Contrôle dynamique multicritère des résultats d'une chaîne de TAL

Grégory Smits

France Télécom division R&D, TECH/EASY/LN
gregory.smits@francetelecom.com

Résumé

Le traitement linguistique d'un énoncé écrit conduit le plus souvent à la prise en compte d'interprétations concurrentes, ainsi qu'à la création d'ambiguïtés artificielles. Le contrôle de ces points d'embarras est indispensable pour garantir une efficacité et une précision convenable du processus d'analyse. L'approche décrite dans ce document exploite le paradigme de l'aide multicritère à la décision dans un contexte de TALN. Elle consiste à optimiser l'apport des méthodes spécifiques de contrôle dans une chaîne de traitement.

Mots-clés : hypothèses concurrentes, architecture de contrôle, aide multicritère à la décision.

Abstract

The linguistic processing of textual utterances often leads to several concurrent interpretations that can raise artificial ambiguities. Controlling these awkward points is essential in order to obtain acceptable efficiency and precision of analysis processes. The approach described in this document makes use of the multicriteria decision aid paradigm in a context of NLP. It relies on the optimisation of the use of specific control methods in a processing chain.

Keywords: concurrent hypothesis, control architecture, multicriteria decision aid.

1. Introduction et motivations

L'application d'une chaîne de traitement linguistique sur un énoncé conduit le plus souvent à la prise en compte d'hypothèses concurrentes et ce à toutes les étapes du processus (analyse lexicale, syntaxique, sémantique, génération, etc.). La présence d'indéterminations et surtout leur propagation affectent la qualité, la précision et l'efficacité du processus d'analyse. Les points d'embarras[1] s'expliquent la plupart du temps par l'absence d'informations discriminantes sur les différentes alternatives, mais ils peuvent également résulter de l'usage de ressources linguistiques (lexiques, règles de construction, thésauri, etc.) imprécises ou trop génériques.

Le principal enjeu dans la résolution des indéterminations réside dans la construction et la mise à disposition d'informations complémentaires et précises, nécessaires pour transformer les points d'embarras en points de décision. Constatant l'aspect artificiel de ce phénomène d'indétermination et en s'appuyant sur les modèles cognitifs issus des travaux en psycholinguistique, des architectures logicielles parallèles ont été proposées : `CARAMEL` (Sabah, 1990), `TALISMAN` (Stefanini, 2004). Cette remise en cause des architectures classiques de traitement

[1] Terme utilisé dans (Sabah, 1989) pour désigner les étapes du processus de traitement où le choix pour "une meilleure" interprétation ne peut être émis à partir des informations disponibles.

automatique (modulaires, séquentielles, linéaires, etc.) se justifie par la recherche de complémentarité des différentes sources de connaissances, qui permet notamment de rendre disponible « simultanément »les interprétations lexicales, syntaxiques, sémantiques, etc. Cependant, les interconnexions ainsi que l'utilisation conjointe de ressources de nature différentes complexifient énormément les structures de données mises en jeu et rendent difficiles leur maintenance et leur développement. La seconde remise en cause concerne l'usage même de ressources linguistiques construites *a priori*. En effet, les approches stochastiques exploitent des corpus représentatifs pour extraire des régularités d'usage spécifiques à un contexte applicatif. Les ressources construites sont ainsi plus précises et pertinentes pour une tâche donnée. Cependant, malgré cette propriété, des ambiguïtés artificielles subsistent, nécessitant l'ajout de ressources génériques (Bourigault et Frérote, 2004). L'absence de corpus significatifs pour certains contextes ou certaines langues constitue également une limite à l'essort de ces méthodes.

Face à ce problème d'absence ou de manque d'informations discriminantes sur les hypothèses concurrentes lors des points d'embarras, de nombreuses méthodes spécifiques de contrôle ont été proposées. Sous la notion de méthode de contrôle, nous englobons toute procédure visant à rajouter, selon un certain point de vue, un critère de jugement sur chacune des interprétations envisagées. À partir de ces critères, des actions peuvent être réalisées pour réduire ou organiser l'espace des hypothèses. L'usage de méthode de contrôle pour superviser un processus d'analyse linguistique est une pratique récurrente, mais très peu de travaux cherchent à les combiner. Ceci peut sans doute s'expliquer par l'absence de formalisation du problème et de cadre générique d'usage de ces informations discriminantes. Pour pallier ce manque, nous proposons une approche décisionnelle à ce problème de gestion de l'indétermination. Nous cherchons notamment à optimiser l'apport des méthodes de contrôle ainsi qu'à faciliter leur utilisation dans un processus d'analyse linguistique. Nous nous intéressons plus particulièrement au comportement de la chaîne de traitement développée par l'équipe `Langues Naturelles` de `France Télécom division R&D` (Chardenon, 2005).

Cet article propose, dans un premier temps, une présentation de cette chaîne de traitement. Dans un second temps, nous présenterons la formalisation envisagée et l'architecture logicielle que nous avons implémentée pour gérer cette nouvelle définition du problème. Nous terminerons par les perspectives de validation et d'évaluation de l'apport de notre approche sur la précision des résultats et l'utilisabilité des méthodes de contrôle.

2. Contrôle d'une chaîne de traitement

2.1. Le processus d'analyse en question

L'équipe `Langues Naturelles` de `France Télécom division R&D` a développé une chaîne de traitement linguistique permettant d'analyser des énoncés écrits selon plusieurs niveaux d'interprétation. Différents modules (segmenteur, analyseur lexical, chunking, dépendance, sémantique, etc.) peuvent être combinés pour obtenir la précision de représentation linguistique désirée. Ce système se présente donc comme une boîte à outils paramétrable, appliquant séquentiellement les différentes étapes de traitement sélectionnées sur les résultats des modules précédents. L'architecture fonctionnelle du système respecte une dichotomie entre les algorithmes génériques et les ressources linguistiques dédiées à une lanque et sélectionnées pour une tâche précise. Ce type de stratégie propose une flexibilité et une adaptabilité intéressantes aux différents contextes applicatifs. Cependant, la couverture des phénomènes linguistiques ainsi que la précision des résultats dépendent essentiellement du jeu de ressources utilisé.

Un travail fastidieux et complexe d'adaptation, de sélection et de complétion des ressources est nécessaire pour chaque nouvel usage de la chaîne. Pour certains cadre applicatifs dits "ouverts", nécessitant de larges ressources, il est impossible de définir précisément les données utiles et pertinentes. De même, dans un souci de couverture maximale, des règles de correction ou des règles génériques de construction sont ajoutées. L'ensemble de ces paramètres : processus séquentiel, ressources imprécises, règles génériques, sont autant de sources d'indéterminisme et d'ambiguïtés artificielles.

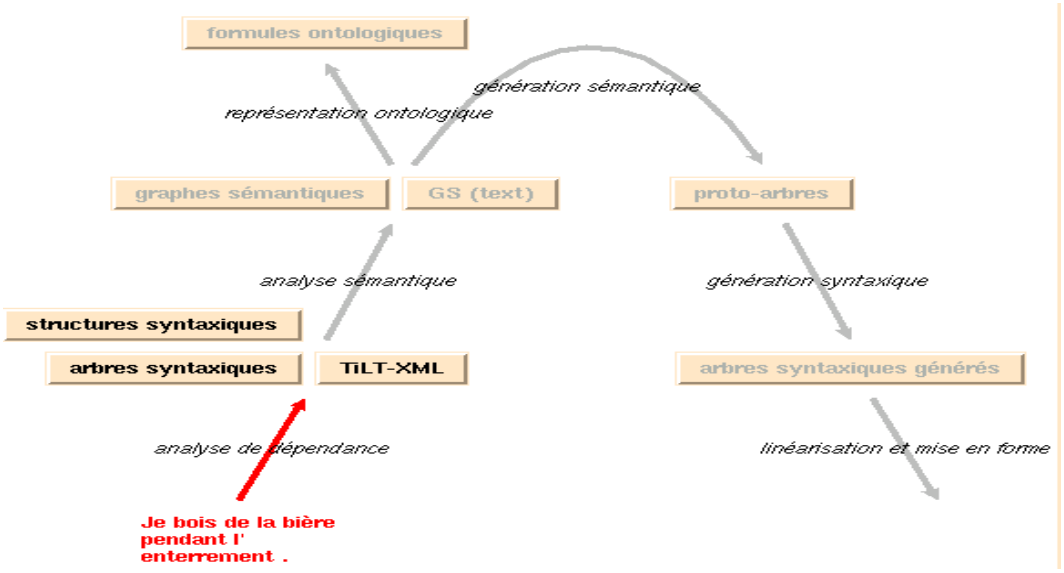

Figure 1. La chaîne de traduction

La figure 1 illustre la chaîne de traitement déployée pour une application de traduction automatique de requêtes du domaine pharmaceutique. Le processus d'analyse et de génération est donc composé d'une succession de modules dont l'objectif final est de disposer « d'une meilleure traduction »possible. De nombreux points de contrôle peuvent donc être ajoutés pour arriver plus rapidement et avec plus de certitude à cet objectif.

2.2. Les méthodes de contrôle

Afin d'ordonner, de filtrer ou de trier les hypothèses concurrentes générées dès les premières étapes de l'analyse, des jugements issus de différentes méthodes de contrôle peuvent être exploitées. La désambiguïsation du sens des mots (DSM) est un exemple d'étape intermédiaire de contrôle nécessaire et difficilement contournable dans un tel contexte de traduction automatique. Différents points de vue peuvent être émis pour privilégier un sens par rapport à d'autres (Ide et Véronis, 1998) :

1. la pertinence ou le partage de sens commun avec les sens du voisinage lexical ;
2. l'influence des catégories morpho-syntaxiques du mot et de ses proches voisins ;
3. le respect des contraintes sémantiques sur les structure prédicatives des verbes conjugués ;
4. des préférences empiriques définies *a priori* pour des domaines sémantiques ;
5. la probabilité d'usage d'un sens pour un mot selon un corpus d'apprentissage.

2.3. Une complémentarité inexploitée

Le contrôle des différents points d'embarras d'un processus d'analyse repose sur l'utilisation des préférences émises par une unique méthode de contrôle. Ce constat est valable pour la chaîne de traitement étudiée, mais également pour la plupart des travaux existants traitant du phénomène d'indétermination. Cependant, la prise de décision notamment pour la levée d'ambiguïtés artificielles repose rarement sur la prise en compte d'un unique point de vue, mais davantage sur l'agrégation de jugements pouvant être complémentaires ou contradictoires. Les trois premiers exemples du tableau ci-dessous illustrent sur notre exemple de la DSM que le choix du sens à privilégier peut reposer sur des critères différents et qu'il est donc indispensable de les prendre tous en considération lors de la décision. Le dernier exemple illustre un cas typique de contradiction entre deux critères.

Énoncé	Mot ambigu	Critères pertinents (Sec. 2.2 page 757)
le cachet d'aspirine	cachet(1-pharm.,2-salaire,3-poste)	Pref.1[1] $1 > 2 \geq 3$
des paris	paris(1-ville,2-jeu)	Pref.2 : $1 > 2$
je mange un avocat	avocat(1-droit,2-légume)	Pref.3 : $2 > 1$
À l'enterrement, je bois de la bière	bière (1-boisson,2-brancard)	Pref.1 : $2 > 1$ Pref.3 : $1 > 2$

Devant l'hétérogénïté des phénomènes d'ambiguïté artificielle, il apparaît indispensable de prendre en compte les différentes préférences émises par les méthodes de contrôle disponibles pour obtenir une décision globale plus fiable. Le fait que très peu de travaux aient cherché à combiner ces jugements s'explique sans doute par l'absence de cadre générique et flexible d'usage de préférences et de critères de jugement dans un contexte de TALN. L'approche présentée dans la section suivante propose de pallier ce manque.

3. Vers une approche décisionnelle du problème de l'indétermination

3.1. Formalisation du problème

Notre problème de contrôle des points d'embarras et des indéterminations, peut être dans une certaine mesure résolu en intégrant des relations de préférences entre les hypothèses concurrentes. Comme nous venons de le voir, les préférences peuvent représenter différents points de vue que nous souhaitons agréger pour améliorer la décision prise (filtrage, ordonnancement ou tri). Le paradigme de l'aide multicritère à la décision (Vincke, 1998) et plus précisément les méthodes de surclassement constituent une approche suffisament flexible et générique pour formaliser notre problème. Le contrôle d'un point d'embarras est alors pris en charge par un opérateur décisionnel qui en fonction d'une configuration établie *a priori* exploite un ensemble de préférences disponibles sur des hypothèses concurrentes de même nature. Ces préférences peuvent être directement établies par les méthodes de contrôle ou bien construites à partir des scores de pertinence calculés. On peut ainsi établir entre deux hypothèses a et b des relations de préférence stricte (aPb), d'indifférence (aIb) ou d'incomparabilité (aRb). La gestion de la complémentarité et des contradictions entre les jugements repose sur un ordre de pertinence

[1] Dans un contexte de traitement de corpus médicaux ou pharmaceutiques, cette préférence du sens 1 sur le sens 2 pourrait être renforcée par les méthodes de contrôle 4 et 5.

relative des critères entre eux selon un contexte donné. Il est ainsi possible en cas de contradiction d'établir selon cet ordre une prédominance d'un jugement par rapport à un autre (en DSM, le jugement basé sur le partage de sens communs avec le voisinage lexical est par exemple beaucoup plus pertinent sur des énoncés techniques que sur des énoncés dits « tout-venant »). Le résultat de l'application d'un opérateur de décision est donc constitué d'une structure de préférence, qui peut être exploitée entre deux modules de traitement pour propager uniquement les hypothèses jugées les plus pertinentes.

3.2. Ajout d'une composante décisionnelle dans l'architecture séquentielle de traitement

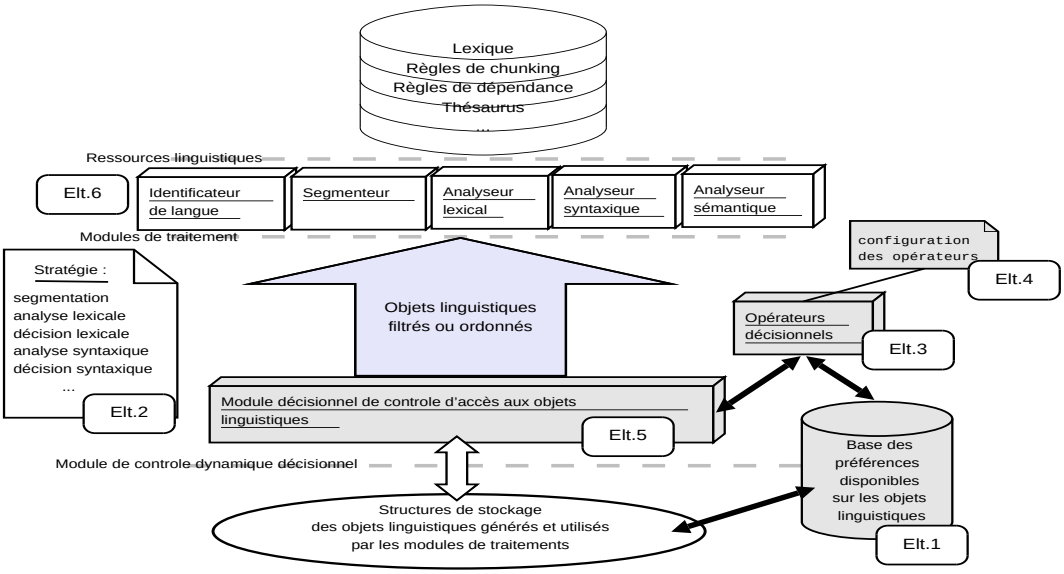

Figure 2. Intégration du contrôle décisionnel dans le processus de traitement

L'ajout d'étapes décisionnelles dans le processus d'analyse a nécessité la conception d'une architecture permettant d'intégrer aisément de nouveaux critères de jugement, d'appeler des opérateurs de décision sur ces données et d'exploiter leur résultat. Ce système décisionnel est tout d'abord composé d'une structure de stockage (Elt.1 sur la figure 2) dédiée aux différentes préférences émises sur les objets linguistiques[2] manipulés. Le comportement du processus de traitement est défini par une stratégie (Elt.2) qui stipule l'ordre d'appel des différents modules de traitement, ainsi que les étapes intermédiaires où un opérateur de décision (Elt.3) doit être appliqué. Le comportement de cet opérateur est externalisé dans un fichier de configuration `xml` (Elt.4), qui contient pour chaque étape de décision : la liste des identifiants des préférences utilisées, la distribution des poids d'importance des critères, la méthode d'agrégation utilisée et l'action à réaliser à partir de la structure de préférences générée. Les méthodes génériques d'accès aux objets linguistiques (Elt.5) exploitent le résultat de l'opérateur ainsi que les instructions de la configuration pour contrôler le comportement des modules de traitement (Elt.6) en filtrant ou en ordonnant les objets concurrents qu'ils prennent en entrée.

[2] Par objet linguistique, nous considérons toute structure d'informations manipulées par les algorithmes de traitement et intervenant dans la construction d'interprétations linguistiques intermédiaires ou finales.

4. Conclusion et perspectives

Afin de contrôler la présence d'hypothèses concurrentes lors des principales étapes du processus d'analyse, nous avons défini une approche basée sur l'aide multicritère à la décision. Nous disposons désormais d'outils permettant d'influencer le comportement de la chaîne de traitement à partir des jugements émis par différentes méthodes de contrôle.

Nous travaillons actuellement sur l'évaluation de l'utilisabilité de notre système ainsi que son apport pour limiter la présence d'ambiguïtés artificielles dans le processus de traitement déployée pour une tâche de traduction automatique. Sur un corpus dont l'annotation est en cours, nous allons sur quelques points d'embarras (DSM, POS-tagging, sélection d'une meilleure analyse en dépendance) évaluer l'apport de chaque jugement disponible actuellement dans la chaîne de traitement sur la décision globale (comparaison du jugement par rapport à la référence). Cette évaluation nous permettra d'ajuster automatiquement la distribution des relations d'importance entre les jugements. Nous pourrons ainsi quantifier l'apport d'une décision basée sur plusieurs critères complémentaires par rapport à l'usage individuel de chaque méthode de contrôle. L'analyse des premiers résultats nous permettra également d'identifier des critères qui ne sont pas disponibles actuellement dans la chaîne actuelle et qui pourrait améliorer la précision des résultats.

Références

BOURIGAULT D. et FRÉROTE C. (2004). « Ambiguïté de rattachement prépositionnel : introduction de ressources exogènes de sous-catégorisation dans un analyseur syntaxique de corpus endogène ». In *TALN 2003*.

CHARDENON C. (2005). « Analyse syntaxique en dépendance et évaluation ». In *TALN 2005*.

IDE N. et VÉRONIS J. (1998). « Word Sense Disambiguation : The State of the Art ». In *Computational Linguistics*.

SABAH G. (1989). *L'IA et le langage*. Hermes.

SABAH G. (1990). « CARAMEL : A flexible model for interaction between the cognitive process underlying natural language understanding ». In *the 13th International Conference on Computational Linguistics*.

STEFANINI M. (2004). « TALISMAN : un Système Multi-Agents pour le TAL ». In *TALN 2004 - AGENTAL "Agents et Langue"*.

VINCKE P. (1998). *Aide multicritère à la décision*. Ellipses Marketing.