Identification automatique des valeurs temporelles dans les textes

Marie Chagnoux, Slim Ben Hazez et Jean-Pierre Desclés

LaLICC – UMR 8139 du CNRS - Université de Paris 4 -Sorbonne
96, Boulevard Raspail, 75006 Paris

Marie.Chagnoux@paris4.sorbonne.fr, Slim.Ben-hazez@paris4.sorbonne.fr,
Jean-Pierre.Descles@paris4.sorbonne.fr

Keywords – Mots Clés

Traitement aspectuo-temporel, temps, temporalité, exploration contextuelle, analyse sémantique de surface, ressources linguistiques.

Aspecto-temporal processing, time, temporality, contextual exploration, semantic analysis, linguistic resources.

Résumé – Abstract

Cet article présente une application qui associe un certain nombre de valeurs sémantiques à des segments textuels en vue de proposer un traitement automatique de la temporalité dans les textes. Il s'agit d'automatiser une analyse sémantique de surface à l'aide de règles heuristiques d'exploration contextuelle et d'une base organisée de marqueurs linguistiques.

This paper presents a system which associates semantic values with textual segments in order to propose an automatic processing of temporality in texts. The purpose is to automate a semantic analysis driven with contextual exploration heuristic rules and an organized data base of linguistic markers.

1 Introduction

A partir d'un texte décrivant des procès selon un ordre syntagmatique, tout lecteur est capable de rétablir une chronologie des événements décrits. Est-il possible d'automatiser ce traitement de la temporalité ? Quelles sont les connaissances linguistiques mises en œuvre ? Les marqueurs aspectuels et contextuels sont-ils suffisants ? L'application que nous présentons propose un certain traitement opératoire de la temporalité dans les textes qui permet de répondre à ces questions.

Cette application fournit deux types de traitement automatisé. Une première tâche assigne un certain nombre de valeurs sémantiques d'ordre temporel à chaque proposition du texte. Une seconde tâche permet ensuite, à partir de ces valeurs, de calculer les interrelations entre segments pour organiser les procès entre eux. Seule la première tâche sera ici décrite.

Dans un premier temps, nous exposerons le cadre théorique et la démarche linguistique mis en œuvre. Ensuite, nous montrerons comment les connaissances modélisées ont été implémentées à l'aide de l'environnement SemanText (Ben Hazez, 2002). Enfin, nous proposerons une première évaluation des résultats obtenus en montrant les limites et les problèmes que pose un tel traitement.

2 Cadre théorique du temps et de l'aspect

Le présent système s'inscrit d'une part dans le prolongement des recherches sur la catégorie aspectuelle et temporelle de J-P. Desclés et Z. Guentcheva (Desclés, 1980) appliquée à l'analyse du français et, d'autre part, dans la technique d'exploration contextuelle ayant conduit à un premier système informatique pour le traitement des temps du passé (Desclés et al, 1991). La description théorique fait appel à des notions aspectuelles fondamentales (état, événement,…), à des notions temporelles et modales (réalisé, non-réalisé, contrefactuel,…), ainsi qu'à des représentations graphiques sous forme de diagrammes temporels permettant la visualisation des relations aspectuo-temporelles associées à un texte.

L'une des difficultés du traitement de temps et de l'aspect vient de la multiplicité des valeurs sémantiques associées aux marqueurs grammaticaux. Ainsi, le passé composé peut, en français, renvoyer à au moins cinq valeurs différentes. Pour construire les valeurs sémantiques, il faut lever l'indétermination de chaque occurrence de tels marqueurs. L'hypothèse mise en œuvre par l'exploration contextuelle revient à rechercher des indices (adverbes, locutions…) dans le contexte, de façon à assigner une certaine valeur.

L'application développée s'appuie sur l'expérience tirée de ces précédents travaux théoriques et appliqués qui ont été complétés afin de proposer un traitement automatique systématique et opérationnel. Il nous a tout d'abord paru nécessaire d'enraciner plus profondément notre étude dans une prise en compte globale de l'énonciation. L'objectif était donc de concilier la dimension textuelle du texte avec une segmentation plus fine en propositions pour repérer localement des phénomènes temporels tout en en calculant leurs relations au sein d'une énonciation globale. Ainsi, le repérage de référentiels et d'une structure textuelle permet d'inscrire les segments dans la dynamique discursive.

De plus, en vue de réutilisations ultérieures par d'autres systèmes (représentation graphique, traduction automatique, etc.), il est apparu que les simples valeurs aspectuo-temporelles n'étaient pas les seuls supports de la sémantique temporelle et que d'autres éléments intervenaient. Sous le terme générique de *temporelles*[1], nous avons donc proposé d'autres catégories afin de fournir un maximum d'informations sémantiques :

- Des informations aspectuelle (état, événement, processus…)
- Des informations modales aspectuelles (réalisé, non réalisé, …)
- Des informations sur les systèmes référentiels (énonciatif, non-actualisé, possible éventuel, possible contrefactuel,…)
- Des informations de répétition (itération, habitude…)

[1] Dans la suite de l'article, *valeurs temporelles* désigne l'ensemble de ces valeurs aspectuelles, référentielles, temporelles, répétitives, etc.

3 Mise en oeuvre informatique : modélisation de l'application

Les connaissances linguistiques dégagées, règles heuristiques et classes d'indices linguistiques, ont été formalisées et implémentées avec SemantText. Cet environnement de développement de ressources linguistiques et de manipulation de textes est composé de ressources linguistiques organisées en plusieurs tâches, d'un formalisme déclaratif unifié qui permet de décrire des données linguistiques et des opérations textuelles, et d'outils logiciels de manipulation de données textuelles. Nous présentons dans cette seconde partie la modélisation dans Semantext qui permet, à partir des marqueurs linguistiques, d'étiqueter le texte avec des informations sémantiques.

3.1 Tâche et sous-tâches : l'architecture

Comme il a été présenté dans la première section, l'application va distinguer différents types de valeurs. Elle se divise donc en un ensemble de sous-tâches élémentaires présenté dans le graphe suivant :

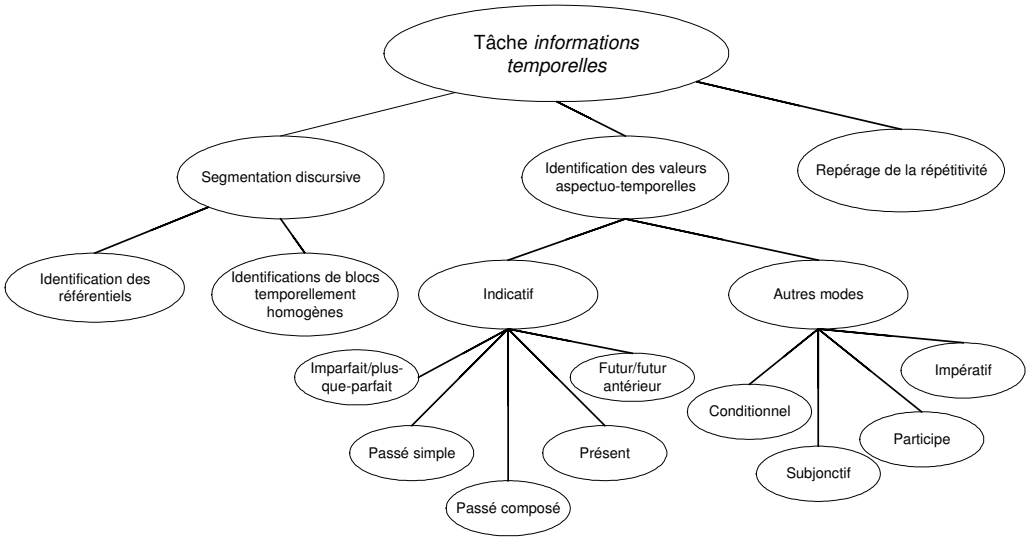

Figure 1 : Arborescence de la tâche de repérage des informations temporelles

Une telle architecture permet :

- d'avoir des tâches indépendantes qui peuvent être soit appliquées simultanément soit appliquées en cascade afin de réutiliser des connaissances obtenues de manière automatique par une première tâche. Ainsi la tâche segmentation discursive doit être achevée avant que les autres ne soient déclenchées, alors que les tâches « Futur » ou « Imparfait » peuvent être activées indépendamment l'une de l'autre.
- de partager ou non des ressources linguistiques : bases de règles et de marqueurs peuvent être hiérarchisées, certaines seront définies au niveau de la tâche parent, d'autres, plus spécifiques seront propres à chaque sous tâche. Des connaissances communes à plusieurs tâches seront donc implémentées au nœud supérieur dans l'arborescence. Ainsi, par exemple, les annotations morpho-syntaxiques sont implémentées à la racine du graphe.

3.2 Organisation des connaissances linguistiques : règles heuristiques et marqueurs

Une tâche élémentaire est composée de deux parties : un ensemble de classes d'indices linguistiques et un ensemble de règles qui permettent d'identifier dans un contexte bien déterminé certains indices déclencheurs et d'étiqueter des segments de texte.

Les indices linguistiques dénotent des séquences textuelles qui peuvent être des morphèmes, des signes de ponctuations, des mots, des locutions, des expressions complexes ou des annotations morpho-syntaxiques. Ces indices sont regroupés dans des classes hiérarchisées exprimées par des motifs linguistiques nommés. Le principe utilisé pour construire l'ensemble des motifs linguistiques de l'application consiste à définir des classes élémentaires et à les réutiliser pour former des classes d'indices de plus en plus complexes. Ainsi un motif comme `&années`, contenant toutes les années du calendrier, sera utilisé pour former d'autres classes d'expressions temporelles plus complexes telle que `&Duree` ou `&Date` qui repéreront respectivement des expressions comme *de 1955 à 1961*ou *le 2 janvier 1759*

Les règles sont définies par le triplet `<Déclencheur, Condition, Action>`. `<Déclencheur>` désigne l'indice (ou le motif) déclencheur de la règle. `<Condition>` désigne le contexte du déclencheur qui conditionne la partie *action*. Une condition permet d'exprimer des restrictions contextuelles. `<Action>` désigne les actions à exécuter par la règle si la partie condition est satisfaite et donc à attribuer une valeur à la proposition contenant l'occurrence du motif déclencheur. Ainsi la règle suivante `<&PC, ($proposition$ contains &Date), {$event=$proposition$}>` peut s'interpréter comme « en présence d'un passé composé, si la proposition courante contient un élément de la classe des motifs `&Date`, alors étiqueter cette proposition comme un événement ».

Pour une tâche donnée, le processus d'interprétation des règles consiste dans un premier temps à analyser le texte en une seule passe pour identifier l'ensemble des occurrences des motifs déclencheurs. Cela permet de compiler tous les motifs linguistiques sous forme d'automates à états finis déterministes avec une priorité accordée aux motifs les plus longs.

Ensuite, le moteur de règles coordonne l'exécution des règles suite aux occurrences reconnues des indices déclencheurs. Pour chaque occurrence, il détermine ensuite l'ensemble des règles candidates à l'exécution puis sélectionne une règle qui est alors activée. L'exécution d'une règle comporte l'évaluation de la partie *condition*, puis l'exécution de l'action si la condition est satisfaite.

Pour chaque tâche élémentaire, les règles sont organisées selon un ordre prioritaire. En effet, dans SemanText, elles sont indexées par la notion de déclencheur et peuvent être examinées dans un ordre choisi. L'organisation des règles permet ainsi de considérer certaines règles comme plus puissantes que d'autres. La stratégie consiste donc à hiérarchiser l'ordre d'exécution des règles et à empêcher une règle de s'appliquer sur la même occurrence d'un indice déclencheur lorsqu'une autre règle s'est déjà appliquée. Enfin, l'ordre de priorité impliqué par cette organisation offre la possibilité d'attribuer une valeur par défaut : si aucun indice complémentaire n'est repéré et que, par conséquent, aucune condition n'est satisfaite pour attribuer une valeur au segment textuel, la dernière règle implémentée attribuera une valeur à tous les segments non étiquetés. Chaque tâche est donc construite par une organisation motivée qui garantie l'étiquetage de chaque segment.

4 Expérimentation

Le système actuel comprend 950 marqueurs linguistiques et 105 règles. Certaines sous-tâches sont encore en cours de construction et ne sont donc pas encore complètement opérationnelles. Les connaissances linguistiques ont été acquises à partir des travaux précédemment cités complétés à l'aide d'un corpus regroupant des articles de journaux (61 textes), des articles en ligne sur Internet (35 textes), des romans (5 textes), 2 pièces de théâtre et des extraits divers. Le système a été ensuite évalué plusieurs fois sur des textes n'ayant pas servi à l'expérimentation. Les textes ont été alors incorporés au corpus d'acquisition et les résultats obtenus ont permis de compléter ou de préciser certaines données linguistiques.

Il aurait été nécessaire pour l'évaluation de ce système d'une part de pouvoir le comparer à d'autres systèmes existants, d'autre part, de quantifier et qualifier les résultats obtenus. Bien que certains travaux s'inscrivent dans la même problématique (Gosselin, 1996), (Moeschler et al, 2000), il n'existe pas, à notre connaissance, actuellement de système opérationnel proposant un traitement similaire. Notre évaluation porte donc uniquement sur les résultats obtenus. Les premières évaluations du système ont fait ressortir les points suivants.

Premièrement, la performance du système est grandement tributaire du traitement en amont et principalement de la reconnaissance automatique des temps morphologiques. Un mauvais étiquetage morpho-syntaxique, un participe passé reconnu comme un verbe conjugué par exemple, entraîne une segmentation erronée, ce qui, au niveau de notre système se traduit par une réponse incorrecte.

Deuxièmement, bien qu'un corpus diversifié ait été utilisé lors de la période d'acquisition des connaissances, certaines informations sémantiques ne sont pas traitées, soit pour une question de marginalité (effet de style ou construction peu usuelle), soit pour des questions d'ambiguïté. Un traitement automatique dont l'objectif est de lever l'ambiguïté proposera parfois une solution unique là où un traitement manuel en considérerait plusieurs ou aucune.

Troisièmement, les ressources du système sont parfois incomplètes. Le besoin de ressources lexicales est indéniable : notre système traite une polysémie essentiellement grammaticale, mais l'existence de ressources lexicales permettraient d'étendre sensiblement le taux de couverture du traitement. Pour l'instant, les classes d'indices ont été complétées au fur et à mesure de l'étude des textes de manière à classer les verbes (classes de verbes statiques, dynamiques,...) mais ceci demeure insuffisant. De plus, la complexité de certains passages temporels dans les textes oblige à examiner en détail les combinatoire d'indices. On peut donc relever trois types d'erreurs : les erreurs liées aux traitements en amont et qui peuvent être parfois corrigés par quelques règles, les erreurs jugées acceptables car liées à une indétermination du texte et les erreurs liées au système qui demandent une révision des connaissances linguistiques mises en place.

Enfin, il est nécessaire de signaler le problème posé par les verbes à l'infinitif, les propositions nominales et les phénomènes d'ellipse verbale : il a été décidé que la segmentation en propositions s'articulait autour d'un verbe conjugué. Cette prise de position nécessaire pour un premier prototype empêche de prendre en considération un nombre d'éléments important pour l'analyse temporel. Ainsi un extrait comme :*L' autre jour, vous m' avez attrapé pour un journal, hier soir pour un autre, les deux fois en pas grand temps*[2] ne pourra être traité de manière adéquate par le système.

---

[2] Extrait du *Journal (1887-1910*) de Jules Renard, Ed L. Guichard et G. Sigaux, Gallimard, p209

5 Conclusion

Le système est actuellement en cours d'élaboration dans le cadre d'une thèse. Au fur et à mesures de l'implémentation, les bases de données linguistiques sont élargies de manière à fournir une couverture plus large des phénomènes textuels. D'autres protocoles d'évaluation et d'enrichissements sont actuellement à l'étude pour permettre de mieux cerner les limites des bases de règles et améliorer les tâches.

L'un des objectifs de cette application est d'intégrer une chaîne de traitement automatique plus complexe. Cette intégration est par exemple menée dans le cadre du projet OLETT[3] où elle permet le filtrage sémantique automatique des informations temporelles. Il serait également souhaitable de pouvoir l'inclure dans une tâche de traduction automatique afin d'évaluer les résultats fournis par le système.

Enfin, il est important de rappeler que l'application contient une autre tâche, non présentée ici, qui calcule les relations temporelles entre les segments étiquetés et que, à l'avenir, ce système sera couplé avec un module de génération automatique de diagrammes temporels en cours de réalisation. Cela permettra, à partir d'un texte brut de générer une représentation graphique utilisable à des fins théoriques et didactiques.

Références

Ben Hazez S. (2002), *Un modèle d'exploration contextuelle des textes : filtrage et structuration d'informations textuelles, modélisation et réalisation informatique (système SemanText)*, thèse de doctorat, Université Paris IV-Sorbonne, Paris.

Borillo A. (1996), Le déroulement temporel et sa représentation spatiale en français, *Cahiers de praxématique* 27, p.109-128.

Culioli A. (2002), *Variations sur la linguistique, entretiens avec Frédéric Frau*, Klincksieck, Paris.

Desclés J-P. (1980), Construction formelle de la catégorie grammaticale de l'aspect (Essai) *Notion d'aspect*, Klincksieck, Paris, J David et R Martin (éds), pp. 198-237.

Desclés J-P., Jouis C., Oh H-G., Reppert D. (1991), *Exploration Contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte*. In Knowledge modeling andexpertise transfer, Amsterdam, D. Herin-Aime, R. Dieng, J-P. Regourd, J-.P. Angoujard (éds), pp.371-400.

Gosselin L. (1996), *Sémantique de la temporalité en français*, Duculot, Louvain-la-Neuve.

Moeschler J. et al (2000), *Inférences directionnelles, représentations mentales et subjectivit*é, Cahiers de Linguistique Française 22, Université de Genève, Genève.

---

[3] OLETT est un projet soutenu par le programme interdisciplinaire du CNRS *Société de l'information* et porte sur *L'identification des événement et des lieux pour l'organisation aspectuo-temporelles sous-jacente aux textes (Application : filtrage sémantique automatique et recherche d'informations sur le Web)*.