Les types d'analyse dans quelques applications du TAL

Marguerite Leenhardt
M2 Professionnel Ingénierie Multilingue
INaLCO Promotion 2007-2008

6 février 2008

Résumé

Du point de vue du Traitement Automatique des Langues (TAL), la notion d'analyse renvoie à l'analyse des formes linguistiques, c'est-à-dire des séquences de symboles constitutives du donné linguistique. Il s'agit de mobiliser un modèle d'interprétation de ces séquences de symboles, défini en fonction de données extérieures que sont, par exemple, les connaissances linguistiques, entre autres conventions d'interprétation des symboles que l'on souhaite analyser. Nous tenterons dans ce devoir de faire état des différents niveaux d'analyse, des transformations qu'elles appliquent aux données en entrée et d'identifier les problèmes de robustesse des applications qui découlent de ces analyses.

1 La notion d'analyse en TAL

Précisons d'emblée que l'on ne peut parler d'analyse en soi, en tant qu'un type d'analyse ne se peut appliquer uniformément aux données linguistiques. En effet, toute analyse est fonction de ce qu'on l'on veut en faire, c'est-à-dire des données auxquelles on veut l'appliquer.
Puisque le TAL a pour objet le donné linguistique, le dénominateur commun aux différents types d'analyses qui en découlent est l'objectif de conférer une structure à une information peu structurée, en fonction d'informations extérieures. De ce point de vue, le problème de la robustesse des applications du TAL mobilisant un ou plusieurs modèles d'interprétation des données linguistiques peut s'envisager en termes de capacité à analyser des entrées non conformes à un modèle donné, compte-tenu de la variabilité du phénomène linguistique. L'exigence de robustesse est donc liée au problème de la détermination du contexte, en termes de connaissances intra et extralinguistiques, qui doit être mobilisé pour une analyse donnée : comment intégrer le contexte à un modèle d'interprétation des données ? La détermination du contexte intégré au modèle d'interprétation est un point crucial, puisqu'il participe de la performance d'une application : dans quelle mesure l'étendue du contexte mobilisé est-elle pertinente, compte-tenu des informations que doit produire une analyse donnée pour une application donnée ?

2 Analyses et applications

Nous choisissons de distinguer dans ce devoir deux types d'applications, selon les analyses qu'elles incluent et les tâches pour lesquelles elles sont développées. Ainsi, la distinction entre *applications atomiques* et *applications complexes* relève de l'étendue de l'objet analysé et du système dans lequel elles sont intégrées. Par exemple, les systèmes de Questions/Réponses ou encore les moteurs de recherche peuvent mobiliser des modules d'analyse syntaxique, lexicale ou sémantique : ce sont de tels systèmes auxquels on réfère par *applications complexes*. Il est entendu que l'on ne prétend pas donner ici un panorama exhaustif des applications du TAL, puisque notre objectif est de donner un aperçu des niveaux d'analyse et d'envisager les problèmes de robustesse des traitements pour quelques applications.

2.1 Analyse syntaxique

L'analyse syntaxique complète et l'analyse syntaxique locale ont pour objectif de produire une représentation arborescente à partir de séquences de texte ; le modèle d'interprétation des données est une grammaire.

2.1.1 Analyse syntaxique complète

L'objectif de l'analyse syntaxique est de produire un arbre syntaxique complet pour une phrase. Il s'agit donc de passer d'une représentation séquentielle à une représentation arborescente. Plusieurs problèmes de robustesse se posent. En premier lieu, le problème de la couverture de la grammaire, que l'on peut assimiler au modèle d'interprétation des données en entrée. Plus la grammaire gère de cas particuliers, donc plus sa couverture est étendue, plus l'ambiguïté augmente. Cela a pour conséquence de multiplier les choix entre les diverses représentations arborescentes possibles pour une même phrase, et pose donc le problème du choix de l'analyse correcte parmi l'ensemble des analyses possibles. De plus, l'analyse syntaxique est peu performante, en termes de rapidité d'exécution, lorsqu'elle est appliquée à de grands corpus, c'est-à-dire sur des corpus de millions ou milliards de mots. En fait, mobiliser une analyse syntaxique dépend du type d'application complexe visé : si elle est pertinente pour une tâche d'interprétation ou de comparaison de phrases – où il s'agit, respectivement, d'associer une représentation syntaxique à une représentation sémantique et de comparer les structures syntaxiques de phrases en langue source et langue cible pour une tâche de traduction automatique ou d'alignement -, l'analyse syntaxique complète n'est pas pertinente pour d'autres types d'applications, pour lesquelles elle fournit plus d'information que nécessaire. Par exemple, pour des tâches d'extraction d'information ou d'acquisition lexicale, une analyse syntaxique locale ou chunking suffit.

2.1.2 Chunking : analyse locale ou analyse syntaxique partielle

Il s'agit également de passer d'une représentation séquentielle à une représentation arborescente, mais la sortie d'une analyse locale est une arborescence moins complexe que celle d'une analyse syntaxique complète : la profondeur de l'arbre en sortie est limitée car certaines relations d'inclusion des constituants

syntaxiques ne sont pas prises en compte, comme par exemple l'association des syntagmes prépositionnels au constituant syntaxique dont ils sont modifieurs. De fait, la grammaire d'une analyse locale est moins étendue, puisqu'elle vise l'identification de syntagmes sans chercher à prendre en compte les relations syntaxiques qu'ils entretiennent au sein d'une phrase. En effet, le chunking consiste à identifier des syntagmes à partir d'une grammaire dont les règles ne sont pas récursives. Une analyse locale propose toujours, en principe, une solution ; elle est plus rapide sur de grands volumes de données qu'une analyse syntaxique complète. De ce point de vue, si l'on évalue la robustesse en termes de capacité à analyser des entrées non conformes à un modèle donné, l'analyse locale est une alternative robuste à l'analyse syntaxique complète. Le chunking est une méthode utilisée, par exemple, pour des tâches d'extraction d'information, d'annotation pour de l'apprentissage automatique, mais aussi dans les systèmes de Questions/Réponses.

2.2 Analyse morphologique ou analyse lexicale

L'analyse morphologique consiste en l'analyse interne de la structure du mot, du point de vue de sa composition en morphèmes. Cette analyse repose donc sur l'identification des morphèmes, étape préalable à un certain nombre de manipulations – la substitution et l'effacement, par exemple –, qui visent à déterminer la valeur distinctive d'un morphème dans le mot. Il s'agit donc de passer d'une représentation séquentielle, c'est-à-dire le mot, à une représentation componentielle ou arborescente, c'est-à-dire une mise en évidence des éléments constitutifs du mot, donc de sa structure ; le modèle d'inteprétation des données est un lexique morphologique ou un ensemble de règles de dérivation lexicale.
Comme on l'a abordé plus haut, le choix des symboles, en tant qu'unités d'analyse, donc le choix d'interprétation du donné linguistique que l'on souhaite analyser, influe sur le type d'analyse, puisque cela relève du choix du modèle d'interprétation des données. Du point de vue du TAL, une des problématiques de l'analyse est de pouvoir prendre en compte une modélisation des données non prévues par le modèle. En l'occurrence se pose le problème des mots inconnus. Traditionnellement, les mots inconnus sont les noms propres, les mots erronés – mal orthographiés, par exemple –, et les néologismes. La définition de règles d'analyse peut être fondée sur des critères linguistiques, par exemple les indices de constructivité lexicale, ou sur des critères formels, par exemple la comparaison de chaînes de caractères par rapport à une référence. Cependant, de la même façon que l'extension de la couverture des règles de grammaire dans le cadre d'une analyse syntaxique génère davantage d'ambiguïté dans les résultats, l'ajout de règles pour l'analyse morphologique n'en améliore pas forcément les performances. Si certaines contraintes permettent d'améliorer la performance d'une règle, elles peuvent du même coup exclure nombre d'analyses pertinentes. Le point de vue de pertinence est donné par les performances requises par l'application qui mobilise une telle analyse, comme nous l'avions suggéré pour l'analyse syntaxique. C'est en effet en fonction des objectifs de l'application qu'il faut déterminer le modèle d'analyse.

2.3 Analyse sémantique

De façon très générale, l'analyse sémantique prend pour objet des phrases ou des textes entiers et a pour but d'en déterminer le sens, à partir de l'examen du sens des mots, mais aussi des expressions et des phrases, en fonction des relations qu'ils entretiennent.
Il s'agit donc de passer d'une représentation séquentielle à une représentation structurale et relationnelle, qui peut prendre, entre autres, la forme d'une structure arborescente ou d'une représentation ensembliste ; le modèle d'interprétation des données est un ensemble règles formalisant les structures sémantiques possibles pour les unités considérées et les relations sémantiques possibles entre elles. Les graphes et la représentation réseaunale sont aussi utilisés pour la modélisation des relations sémantiques.
Dans ce type d'analyse, la dépendance au contexte est primodiale, puisque la caractérisation du sens est contextuellement déterminée. L'analyse sémantique est parfois mobilisée pour lever l'ambiguïté d'un sous-ensemble d'analyses syntaxiques possibles pour une phrase. Elle est par exemple aussi utilisée dans des applications liées au marketing, pour caractériser l'opinion des consommateurs sur un produit, un service ou une entreprise. Plus que tout autre domaine d'analyse du donné linguistique, l'analyse sémantique est confrontée à la variabilité des productions langagières. Les ressources qu'elle mobilise sont très larges et les procédés sur lesquels elle se fonde très divers. La difficulté provient essentiellement du fait qu'un modèle sémantique sous-tend une représentation du monde qui peut ou non être en adéquation avec le fait linguistique même. Un modèle fondé sur une représentation en termes de concepts peut être soumis à une critique fondamentale, linguistiquement parlant, qui tient au fait d'une réduction de la compexité du sens dans la notion de concept, notion qui tend à s'abstraire de cette complexité en érigeant des invariants. Bien qu'une représentation conceptuelle, partant, logiciste du sens puisse être efficiente pour certaines applications telles que la constitution d'ontologies, par exemple, il est admis qu'envisager le sens en termes prédicatifs est un parti pris impuissant pour rendre compte de la diversité des sens d'une unité, puisque celui-ci est fonction de la variation contextuelle. La représentation en structures conceptuelles du sens évacue le processus interprétatif, qui donne pourtant sa valeur distinctive à une unité de sens dans la langue. Cependant, il n'est pas surprenant que nombre d'applications favorisent la représentation conceptuelle, compte-tenu de la complexité de pouvoir traduire à l'aide d'une machine – qui est fondamentalement un outil de calcul symbolique –, l'ensemble des processus cognitifs de la détermination du sens par un locuteur humain. La contrainte de robustesse des systèmes se satisfont de représentations conceptuelles ; on peut par exemple penser aux applications telles que la recherche d'information ou la fouille de textes, qui se fondent, entre autres, sur des représentations en graphes sémantiques, dans le cadre de systèmes tels que les moteurs de recherche, le résumé automatique ou encore la classification automatique de textes ou de documents.

2.4 Analyse de l'oral

L'analyse de l'oral est l'analyse de séquences sonores, en fonction de caractéristiques acoustiques ; le modèle d'interprétation des données est un ensemble de règles permettant de caractériser un phonème sur la base d'indices acous-

tiques : on passe donc d'une représentation séquentielle à une représentation séquentielle. En effet, par exemple pour une tâche d'annotation phonétique automatique, on passe d'une séquence de son brute, néanmoins acoustiquement caractérisée dans un spectrogramme, à une séquence de son enrichie d'une annotation phonétique ou phonologique. Pour une tâche de transcription automatique, on passe d'une séquence de son à une séquence de texte, qui peut être soit une suite de symboles phonétiques ou phonologiques, soit une séquence de symboles alphabétiques. De tels résultats peuvent être les données en entrée pour une analyse syntaxique de l'oral, par exemple. Un autre type d'application, la synthèse vocale, passe d'une représentation séquentielle, c'est-à-dire une suite de symboles alphabétiques ou phonétiques, à une représentation séquentielle, c'est-à-dire une séquence sonore.
Sans aller plus avant dans le détail des applications issues du traitement de l'oral, le point à souligner est que l'analyse prend en entrée une séquence et produit en sortie une séquence.

3 Modèles d'interprétation, problèmes de robustesse

Les applications que nous qualifions d'*atomiques* - analyses syntaxique, lexicale, sémantique, de l'oral - mobilisent des modèles d'interprétation à base de règles, on retrouve donc bien la conception sous-jacente de langage formel en tant que formalisation interprétant une suite de symboles en fonction d'un ensemble de règles donné. A l'exception de l'analyse de l'oral, la majorité des analyses du donné linguistique partent d'une représentation séquentielle pour produire une représentation structurée, sous forme de graphes essentiellement, un arbre étant un graphe particulier. L'écrit et l'oral sont des donnés linguistiques différents, les modèles d'interprétation sous-tendus par leurs analyses respectives ne produisent donc pas les mêmes représentations.
Si l'on entend par *robustesse* la capacité à pouvoir analyser des entrées non conformes à un modèle donné, on peut se demander si l'analyse de l'oral relève de la problématique de la robustesse : toute séquence sonore est conforme à un modèle donné, en tant qu'elle est acoustiquement caractérisable.
Le problème du rapport entre l'étendue de l'ensemble de règles d'interprétation pour une séquence textuelle et l'ambiguïté des analyses produites, comme nous l'avons évoqué par exemple pour l'analyse syntaxique, est par contre un problème de robustesse. Il semble que, de la même façon que la nature du phénomène que l'on souhaite analyser, comme la nature des informations produites par l'analyse en fonction des besoins de l'application dans laquelle elle s'insère, participent de la détermination du modèle d'interprétation, la notion de *robustesse* soit aussi conditionnée par l'ambiguïté moindre du résultat de l'analyse. L'analyse robuste aurait donc pour objectif, non seulement de gérer des entrées non conformes à un modèle donné, mais aussi de pouvoir associer à une entrée donnée un résultat, et ce d'une façon la plus univoque possible. Une analyse robuste serait donc analogue à une relation qui, à un élément de l'ensemble de départ, associe un élément de l'ensemble d'arrivée.