Approche linguistique

Sommaire

Résumé

Ce chapitre s'intéresse aux principaux fondements de l'approche linguistique du problème de la compréhension. L'application du principe de compositionnalité est exposé en 1.2. La partie 1.3 présente les grammaires formelles de Chomsky. Les évolutions et les premières applications de ces grammaires sont détaillées dans la section 1.4. Enfin, la partie 1.5 présente les grammaires stochastiques et quelques systèmes les utilisant.

1.1 Introduction

Les systèmes développés selon l'approche linguistique se basent sur l'application du principe de compositionnalité de Frege à partir d'une analyse syntaxico-sémantique de la proposition. Après de brefs rappels théoriques, quelques-uns de ces systèmes sont présentés dans ce chapitre. Les premiers d'entre eux ont été réalisés sous la contrainte technique d'une puissance de calcul restreinte et sont pionniers dans le domaine de l'interaction homme-machine.

Le principe de compositionnalité de Frege est explicité en 1.2. Les grammaires formelles de Chomsky, créées dans ce contexte réflexif et appliquées à la construction des arbres syntaxiques, sont présentées en 1.3. La partie 1.4 expose des évolutions de ces grammaires orientées vers la prise en compte de connaissances sémantiques. L'introduction de paramètres stochastiques dans les approches à base de grammaires est exposée dans la section 1.5.

1.2 Principe de compositionnalité

Dans la plupart des systèmes issus de l'approche linguistique, tous les sens possibles de chaque mot sont considérés. Ces informations sont ensuite composées sous la contrainte d'obtenir un sens cohérent pour chaque proposition. Une approche de ce type, décrite dans (Allen, 1988), consiste à analyser une phrase écrite pour obtenir l'arbre syntaxique qui lui est associé. Un ensemble de règles fait ensuite correspondre les blocs de l'arbre à des fragments de représentations sémantiques définis au sein d'une ontologie structurée.

Cette approche, issue des travaux de Frege, est justifiée par l'hypothèse que chaque constituant syntaxique important d'une phrase correspond à un constituant conceptuel, la réciproque étant fausse. La figure 1.1 présente l'exemple de l'arbre sémantique associé à la proposition "Je cherche un hôtel Sofitel pour le soir du 25 octobre"

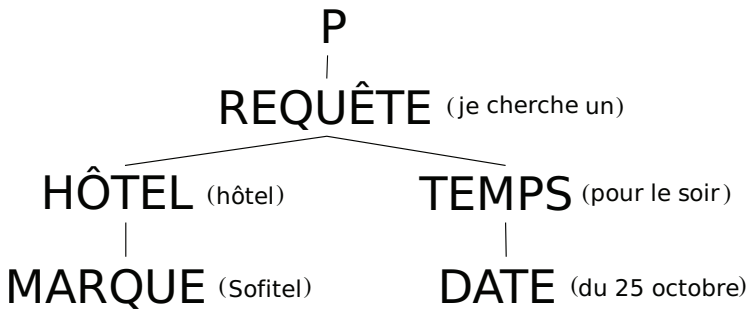

FIGURE 1.1 – *Arbre sémantique associé à la proposition "Je cherche un hôtel Sofitel pour le soir du 25 octobre".*

Ce formalisme est basé sur un ensemble de catégories. Chaque catégorie peut être détaillée par une fonction et un argument. Dans l'exemple ci-dessus, la catégorie DATE

est associée à la fonction *du* et l'argument *25 octobre*. L'expression P peut être obtenue à partir d'une structure syntaxique telle que :

G : P[VP [PR je, V cherche] NP [ART un, N hôtel, N Sofitel] NP [PREP pour, ART le, N soir] NP [PREP du, ADJ vingt-cinq, N octobre]]]

Une partie de l'arbre syntaxique de cette structure est présentée dans la figure 1.2

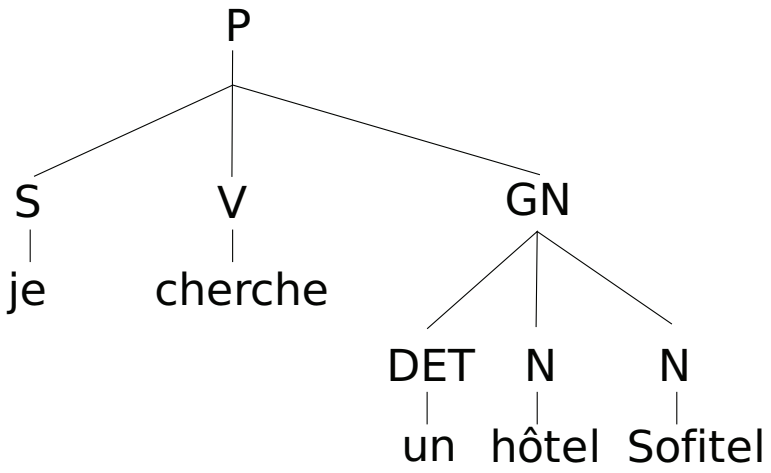

FIGURE 1.2 – *Arbre syntaxique associé à la proposition "Je cherche un hôtel Sofitel".*

Selon les domaines d'application, des représentations sémantiques peuvent être associées à des nœuds non terminaux de l'arbre syntaxique et l'interprétation de la phrase peut être réalisée en utilisant les étiquettes sémantiques de ces associations.

1.3 Grammaires formelles

Ces modélisations sont issues des grammaires formelles de Chomsky. Ces grammaires s'inspirent du langage formel et tentent d'intégrer les caractéristiques du langage humain à l'aide de règles d'association des mots. Composées d'un nombre fini de règles de production (règles de réécriture), elles permettent de générer et d'analyser un langage donné.

Par définition, une grammaire formelle G est un quadruplet (V_N, V_T, R, P) avec :
V_T : vocabulaire terminal (ensemble des symboles terminaux)
V_N : vocabulaire non terminal (ensemble des symboles non terminaux)
V : vocabulaire ($V = V_T \cup V_N$)
P : axiome, élément de V_N
R : ensemble de règles de réécriture de la forme : $\alpha \rightarrow \beta$ telles que $(\alpha, \beta) \in V^* \times V^*$ et $\alpha \neq \epsilon$

On appelle *langage engendré par G* l'ensemble de toutes les suites de symboles qui dérivent de l'axiome P de G : $L(G) = \{x, x \in V^* etP \Rightarrow^* x\}$ (i.e. x peut être obtenu à partir de P grâce à une succession de réécritures).

Un langage est dit *décidable* si pour toute phrase, on peut savoir au bout d'un temps fini si elle appartient ou non au langage.

Les grammaires formelles sont classées hiérarchiquement par Chomsky et Schützenberger (Chomsky et Schützenberger, 1963) par ordre d'expressivité décroissante.

◦ **Les grammaires de type 0** se définissent par des règles du type :

$$\alpha \rightarrow \beta \quad avec \quad \alpha, \beta \in V^*$$

Ces grammaires ne sont pas décidables.

◦ **Les grammaires de type 1**, dites grammaires contextuelles, se définissent par des règles du type :

$$\alpha A \beta \rightarrow \alpha \gamma \beta \quad avec \quad A \in V_N \quad \alpha, \beta, \gamma \in V^*, \gamma \neq \epsilon$$

Toute règle comporte un symbole non terminal entre deux mots que l'on retrouve après la dérivation. Le non terminal est transformé de façon non nulle. Les mots qui encadrent le non terminal représentent son contexte qui va influer sur sa dérivation. Les grammaires de type 1 sont décidables. Pour déterminer si une phrase de longueur n appartient au langage, il suffit de réaliser toutes les dérivations comportant n symboles ou moins, ce qui nécessite un temps fini. Cependant, la génération est de complexité exponentielle en n (le temps d'analyse est proportionnel à l'exponentielle du nombre de mots à analyser).

◦ **Les grammaires de type 2**, dites grammaires algébriques ou hors-contexte, se définissent par des règles de la forme :

$$A \rightarrow \gamma \quad avec \quad A \in V_N \quad \gamma \in V^*$$

Bien que limitées par leur incapacité à traiter les dépendances longues distances, l'usage des grammaires hors-contexte est souvent privilégié en TALN, essentiellement en raison du bon compromis entre leur capacité descriptive et leur complexité (polynomiale en $O(n^3)$, où n est le nombre de mots à analyser, d'après l'algorithme CYK).

◦ **Les grammaires de type 3** ou grammaires régulières sont linéaires gauches ou droites. Les règles qui les définissent sont de la forme :

grammaire linéaire gauche	*grammaire linéaire droite*
$A \rightarrow Ba$	$A \rightarrow aB$
$A \rightarrow a$	$A \rightarrow a$
$avec\ A, B \in V_N, a \in V_T$	$avec\ A, B \in V_N, a \in V_T$

Les grammaires régulières sont utilisés en TALN pour la représentation compacte des lexiques, la construction de correcteurs orthographiques, la composition de grammaires

locales (traitement des nombres). Dans le cadre du dialogue et pour des applications à des domaines restreints, elles sont souvent choisies en raison de leur complexité linéaire.

◦ **Les grammaires de type 4**, dites grammaires à choix finis, se définissent par des règles de la forme :

$$A \rightarrow a \quad avec \quad A \in V_N \quad a \in V_T$$

Ces grammaires ne permettent que l'énumération des phrases de leur langage sur V_T.

1.4 Évolutions

Conscient de l'incapacité des grammaires hors-contexte à modéliser toutes les subtilités du langage naturel (Chomsky, 1964), Woods propose l'utilisation des grammaires à base de réseaux de transitions augmentés (GRTA) dans les procédures de décomposition syntaxique (Woods, 1970). Dans la perspective de mieux modéliser richesse et complexité du langage naturel, ces grammaires contiennent des connaissances sémantiques sensibles au contexte et leurs stratégies de décomposition syntaxique incluent des processus d'inférence logique.

Ces grammaires sont une extension des grammaires à base de réseaux de transitions (GRT). Les GRT sont faiblement équivalentes aux grammaires hors-contexte dont elles ne diffèrent en équivalence forte que par leur aptitude à caractériser les arborescences redondantes du type S[S et S et ... et S]. Elles intègrent, via des réseaux de transitions, les caractéristiques que les grammaires de transitions ajoutent aux grammaires hors-contexte.

Composées d'états reliés par des arcs (graphes orientés), ces grammaires ont l'expressivité des grammaires hors-contexte à laquelle s'ajoute la capacité de déplacer des fragments de structure, de les recopier, de les supprimer : Ces actions sont généralement dépendantes du contexte dans lequel les fragments apparaissent. La chaîne d'entrée est analysée de gauche à droite durant la décomposition, mot par mot. Le mot entrant et l'état courant détermine l'arc emprunté par le processus. Des GRT sont utilisées pour la compréhension de la parole spontanée par (Young et al., 1989).

Dans les GRTA, des tests conditionnels peuvent être associés à certains arcs et un ensemble de structures de construction peuvent être effectuées si l'arc est emprunté (composition d'arbres, génération d'interprétations sémantiques). En effet, le réseau de transitions augmenté fournit une description structurelle partielle de la phrase à chaque état. Ces descriptions sont stockées dans des registres mis à jour au fil de l'analyse. Le contenu des registres est composé des valeurs des caractéristiques linguistiques et peut aussi être utilisé pour construire les arbres d'analyse. Une approche de ce type est décrite dans (Woods et al., 1976) et est proposée dans le projet ARPA de 1971, détaillé dans (Klatt, 1977). Il inclut des approches essentiellement basées sur l'Intelligence Artificielle (IA) pour combiner analyse syntaxique et représentation sémantique en logique formelle. Les systèmes de ce projet génèrent des hypothèses de séquences de mots grâce

à un système de reconnaissance automatique de la parole puis produisent une interprétation avec les mêmes approches que celles utilisées pour le traitement de l'écrit.

Les concepts et relations d'un réseau sémantique peuvent aussi être implémentés dans le formalisme des *cadres sémantiques* (Fillmore, 1985). Le concept linguistique original des cadres sémantiques (ou cadres de cas, *case-frame*, ou encore grammaires de cas, *case-based grammar*), comme proposé par (Fillmore, 1968), est basé sur la définition d'un ensemble de cas universels qui permet de mettre en avant la relation entre un verbe et ses composants nominaux. En se référant à la terminologie de (Bruce, 1975), un *cas* est une relation entre un prédicat (en général le verbe, mais pas exclusivement) et un des ses arguments. Un marqueur de cas est un indicateur de structure de surface (par exemple une préposition) pour le cas concerné. Un cadre de cas d'un prédicat est constitué de l'ensemble des cas propres à ce prédicat. Une grammaire de cas est un jeu complet de cas pour un langage entier.

L'approche par grammaire de cas est appropriée pour les systèmes de compréhension de la parole où le besoin d'un support sémantique lors de l'analyse est fondamental. L'analyseur sémantique réalise une analyse par cas pour déterminer le sens d'une requête, et construit la représentation en cadres correspondante. L'historique du dialogue est utilisé pour compléter le cadre sémantique en cas de besoin. Les cas définissant la grammaire complète ainsi que les mots-clés associés (*trigger keywords*) sont dans une large mesure dépendants de la tâche et du domaine du système de dialogue. Les grammaires de cas ont été appliquées avec succès dans de nombreux systèmes (par exemple (Hayes et al., 1986; Ward, 1991)) et largement popularisées à partir des années 90 par leur utilisation dans les systèmes du LIMSI (Matrouf et al., 1990; Bennacef et al., 1994, 1996; Lamel et al., 1999). Un exemple d'un cadre utilisé pour la tâche ATIS en français (Bennacef et al., 1994) est donné dans la figure 1.3.

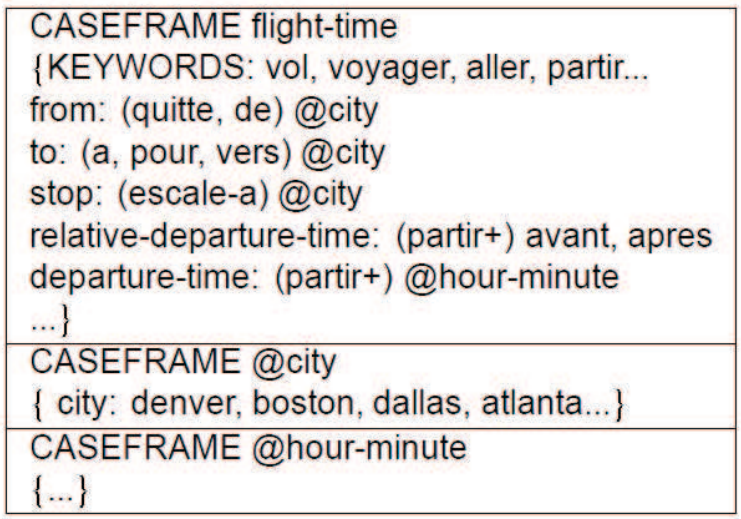

FIGURE 1.3 – *Exemple de cadre sémantique pour la tâche ATIS*

Le formalisme des cadres sémantiques servant de base à l'approche présentée dans

nos travaux, nous y reviendrons plus longuement dans la section 3.3.

Des exemples récents d'application du formalisme des grammaires à la problématique de la compréhension dans les systèmes de dialogues peuvent être trouvés dans (Villaneau et al., 2004; Denis et al., 2006). Dans (Denis et al., 2006), des grammaires d'arbres disjoints sont utilisées pour modéliser la tâche de dialogue. L'approche, basée sur l'analyse syntactique profonde (*deep-parsing*) et la logique de description, se décompose en 2 étapes :

- un analyseur LTAG (Crabbe et al., 2003) produit une analyse syntactique de la phrase. Seules les derivations partielles sont recherchées et les plus longues sont conservées ;
- un constructeur sémantique produit un graphe conceptuel à partir des analyses syntactiques précédentes. Le graphe conceptuel est ensuite réévalué en confrontation avec une ontologie interne, de sorte à éliminer les relations inconsistantes.

1.5 Grammaires stochastiques

Pour prendre en compte l'ambigüité d'analyse liée aux spécificités structurelles des messages oraux et surtout l'imprécision des transcriptions issues de la reconnaissance de la parole, les grammaires sémantico-syntaxiques présentées précédemment évoluent vers des grammaires stochastiques en utilisant un corpus d'apprentissage.

Des analyseurs CHART peuvent être utilisés pour stocker les forêts de sous-arbres représentant les résultats intermédiaires de l'analyse syntaxique lorsque les erreurs de reconnaissance ont empêché l'analyse complète d'aboutir. Une grammaire hors contexte probabiliste peut alors estimer la probabilité d'une analyse partielle. Cette estimation s'appuie sur un algorithme polynomial (en $O(n^3)$), ce qui rend possible l'utilisation de ces grammaires dans les systèmes de compréhension opérationnel (Corazza et al., 1994).

L'analyseur syntaxico-sémantique TINA développé à l'institut de technologie du Massachussetts (MIT) utilise ainsi une grammaire hors contexte probabiliste de type GRTA (Seneff, 1989). Cette grammaire est automatiquement convertie en un graphe dont les sommets sont les catégories syntaxiques ou sémantiques et les arcs portent les probabilités des règles, estimées sur un corpus d'apprentissage. Les constructions les plus fréquemment rencontrées sont donc privilégiées au cours de l'analyse. En cas d'échec de l'analyse complète, les analyses partielles sont réalisées à partir de chaque mot du message. D'autres approches à base de grammaires et d'analyse robuste ont été proposées, tel le Structured Language Model décrit dans (Chelba, 1997) et adapté à l'analyse sémantique par (Bod, 2000).

L'approche par analyse de surface de (Gildea et Jurafsky, 2002) permet de détecter les relations sémantiques contenues dans un message. Ces relations, nommées rôles sémantiques, sont formalisées par une représentation sémantique de haut niveau, indépendante de la tâche considérée.

Le système proposé par (Gildea et Jurafsky, 2002) est basé sur des classifieurs entrainés sur les phrases annotées manuellement en rôles sémantiques (selon les standards du projet FrameNet décrit en 3.4). Les phrases d'entraînement sont soumises à un analyseur syntaxique. Leurs sont alors associées leurs caractéristiques lexicales et syntaxiques ainsi que les probabilités a priori des différentes combinaisons des rôles sémantiques qu'elles contiennent. Les phrases testées sont analysées et annotées avec les caractéristiques extraites durant cette analyse puis soumises aux classifieurs, fournissant les étiquettes sémantiques de leurs constituants.

Ces méthodes sont reprises par (Pradhan et al., 2004) qui proposent un algorithme d'apprentissage pour l'analyse sémantique de surface basé sur des classifieurs à noyaux de type machines à vecteurs de support (séparateurs à vaste marge - SVM).

Dans le contexte de l'étiquetage sémantique, les travaux présentés dans (Moschitti, 2006; Moschitti et al., 2008) s'appuient sur l'utilisation de fonctions à noyaux adaptées aux traitement des arbres. L'analyseur utilisé génère un arbre syntaxique dans lequel les feuilles sont tout d'abord annotées. L'information sémantique est ensuite propagée vers la racine selon différentes stratégies pour produire un arbre syntaxico-sémantique initial. Tous ses sous-arbres sont alors extraits par des fonctions à noyaux spécifiques (*tree kernels*). L'usage de noyaux différents entraîne l'extraction de différents types de sous-arbres. Des classifieurs SVM, appris sur ces ensembles de sous-arbres, permettent de décider de l'arbre syntaxico-sémantique final à privilégier, de reclasser les hypothèses d'arbres et également d'évaluer les heuristiques de propagation de l'information sémantique.

1.6 Conclusion

Dans le cadre de la compréhension du dialogue oral, les approches strictement linguistiques fondées uniquement sur des grammaires sont donc souvent mises en défaut par la structure même du message véhiculé. En effet, les messages oraux ne sont pas formulés selon les mêmes normes que les messages écrits. Ces messages sont souvent agrammaticaux, contiennent des répétitions, des phrases inachevées. Une part importante des informations contenues dans les échanges oraux est perdue lors de leur transcription. Il en est ainsi par exemple des informations prosodiques indiquant souvent le mode de l'échange (interrogatif, affirmatif...) ou des informations implicites telles que les silences ou les hésitations.

De plus, les performances des systèmes de reconnaissance de la parole sont telles que de nombreuses erreurs émaillent encore les transcriptions automatiques des messages. Les méthodes d'analyse robuste présentées précédemment améliorent la qualité et la couverture des interprétations sémantiques obtenues par les approches linguistiques.

L'approche stochastique de la compréhension du dialogue oral est actuellement l'alternative principale aux méthodes linguistiques. Basée sur l'apprentissage, elle permet

de concevoir des systèmes mieux adaptés aux spécificités de l'oral et notamment aux erreurs de transcription. Cette approche est présentée dans le chapitre 2 suivant.