les contenus Web indexés et d'offrir des fonctions de navigation dans la base de connaissance. Enfin, la plateforme a pour but d'assister l'utilisateur dans la construction d'un voyage à partir des ressources sélectionnées.

L'ontologie de tourisme que nous avons modélisée dans ce projet décrit les ressources touristiques, au sens large (activités, hébergements, logistique, patrimoine, territoires, voyages et itinéraires...) et organise la base de connaissance du territoire. Cette base de connaissance permet en effet de présenter les ressources touristiques dans leur contexte, de suggérer des voyages, itinéraires, séjours, activités aux utilisateurs. Elle permet au territoire de valoriser des ensembles cohérents de ressources, de créer des offres nouvelles et composites (circuit des peintres, itinéraires de l'aventure...) et de piloter le marketing du territoire par d'autres biais que le prix. Dans cette ontologie nous nous sommes particulièrement intéressés aux propriétés temporelles de ces ressources touristiques. En effet, il nous semblait important de pouvoir renseigner l'utilisateur sur les informations d'ouverture et de fermeture d'une offre touristique. Ces informations pouvant être obtenues à partir des contenus Web des ressources touristiques du territoire, c'est dans ce contexte que nous avons développé notre outil de repérage et d'annotation automatique des informations temporelles d'offres touristiques.

Dans la suite de cet article, nous exposerons tout d'abord un bref état de l'art concernant le domaine de travail et les méthodes utilisées. Puis nous présenterons l'outil de repérage et d'annotation des informations temporelles suivi de la plate-forme de peuplement d'ontologie grâce aux connaissances extraites automatiquement. Nous finirons cet article par l'évaluation de notre outil d'annotation avant de conclure et d'envisager les travaux futurs.

2 L'apport du Web Sémantique et des outils de TAL

Les données actuelles des contenus Web sont souvent encore écrites en langage naturel, car elles sont destinées aux humains. Le langage naturel étant par essence trop ambigu, des alternatives formelles et sémantiquement explicites doivent être mises en place pour lever ces ambiguïtés, aussi bien dans le contenu que dans ses annotations. Le Web Sémantique consiste à décrire les contenus documentaires en les annotant avec des informations non ambiguës afin de favoriser l'exploitation de ces contenus par des agents logiciels (Prié et Garlatti, 2004). Les ontologies, originaires des techniques de modélisation de la connaissance notamment développées en intelligence artificielle, fournissent les moyens d'exprimer les concepts d'un domaine en les organisant hiérarchiquement et en définissant leurs propriétés dans un langage de représentation des connaissances formel favorisant le partage d'une vue consensuelle sur ce domaine entre les applications informatiques qui en font usage (Bourigault et al., 2004).

L'exploitation des outils du Web Sémantique, et notamment des ontologies, pour la tâche d'enrichissement automatique de bases de connaissance est encore innovante. La mise en œuvre de ce peuplement passe le plus souvent par le traitement du langage naturel à partir des

attendues par l'utilisateur – et comment les annoter pour que celles-ci soient exploitables et intégrables à la base de connaissance.

Les expressions temporelles que l'on souhaite repérer et annoter ont une visée informative et pratique. Il ne s'agit pas de dates historiques ou d'expressions descriptives du type *la nuit d'avant* mais d'informations pratiques dans le domaine du tourisme. Il peut ainsi s'agir d'horaires d'ouverture, de dates, de périodes, etc. On peut classer ces expressions en deux catégories principales (Weiser et al., 2008) : les informations temporelles qui concernent un événement particulier et les informations temporelles répétitives. La première comprend des dates (*concert le 1^{er} octobre*), des périodes (*festival de mai à juin*), des heures (*le concert commence à 20h*). La seconde comprend des horaires (*le musée ouvre à 10h*), des périodes (*le restaurant est ouvert du lundi au samedi*) et des exceptions (*le camping est ouvert toute l'année sauf en janvier*). Des exemples d'une complexité plus grande peuvent également prendre place dans cette classification comme *de mai à juin, ouvert tous les jours sauf le mardi*. La complexité des expressions temporelles varie énormément : certaines expressions sont très simples, d'autres peuvent devenir complexes, jusqu'au point d'être floues ou ambiguës.

Une fois ces expressions repérées, elles doivent être stockées dans la base, contrainte par l'ontologie de tourisme, intégrant, pour les besoins du projet, une modélisation du « temps du tourisme ». L'ontologie constitue donc un pivot, une articulation dans ce projet et de nombreux ajustements ont été effectués pour pallier aux contraintes engendrées par cette modélisation. Par exemple, en ce qui concerne les jours et les heures, nous avons décidé de modéliser chaque jour de la semaine, et chaque journée est découpée en « parties de journée » ; on a donc *lundi matin, lundi midi, lundi après-midi et lundi soir*. Cette modélisation est indispensable, entre autres pour représenter les horaires d'ouverture d'un restaurant. Par soucis de simplicité et d'efficacité, le format XML a été choisi pour accomplir la tâche d'annotation. Un jeu de balises a donc été défini en fonction de ce qu'il était utile et possible d'annoter. Sa structure est calquée sur ce que l'ontologie modélise et une DTD⁴ a été établie afin de la fixer. Cette DTD est détaillée dans (Weiser, 2008) et on la présentera dans la partie suivante.

4 Annotations des informations temporelles

Le jeu de balises d'annotation et la DTD qui lui est associée ont été développés en concertation avec les différents partenaires du projet, le but étant que les balises collent aux données, c'est-à-dire qu'elles permettent d'annoter les données des pages Web le plus finement possible. Mais les données annotées doivent ensuite être intégrées à la base de connaissance. Il a donc fallu trouver un juste milieu entre ce que l'on pouvait annoter et ce que l'on pouvait exploiter. Les expressions temporelles sont donc annotées avec les balises : *période-ouverture*, *période-fermeture*, *exception* et *incertitude*. La balise *exception* permet d'annoter la chaîne textuelle décrivant une exception (comme *sauf le mardi*) et ainsi de garder l'information telle quelle afin de la fournir textuellement à l'utilisateur. La balise *incertitude* permet d'indiquer que le résultat n'est pas fiable : il est flou ou comprend une ambiguïté. De plus, une balise *description* reprend l'expression textuelle en entier.

En ce qui concerne les balises *période-ouverture* et *période-fermeture*, elles permettent de définir plus précisément l'expression repérée et peuvent inclure les balises *date*, *date-début*, *date-fin*, *jour*, *heure-début*, *heure-fin* et *partie-de-journée*. La balise *date* sert à annoter les dates seules ; dans la base de connaissance, on considérera que la date de fin est alors la

⁴

Une DTD (Document Type Definition) permet de décrire un modèle de document XML.

même que la date de début. La balise *jour* permet d'annoter les jours de la semaine tandis que *heure-début* et *heure-fin* annotent les heures et que *partie-de-journée* annote les informations du type matin et après-midi. À terme, dans la base de connaissance, toutes les informations d'horaires seront converties en parties de journées.

Cette DTD permet une grande liberté, nécessaire car les informations temporelles présentes dans les pages Web peuvent apparaître dans n'importe quel ordre. Notre outil permet ainsi de repérer et d'annoter un grand nombre d'expressions temporelles liées au domaine du tourisme. Certaines expressions ne sont malgré tout pas repérées à l'heure actuelle. Par exemple, une expression temporelle donnée peut apparaître tantôt au sujet d'un objet touristique et tantôt dans un contexte tout à fait différent. Des dates seules, sans contexte direct se trouvent parfois dans une page Web, au sujet d'un objet touristique mais le plus souvent elles n'ont aucun lien avec le tourisme et il serait donc trop « risqué » de les prendre en compte. Nous avons donc choisi de ne pas les repérer. D'autres expressions ne sont pas prises en compte alors qu'elles devraient l'être. Il s'agit alors en général d'un manque de richesse au niveau des marqueurs lexicaux et les graphes pourront encore être enrichis. On trouve par exemple dans une page l'expression *Réception dès 6 h 00* mais le marqueur « réception » n'apparaît pas dans nos graphes et l'expression est donc manquée.

5 Le peuplement d'ontologie

Bien que faisant déjà l'objet de nombreuses recherches, cette tâche de peuplement d'ontologie reste un véritable challenge, pas seulement dans le domaine du tourisme ou des informations temporelles mais quel que soit le domaine étudié. En effet, le rôle des humains demeure bien souvent irremplaçable et l'automatisation reste un des plus grands besoins pour de tels outils, particulièrement lorsqu'il s'agit d'annoter de grandes collections de documents (Uren et al., 2006). La plupart de ces outils ont évolué vers des environnements de plus en plus automatisés grâce aux méthodes issues des champs de l'Extraction d'Information et de l'Apprentissage Automatique (Corcho, 2006). Parmi les approches les plus abouties, citons les outils OntoSophie (Valarakos et al., 2004), KIM (Kiryakov et al., 2005) et OntoPop (Amardeilh, 2007). Pour plus de détails sur les plateformes existantes, nous référons le lecteur aux études (Uren et al., 2006) et (Reeve et Han, 2005).

Dans le cadre de notre projet, nous avons décidé d'utiliser l'outil CA Manager, nouvelle version de la plateforme OntoPop et développé dans le cadre du projet de recherche européen TAO⁵. La principale différence entre les outils précédemment cités et le CA Manager repose sur le fait que ce dernier préserve l'indépendance entre les outils de TAL et la base de connaissance utile à la tâche de peuplement. Pour ce faire, il exploite à la fois l'infrastructure modulaire proposée par UIMA⁶ et les langages et techniques proposés par la communauté Web Sémantique. Ceci lui procure une plus grande flexibilité et une capacité d'adaptation aux différents besoins, comme cela est souvent demandé pour les applications industrielles. Enfin, contrairement aux autres approches, il est aussi capable de contrôler la qualité et la validité des résultats de l'extraction d'information par rapport à une ontologie donnée, de les confronter avec d'autres ressources existantes (internes ou externes) et de les enrichir.

⁵ Transitioning Applications to Ontologies (TAO) du EU Sixth Framework Program (FP6-026460).

⁶ Unstructured Information Management Architecture (<http://www.alphaworks.ibm.com/tech/uima>) : APIs open source pour la construction de flux de traitement naturel du langage proposé par IBM.

5.1 Les étapes de la chaîne de peuplement des informations temporelles

Le CA Manager agit donc comme un médiateur entre des outils d'analyse de contenus, notamment linguistiques, et des bases de données sémantiques (comme celle proposée par ITM⁷ dans ce projet) et ce, quelle que soit l'ontologie du domaine. Il se compose d'un module générique permettant d'organiser la chaîne de peuplement d'ontologie, illustrée dans la Figure 2, en fonction de l'extraction de connaissances, de la consolidation des résultats issus de l'extraction, et du stockage des instances extraites dans des bases de connaissance.

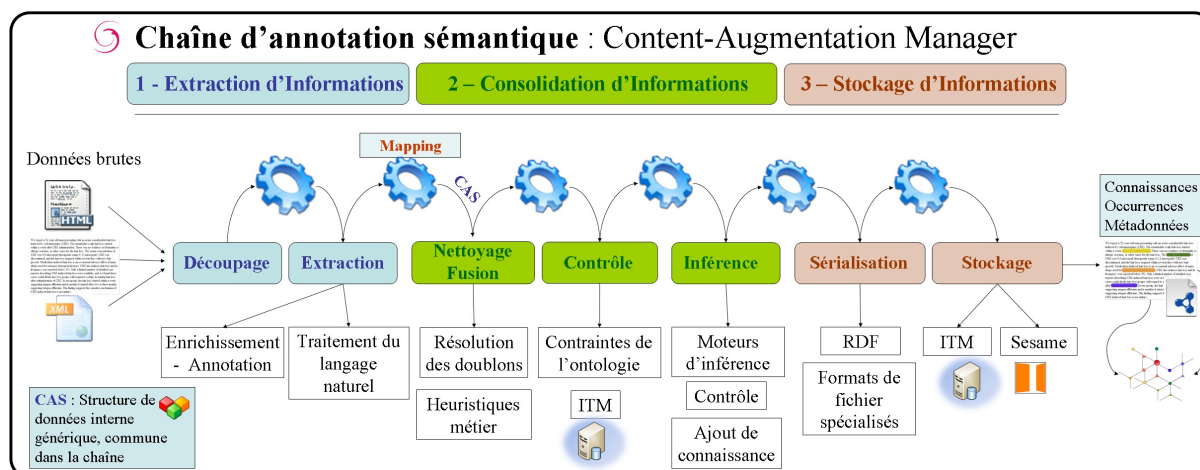


Figure 2 : Étapes de la chaîne d'annotation sémantique CA-Manager

Nous avons donc paramétré la chaîne de traitement du CA Manager pour utiliser l'outil d'annotation des informations temporelles d'une part et la base de données sémantique ITM dans laquelle nous avons chargé à la fois le modèle (classes et propriétés) et les instances de l'ontologie touristique du territoire. Comme vu précédemment, l'analyse par transducteurs (Unitex) produit un fichier XML comportant les annotations temporelles. A travers l'étape d'extraction, la chaîne propose le moyen d'intégrer les annotations ainsi produites à l'aide de règles d'acquisition de connaissances formulées manuellement et détaillées dans la partie suivante. Vient ensuite la phase de consolidation, permettant de confronter à l'ontologie du domaine les instances susceptibles d'y être ajoutées, et d'améliorer la qualité des informations. La consolidation comporte l'étape de fusion qui consiste principalement en la détection des doublons. En effet, on ne veut pas ajouter à une offre touristique existante des informations temporelles qui s'y trouvent déjà. L'étape suivante est celle de contrôle dans l'ontologie qui permet de préserver l'intégrité de la base de connaissance, et doit être effectuée avant la création des instances dans le référentiel. Chaque algorithme de consolidation du CA Manager prend en compte deux axes : la ressource ontologique concernée (instance de classe ou valeur de propriété) d'une part et les contraintes à contrôler (restrictions de domaine et de couverture, cardinalités des propriétés) d'autre part. Pour plus de détails sur ces algorithmes de consolidation, voir (Amardeilh, 2008). Enfin, la dernière étape permet de sérialiser les informations dans le format accepté par la base de connaissance utilisée et de les y stocker, ce qui correspond au peuplement d'ontologie.

5.2 Définition des règles d'acquisition des connaissances

Le processus d'acquisition des connaissances de l'étape d'extraction, aussi appelé « mapping », permet de localiser les informations annotées dans le fichier XML, et de définir la manière dont ces informations seront ajoutées dans la base de connaissance. Ce mapping

⁷

Intelligent Topic Manager (ITM), outil de gestion de base de connaissance proposé par Mondeca

est créé manuellement à partir de la sortie XML de l'outil d'annotation et des concepts modélisés dans l'ontologie du domaine. Il définit les classes d'entités et leurs propriétés, en déclarant les balises XML des annotations qui les renseignent. Chaque entité décrite donnera lieu à la création d'une instance de classe. Elle possède donc un identifiant unique généré à la volée selon les standards du Web Sémantique (URI), ainsi que sa classe d'appartenance. Puis les propriétés de chacune de ces entités sont également définies en fonction du modèle de l'ontologie et de la localisation de sa valeur dans le XML d'annotation linguistique.

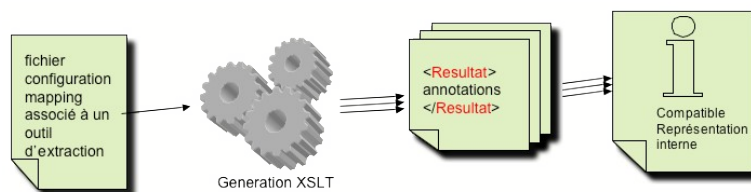


Figure 3 : Mécanisme d'acquisition de connaissances

Ce mapping est implémenté sous la forme d'un ensemble de règles d'acquisition de connaissance (RAC) tel que décrit dans (Amardeilh, 2007). Il prend la forme d'un fichier de configuration XML déclarant ces règles. L'exécution de ce fichier permet de générer automatiquement une feuille de style XSLT, transformant les annotations XML dans un format compatible avec la base de connaissance comme RDF ou OWL (cf. Figure 3).

Dans le cadre de notre étude, nous avons créé des règles d'acquisition de connaissance spécifiques à l'analyse d'informations temporelles. Pour illustrer le mécanisme de mapping et la déclaration des règles d'acquisition de connaissance, reprenons l'exemple d'annotation d'horaires présenté dans la partie 2 :

```
<UT> <periode_Ouverture> Horaires d'ouverture <jour> Lundi</jour> et<jour> mercredi</jour> <heure_debut> 9h</heure_debut>-<heure_fin>12h</heure_fin><jour> Jeudi</jour><heure_debut> 14h </heure_debut>- <heure_fin>17h </heure_fin> <jour> Vendredi</jour><heure_debut> 14h</heure_debut>-<heure_fin>16h </heure_fin> </periode_Ouverture> </UT> <description>"Horaires D'ouverture Lundi et mercredi 9h-12h Jeudi 14h-17h Vendredi 14h-16h"</description>
```

```
<classMapping id="1">
  <types>
    <type ispath="false">http://www.projet-eiffel.org/ontology#PeriodeOuverture</type></types>
  <expressions>
    <expression> <xpath>Content/Text/UT/periode_Ouverture</xpath> </expression>
  </expressions>
  <propertyTypeMappings>
    <propertyTypeMapping id="22">
      <dataType>label</dataType>
      <types> <type>eiffel:jour</type> </types>
      <expressions> <expression> <xpath>jour</xpath> </expression> </expressions>
    </propertyTypeMapping>
    <propertyTypeMapping id="23">
      <dataType>label</dataType>
      <types> <type>eiffel:heure_debut</type> </types>
      <expressions> <expression> <xpath>heure_debut</xpath> </expression> </expressions>
    </propertyTypeMapping>
    <propertyTypeMapping id="24">
      <dataType>label</dataType>
      <types> <type>eiffel:heure_fin</type> </types>
      <expressions> <expression> <xpath>heure_fin</xpath> </expression> </expressions>
    </propertyTypeMapping>
  </propertyTypeMappings>
</classMapping>
```

Figure 4 : Fichier de mapping

Le fichier de mapping (cf. Figure 4) permet de créer une instance de la classe « Période d'ouverture », elle-même sous-classe de « Unité de Temps (UT) » dans l'ontologie de tourisme. Cette instance est renseignée par le contenu du fichier XML d'annotation, par le biais des chemins Xpath déclarés. Ainsi le libellé de cette instance correspondra à la valeur de la balise « periode_Ouverture » du fichier XML d'annotation. Les propriétés d'une période d'ouverture, i.e. les jours, les heures de début et de fin, sont également décrites dans le

mapping et prendront respectivement comme valeur celles des balises « jour », « heure_debut » et « heure_fin ».

6 Évaluation du système

Dans cette partie, nous présentons une évaluation de notre système de repérage et d'annotation, en termes de rappel et précision. Pour plus de détails sur les questions d'évaluation, voir (Popescu-Belis, 2007). Si ces mesures de rappel et précision sont souvent utilisées en TAL, elles sont aussi parfois controversées, que ce soit dans le domaine de l'extraction d'information (Lavelli et al., 2004) ou dans celui de l'annotation sémantique (Maynard, 2005). En effet ces mesures sont trop rigides et ne permettent pas de rendre compte de résultats imparfaits ou partiels. Certaines solutions ont été envisagées mais elles n'ont pas donné lieu à des mesures standardisées tel le rappel et la précision que nous utilisons. Par exemple, (Freitag, 1998) avait proposé de classer les résultats dans trois catégories : résultats exacts ; résultats dans lesquels l'information attendue est contenue dans le résultat obtenu ; résultats imbriqués où l'information obtenue dépasse l'information attendue.

6.1 Protocole d'évaluation

Pour procéder à une évaluation de notre système de repérage et d'annotation, nous lui avons donné un échantillon de 250 pages sélectionnées au hasard à traiter. Notre évaluation a pour but d'évaluer aussi bien le repérage que l'annotation. Au niveau des taux de rappel et précision, on considère comme correctes les expressions repérées à bon escient, même si elles le sont de façon incomplète ; en effet un repérage incomplet n'est pas toujours nuisible pour le système (par exemple, pour *ouvert toute l'année le vendredi et le samedi*, si on ne repère pas la première partie, *ouvert toute l'année*, cela ne change rien au résultat final). On comptabilise les expressions manquées et les expressions repérées à tort. Au niveau de l'annotation, on considère comme bien annotées les expressions qui sont bien catégorisées (où on ne confond pas ouverture et fermeture par exemple).

Une fois les pages analysées par le système, nous distinguons celles dans lesquelles un repérage (et donc une annotation) a été effectué de celles dans lesquelles il n'y a pas de résultat. Pour celles qui ne donnent pas de résultat, nous vérifions si elles ne contiennent pas tout de même des expressions temporelles touristiques que l'on aurait dû repérer, afin de comptabiliser les expressions manquées. Les pages dans lesquelles un repérage et une annotation ont eu lieu sont également analysées manuellement. On cherche si elles contiennent des expressions manquées et, pour les expressions repérées et annotées, on vérifie si le repérage est effectué à bon escient, si l'expression est repérée dans son ensemble et si elle est bien annotée.

6.2 Résultats

Dans les 250 pages, 191 expressions sont dans un premier temps considérées comme à repérer⁸. Notre système repère 115 expressions : 67 expressions sont repérées à bon escient et 48 expressions sont repérées à tort. 124 expressions temporelles touristiques sont donc manquées par le système. Sur les 67 expressions repérées, 44 sont bien repérées et 23 sont repérées partiellement. Ces chiffres mènent aux taux de rappel et précision présentés dans la

⁸ Pour arriver à ce chiffre (191) on additionne les expressions repérées à bon escient et les expressions manquées et on soustrait les expressions repérées à tort.

première colonne du Tableau 1. En qui concerne l'annotation, sur les 67 expressions repérées, 64 sont bien annotées et 3 comportent des erreurs d'annotation ; cela donne un pourcentage de 95,5 % d'annotations correctes. Certaines pages ayant des propriétés particulières, nous allons faire un bilan selon le nombre d'expressions par page et selon le nombre de pages. Sur les 250 pages analysées, 61 seulement contiennent des expressions temporelles touristiques. De nombreuses expressions temporelles non touristiques se trouvent aussi dans ces pages et mènent parfois à des repérages fautifs mais beaucoup ne sont pas repérées – à raison.

Parmi les 61 pages contenant des expressions à repérer, 18 pages ne sont pas annotées, ce qui mène à 51 expressions manquées. 23 pages sont traitées semi-correctement, c'est-à-dire qu'elles peuvent contenir des expressions bien repérées et bien annotées mais elles contiennent également des expressions repérées partiellement ou des expressions non repérées (manquées). Dans ces pages, 43 expressions sont repérées : 20 sont bien repérées et 23 sont repérées de façon incomplètes. De plus, 73 expressions sont manquées. 20 pages sont traitées correctement, pour un total de 24 expressions bien repérées et bien annotées. Nous remarquons que certaines pages sont à l'origine de beaucoup d'erreurs.

Premièrement, en dehors de ces 61 pages, des expressions repérées à tort sont présentes dans 6 pages, comprenant donc 48 faux-positifs. Certaines de ces expressions apparaissent dans des pages non touristiques ; elles n'auraient donc pas dû se trouver à l'entrée de notre système et relèvent d'une erreur du module d'aspiration des pages. Nous décidons donc d'exclure ces pages (au total 3 pages contenant 41 expressions repérées sont exclues). Les taux ainsi obtenus sont présentés dans la deuxième colonne du Tableau 1.

Deuxièmement, nous nous sommes intéressés au grand nombre d'expressions manquées par notre système. Une même page contient par exemple 49 expressions manquées (sur 124 expressions manquées au total). Cette page présente un agenda. Elle ne peut pas être considérée comme non pertinente, car il s'agit tout de même d'un agenda touristique. En revanche, on peut avancer qu'à l'heure actuelle notre système n'est pas conçu pour analyser de telles pages. Il faudrait donc les repérer et en faire un traitement propre, prenant en compte le fait qu'il s'agit d'un agenda et que la page contient donc un grand nombre d'expressions temporelles. De plus les expressions temporelles contenues dans des pages-agenda ont souvent des formes très différentes de ce que l'on considère comme des « expressions temporelles touristiques ». Nous choisissons donc d'éliminer pour l'instant les pages-agenda de notre évaluation. Une fois ces pages exclues, on obtient les taux de rappel et précision de la troisième colonne du Tableau 1. Ces taux, bien meilleurs que les premiers taux bruts calculés, sont plus représentatifs des capacités de notre outil car nous avons éliminé quelques pages marginales qui avaient beaucoup d'incidence sur le rappel et la précision.

Taux bruts		Taux après exclusion des pages non pertinentes		Taux après élimination des pages-agenda	
Rappel	Précision	Rappel	Précision	Rappel	Précision
67 / 191 35 %	67 / 115 58,2 %	67 / 191 35 %	67 / 74 90,5 %	58 / 95 61 %	58 / 65 89,2 %

Tableau 1 : Taux de rappel et précision

7 Conclusion et perspectives

Nous avons présenté la chaîne de traitement linguistique développée pour un portail touristique sur Internet. En entrée de la chaîne sont fournies des pages Web touristiques

converties en XML. Nous analysons automatiquement ces pages afin d'y repérer et d'y annoter les expressions temporelles liées aux objets touristiques. Le format d'annotation a été développé en fonction de l'ontologie de tourisme du projet. Nous avons ensuite présenté le mécanisme des règles d'acquisition de connaissances permettant de peupler l'ontologie à l'aide des informations annotées. Après avoir passé en revue les différents problèmes de modélisation et d'interaction entre nos modules, nous avons proposé une évaluation du module de repérage et d'annotation des expressions temporelles. Nous nous sommes appuyés sur les mesures de rappel et précision largement utilisées en TAL mais nous pourrions maintenant essayer de mettre au point d'autres mesures plus appropriées, permettant de juger également des résultats partiels.

Outre le fait de continuer nos travaux sur le repérage proprement dit qui consiste à enrichir encore les données linguistiques contenues dans les transducteurs afin d'élargir le nombre d'expressions repérées, les règles d'acquisition des connaissances devront être complétées et évaluées. Nous sommes aussi en train de travailler, avec un autre partenaire du projet, à l'élaboration d'un moteur de raisonnement qui permettrait de calculer automatiquement les périodes d'ouverture en fonction des instances des périodes de fermeture créées dans la base de connaissance à partir des annotations.

Remerciements

Ce travail a été partiellement financé par le projet Eiffel ANR-05-RNTL-007.

Références

AMARDEILH F. (2007). *Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle*. Thèse de doctorat. Université Paris-Sorbonne.

AMARDEILH F. (2008). Semantic Annotation & Ontology Population. In J. CARDOSO & M. D. LYTRAS Eds. *Semantic Web Engineering in the Knowledge Society*, Idea Group Publishing.

BATTISTELLI D., MINEL J.-L., SCHWER S. (2006). Représentation des expressions calendaires dans les textes : une application à la lecture assistée de biographies. *Traitement Automatique des Langues* 47, 3, 1-26.

BOURIGAUT D., AUSSÉNAC-GILLES N., CHARLET J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas, In *Techniques Informatiques et Structuration de Terminologies*, PIERREL J.-M. ET SŁODZIAN M. (Eds.), Numéro Spécial de la Revue d'Intelligence Artificielle (RIA), 18(1), Hermès, Paris, 87-110.

CORCHO O. (2006). Ontology based document annotation: trends and open research problems. In *Int. J. Metadata, Semantics and Ontologies*, 1(1), Inderscience. 47-57.

FREITAG D. (1998). *Machine Learning for Information Extraction in Informal Domains*. Thèse de doctorat, Université Carnegie Mellon.

KIRYAKOV A., POPOV B., TERZIEV I., MANOV D., KIRILOV A., GORANOV M. (2005). Semantic annotation, indexing, and retrieval. In *J. Web Semantics, Science, Services and Agents on the WWW*, 2(1), Elsevier, 49-79.

CTL : du repérage d'expressions temporelles au peuplement d'une ontologie de tourisme

HEARST M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *International Conference on Computational Linguistics (COLING'92)*.

LAVELLI A., CALIFF M. E., CIRAVEGNA F., FREITAG D., GIULIANO C., KUSHMERICK N., ROMANO L. (2004). IE evaluation: Criticisms and recommendations, In *Proceedings of the AAAI 2004 Workshop on Adaptive Text Extraction and Mining (ATEM 2004)*.

MAYNARD D. (2005). Benchmarking ontology-based annotation tools for the Semantic Web, in *Proceedings of the Workshop "Text Mining, e-Research and Grid-enabled Language Technology" in the UK e-Science Programme All Hands Meeting (AHM2005)*.

MORIN E. (1999). Acquisition de patrons lexicosyntaxiques caractéristiques d'une relation sémantique. *Traitement Automatique des Langues* 40, 1, 143-166.

POPESCU-BELIS A. (2007). Le rôle des métriques d'évaluation dans le processus de recherche en TAL. *Traitement Automatique des Langues* 48, 1, 67-91.

PRIÉ Y., GARLATTI S. (2004). Méta-données et annotations dans le Web sémantique, in *Le Web sémantique*, CHARLET J., LAUBLET P. et REYNAUD C. (Ed.), Hors série de la *Revue Information - Interaction - Intelligence (I3)*, 4(1), Cépaduès, 45-68.

REEVE L., HAN H. (2005). Survey of semantic annotation platforms. In *Symposium on Applied Computing (SAC'2005)*, 1634-1638.

UREN V., CIMIANO P., HANDSCHUH S., VARGAS-VERA M., MOTTA E., CIRAVEGNA F. (2006). Semantic annotation for knowledge management: requirements and a survey of the state of the art. In *J. Web Semantics, Science, Services and Agents on the WWW*, 4(1). 14-26.

STERN R.-D. (2007). *Expression linguistique du temps et représentation ontologique : OWL-Time et étude des adverbiaux temporels*. Mémoire de master. Université Paris-Sorbonne.

VALARAKOS A., PALIOURAS G., KARKALETSIS V., VOUROIS G. (2004). Enhancing the Ontological Knowledge through Ontology Population and Enrichment. In *14th Int. Conf. on Knowledge Engineering and Knowledge Management (EKAW'04), Lecture Notes in Artificial Intelligence*, Vol. 3257, Springer-Verlag, 144-156.

WEISER S. (2008). Informations spatio-temporelles et objets touristiques dans des pages Web : repérage et annotation. Actes de *Recital 2008*, 131-140.

WEISER S., LAUBLET P., MINEL J.-L. (2008). Automatic identification of temporal information in tourism web pages. Actes de *LREC'08, the Sixth International Language Resources and Evaluation*, 127-131.