École Nationale Supérieure des Mines de Paris
Corps Techniques de l'État

Panorama de la recherche en traitement automatique du langage écrit au Japon

Jean-Philippe Vert
Ingénieur des Mines

Juillet 1998

Table des matières

Chapitre 1

Introduction

Le traitement automatique du langage naturel (TALN) a progressé de manière spectaculaire avec l'apparition d'ordinateurs toujours plus puissants, et le développement parallèle de méthodes et d'outils aux confluents de la linguistique, des mathématiques et de l'informatique. Et si le chemin reste long avant de parvenir à résoudre des problèmes aussi naturels que la traduction automatique (TA) où la création d'opérateurs virtuels capables de dialoguer en langage naturel, les enjeux économiques liés aux avancés de la discipline sont colossaux. Les principales grandes entreprises du secteur de l'informatique, notamment américaines et japonaises, investissent d'ailleurs des sommes considérables et tissent des réseaux de recherche dans ce domaine, également promu à travers des grands programmes par la DARPA aux États-Unis, le MITI au Japon, ou la Commission Européenne.

L'objet de ce rapport est de présenter un panorama de la recherche effectuée au Japon dans le domaine du TALN, limitée au traitement du langage écrit. Il vise tout autant à évoquer les thèmes de recherche actuels qu'à indiquer les laboratoires concernés, afin de promouvoir les échanges avec les laboratoires francophones intéressés. Il est le fruit d'un voyage d'études réalisé en juillet 1998 par un ingénieur du corps des Mines, en dernière année de formation à l'École Nationale Supérieure des Mines de Paris, qui lui a permis de rencontrer les principaux acteurs de la discipline et de visiter leurs laboratoires.

Ce travail a été financé par la Direction de l'Action Régionale et de la Petite et Moyenne Industrie (DARPMI), au sein du Ministère Français de l'Économie, des Finances et de l'Industrie. Il a été demandé par le Centre de Mathématiques Appliquées de l'École Nationale Supérieure des Mines de Paris, et n'aurait pu voir le jour sans l'extrême gentillesse et ouverture dont ont fait preuve tous les laboratoires japonais sollicités. Ils en sont grandement remerciés.

Ce rapport commence par examiner les approches développées dans l'analyse du langage écrit, incluant les outils d'analyse automatique des structures

morphologiques et syntaxiques des phrases. Il examine ensuite les stratégies suivies dans le domaine de la représentation sémantique du langage. Dans une troisième partie, il fait un rapide tour d'horizon des principales applications en cours de développement en 1998, puis présente quelques ressources linguistiques largement utilisées. La dernière partie de ce rapport, enfin, donne la liste des laboratoires qui ont accepté de contribuer à ce rapport en invitant son auteur à les visiter.

Cette présentation reste volontairement généraliste, et cherche à fournir au lecteur une vision d'ensemble de la recherche effectuée au Japon en traitement automatique du langage écrit, aux dépens de la précision que l'on attendrait d'un ouvrage scientifique. Cependant, le lecteur intéressé par une question précise évoquée dans ce rapport est vivement invité à consulter les serveurs Internet des laboratoires, dont les adresses sont indiquées dans la dernière partie de cet ouvrage, et qui contiennent souvent des bibliographies. Il peut également prendre contact avec les laboratoires, où bien avec l'auteur de ce rapport à l'adresse suivante :

`jean-philippe.vert@mines.org`

Chapitre 2

Analyse morpho-syntaxique

En préalable à la manipulation de textes écrits en langage naturel, quelle que soit la langue utilisée, il est souvent nécessaire d'effectuer une première analyse de la suite de caractères formant ces textes. Cette analyse doit très grossièrement accomplir les tâches suivantes :

- le regroupement des caractères formant les entités linguistiques de base, à savoir les mots (phase de *segmentation*),
- la reconnaissance de la morphologie de ces entités, souvent grâce à un dictionnaire contenant toutes les formes pouvant apparaître dans la langue utilisée et précisant les informations morphologiques associées à chaque forme (phase d'*analyse morphologique*),
- la reconnaissance de la structure grammaticale des phrases, généralement à travers des formalismes utilisant différents types de grammaires (phase d'*analyse syntaxique*).

2.1 Segmentation et analyse morphologique

Il convient ici de préciser une particularité du japonais, qui est *l'absence d'espaces pour séparer les mots*. La phase de segmentation d'un texte en japonais est d'autant plus cruciale qu'un mauvais regroupement des caractères peut changer complètement le sens d'une phrase.

L'approche la plus répandue pour effectuer la segmentation et l'analyse morphologique du japonais est de combiner ces deux tâches en une seule opération, en se servant d'un lexique contenant la liste des suites élémentaires de caractères admissibles, accompagnés de leurs caractéristiques morphologiques. Ce lexique peut fournir, pour une phrase donnée, une liste des segmentations et caractéristiques morphologiques admissibles, et un ensemble de contraintes est généralement chargé d'éliminer des possibilités de cette liste ou de les classer selon leur caractère plus ou moins probable. Ces contraintes peuvent par exemple être représentées sous la forme de contraintes sur les suites de caractéristiques morphologiques (ex : une terminaison verbale doit

suivre la racine d'un verbe), comme dans le cas du logiciel JUMAN, ou être probabilisées sous la forme de n-grammes portant sur les suites de n caractères, comme dans le cas de l'analyseur du NAIST.

2.1.1 JUMAN : la référence

JUMAN est un analyseur morphologique pour le japonais qui a été développé à l'université de Kyoto sous la direction du professeur Nagao. Lorsqu'une phrase en japonais lui est présentée, il la segmente en morphèmes et indique la classe morphologique de chaque morphème (nom, terminaison verbale...). Pour effectuer cette opération, il s'appuie sur un ensemble de deux dictionnaires :

- *le dictionnaire de la grammaire*, qui contient la liste des classes et sous-classes de morphèmes, un dictionnaire de conjugaison, et une liste des règles de connection entre morphèmes.
- *le dictionnaire des morphèmes*, qui contient la liste des morphèmes du japonais avec différentes informations : leur classe et sous-classe, leur prononciation, leur signification etc...

La force de cet outil est sa *modularité* et son *adaptabilité*. En effet, il existe de nombreux formalismes morphologiques en japonais, de même qu'il existe maintes grammaires différentes et encore plus de dictionnaires de morphèmes. Pour assurer son adaptabilité, JUMAN a été conçu comme un noyau se servant des dictionnaires définis par chaque utilisateur. Et s'il est fourni avec une configuration de base (avec un dictionnaire de 120.000 mots et 14 classes de morphèmes), il peut ainsi être configuré pour n'importe quel formalisme et avec n'importe quel dictionnaire. C'est ainsi qu'il est actuellement utilisé a l'université de Kyoto avec le dictionnaire EDR, contenant 230.000 entrées et 3.000 classes de morphèmes. Lorsqu'une phrase est entrée, JUMAN recherche l'analyse morphologique compatible avec ses dictionnaires et qui contient le moins de mots inconnus, de morphèmes et de mots indépendants.

JUMAN est utilisé par de nombreux laboratoires comme analyseur morphologique du japonais, aussi bien au Japon qu'à l'étranger (au SRI de Stanford par exemple). Il est gratuit et peut être téléchargé électroniquement sur le site du laboratoire du docteur Kurohashi, à l'université de Kyoto, ou sur celui du professeur Matsumoto, au NAIST.

2.1.2 Une approche statistique au NAIST

Plutôt que de baser l'analyse morphologique sur une grammaire prédéfinie, le laboratoire du professeur Matsumoto, au NAIST, a développé une approche statistique pour cette tâche. Pour ce faire, l'analyseur commence par découper une phrase quelconque en mots, grâce à un dictionnaire, en conservant toutes les ambiguïtés possibles. Puis un coût est donné à chaque bigramme, en fonction des fréquences d'apparition calculées sur un corpus.

Enfin, la solution de moindre coût est choisie comme étant la plus probable. Le formalisme utilisé contient 14 catégories morphologiques, divisées en sous-catégories (ex : nom commun, nom de pays, nom de personne...). La matrice de coûts des bigrammes a été en partie remplie à la main, mais une automatisation est en cours à partir du corpus EDR.

2.1.3 Des arbres de décision qui n'utilisent pas de dictionnaire à ATR

Constatant la prolifération de systèmes d'annotations et de dictionnaires associés pour l'analyse morphologique du japonais, l'institut ATR (Advanced Telecommunications Research Institute) propose une analyse morphologique *robuste* basée sur des arbres de décision et n'utilisant pas de dictionnaire, pour palier aux inconvénients suivants :

- les mots rencontrés dans le texte qui ne sont pas dans les dictionnaires ;
- les mots qui sont dans le dictionnaire mais qui n'ont pas l'annotation correcte dans le contexte d'une phrase donnée.

La méthode consiste à introduire des informations aidant à déterminer les limites des mots et les annotations correspondantes. Par exemple, en français, une terminaison en "-er" d'un mot est un indice que le mot peut être un verbe du premier groupe à l'infinitif. En japonais, de nombreuses structures systématiques aident également à repérer les coupures entre les mots (comme par exemple le suffixe "na" à la fin de certains types d'adjectifs). Les informations se présentent sous la forme de questions sur le mot visé ou bien sur les relations avec les caractères qui l'entourent, et peuvent être combinées (ex : "le mot se termine en caractères *hiragana* et le caractère suivant est un *kanji*")

De cette manière, de nombreuses questions peuvent être imaginées, et des arbres de décisions sont créés selon des méthodes classiques, à partir de ces questions, et en utilisant un corpus annoté comme ensemble d'entraînement pour les arbres. Les annotations utilisées dans les expériences contiennent 209 annotations possibles, regroupées en 18 annotations de base (nom commun, nom propre, verbe...).

Une fois entraîné, chaque feuille finale de l'arbre de décision contient une distribution de probabilités sur la structure morphologique (segmentation et annotation) des phrases que l'on traite caractère par caractère.

2.2 Analyse syntaxique

L'analyse syntaxique d'une phrase (*parsing* en anglais) est l'opération qui consiste à assigner une structure à la dite phrase, dans un formalisme syntaxique donné. Les formalismes sont des *grammaires* qui sont censées définir le langage. Alors que la segmentation du japonais se heurte à des obstacles propres à la langue, notamment l'absence d'espace, son analyse

syntaxique se rapproche grandement de l'analyse des langues européennes et utilise les mêmes classes de grammaires.

2.2.1 KNP : un analyseur pour les longues phrases, développé à l'université de Kyoto

L'algorithme KNP (*Kurohashi-Nagao parser*, du nom de ses inventeurs) est un algorithme qui a été développé dans le but d'analyser des longues phrases, tâche traditionnellement très difficile si des relations existent entre des mots éloignés. L'algorithme part de l'hypothèse qu'une longue phrase contient souvent des *structures conjonctives* reliant des groupes nominaux ou des propositions. Par exemple, dans la phrase "Paul a aimé le livre et adoré le film", la conjonction "et" sert à relier les deux propositions qui sont parallèles. Ce genre de structure est très fréquent dans la langue japonaise, et crée de nombreuses ambiguïtés pour les analyseurs syntaxiques.

Pour découvrir ces structures conjonctives, les membres du laboratoire du professeur Nagao, à l'université de Kyoto, ont proposé l'algorithme KNP qui calcule la *similarité entre deux séries arbitraires de mots* à gauche et à droite d'une conjonction, et qui sélectionne les deux séries de mots les plus similaires qui peuvent raisonnablement être considérées comme des éléments d'une structure conjonctive. Cette sélection utilise des techniques de programmation dynamique.

Lorsque plusieurs structures conjonctives ont été découvertes, l'algorithme se charge de "démêler" les relations qui se croisent. Enfin, une analyse syntaxique classique est opérée au sein de chaque bloc, et l'analyse de la phrase entière est obtenue en considérant les structures conjonctives comme des nœuds simples.

La mesure de similarité entre deux blocs de mots, qui est la clé de l'algorithme, utilise des comparaisons entre les catégories morphologiques, les caractères (japonais) utilisés, et les catégories sémantiques des éléments des blocs, obtenues grâce à un thésaurus.

2.2.2 Un analyseur pour l'anglais basé sur des arbres de décision à ATR

Les chercheurs d'ATR ont développé un analyseur syntaxique pour l'anglais, en utilisant la "ATR English Grammar". Cette grammaire contient près de 3000 étiquettes possibles pour les mots, et est basée sur environ 1.100 règles. La structure syntaxique d'une phrase apparaît comme un arbre dont chaque noeud contient la règle de grammaire qui l'a généré, s'il est non-terminal, et l'étiquette du mot concerné, s'il est terminal. Un corpus de textes analysés selon cette grammaire, le "ATR/Lancaster Treebank of General English" (développé avec Lancaster University, au Royaume-Uni), représente un ensemble permettant l'entraînement d'arbres de décision construisant auto-

matiquement l'arbre associé à une phrase donnée. Les questions utilisées dans l'arbre de décision concernent les valeurs attachées aux feuilles des arbres, qu'elles soient terminales ou non, ainsi que des caractéristiques des mots bruts ou des phrases (ex : taille de la phrase).

Une analyse est le résultat d'une succession d'analyses partielles, représentant des états successifs dans le processus de détermination de l'arbre. Le passage d'un état à un autre se produit quand un nouveau nœoeud est étiqueté ou quand il est décidé qu'un nœoeud est terminal. Ces décisions sont précisément prises grâce à des arbres de décision, qui sont entraînés pour les prendre avec les corpus d'entraînement.

2.2.3 L'utilisation du formalisme HPSG à l'université de Tokyo

Le laboratoire du professeur Tsujii, à l'université de Tokyo, travaille sur le développement d'un cadre de travail et d'un environnement basé sur le formalisme de grammaires syntagmatiques guidées par la tête (Head Driven Phrase Structure), afin de préparer le développement de diverses applications en TALN, telles l'acquisition de connaissance, la traduction automatique ou l'extraction d'information.

HPSG est un formalisme encore guère utilisé pour les applications pratiques, car il est trop inefficace par rapport à des approches utilisant des CFG par exemple. Il possède cependant certains avantages, parmi lesquels celui de minimiser le noyau de la grammaire, qui contient seulement 6 règles[1] dans le cas de la grammaire japonaise développée dans le laboratoire, et de maximiser le rôle des informations lexicales.

La grammaire est implémentée dans le langage de programmation LiLFeS, et utilise un algorithme d'analyse de structure en deux étapes :

- un analyseur énumère les arbres de structure possibles, en utilisant les règles de CFG compilées à partir des entrées lexicales du formalisme HPSG ;
- puis l'analyseur résout les contraintes qui ne sont pas prises en compte dans la grammaire CFG.

Divers développements sont en cours pour favoriser l'utilisation du formalisme HPSG :

- création d'automates à états finis ;
- création d'algorithmes parallèles (permettant d'augmenter la vitesse par 10) ;
- outil d'extraction d'information ;
- création automatique d'un thésaurus.

Ce projet de développement du formalisme HPSG donne lieu à de nombreuses collaborations notamment au Japon (EDR...), aux États-Unis (Stan-

[1] Ces schémas correspondent aux règles de réécriture des CFG

ford...) ou en Europe (DFKE en Allemagne...).

2.2.4 POWER, un analyseur orienté objet à l'université de Niigata

Le laboratoire du professeur Miyazaki, à l'université de Niigata, concentre ses efforts depuis plusieurs années sur la réalisation d'un nouveau type d'analyseur syntaxique pour le japonais et l'anglais : POWER. Cet analyseur, écrit en PROLOG, est orienté objet. Dans cette approche, les unités élémentaires avec lesquelles le programme travaille sont des objets, qui sont les mots auxquels sont attachées une classe (ex : nom) et une catégorie (ex :sujet). Suivant leurs caractéristiques, les objets envoient des messages à d'autres objets, qui effectuent certaines actions et répondent. Par exemple, un certain type de verbe va envoyer des messages vers un certain type de sujet ou de complément, et attendre des réponses. Les actions constituent la construction progressive de l'arbre de structure.

POWER doit être appliqué à une phrase en japonais déjà analysée morphologiquement. Les arbres de structure sont classés par "coût" censé représenter leur probabilité de correspondre à une phrase donnée. Un avantage de POWER est sa grande flexibilité quant au formalisme à utiliser et aux actions à effectuer suite à un message.

2.2.5 Un analyseur syntaxique probabilisé a Tokodai

Le laboratoire du professeur Tanaka, au Tokyo Institute of Technology (Tokodai en abrégé), a généralisé la méthode d'analyse syntaxique GLR (Generalized Left-Right) en incluant une estimation de la probabilité des différents arbres obtenus avec cet algorithme. En cas ambiguïté entre plusieurs arbres, le plus probable peut alors être choisi, et une connaissance de la fiabilité d'une analyse pour une phrase donnée est également obtenue. De plus, les arbres trop peu probables peuvent être éliminés en cours de construction, ce qui augmente la vitesse de l'analyse

La méthode est plus précise que l'approche PCFG (Probabilistic Context-Free Grammar), qui consiste à assigner une probabilité à chaque règle de la grammaire. Pour obtenir un modèle légèrement sensible au contexte, les probabilités sont estimées pour les actions de la table LR, utilisée par l'automate pour analyser les phrases et obtenue à partir des règles de la grammaire. La probabilité d'une dérivation est définie comme le produit des probabilités des actions liées à cette dérivation. Les probabilités sont estimées à partir des fréquences des actions effectuées pour analyser un corpus de phrases correctement analysées (en l'occurence un ensemble de 10.000 phrases du corpus d'ATR)

2.3 Analyses combinées

Alors que les phases de décomposition morphologique et d'analyse syntaxique sont souvent séparées, la seconde partant des résultats de la première, certains chercheurs tentent de réconcilier ces deux analyses en une seule, se fondant sur la constatation que des informations de l'une peuvent bien souvent aider l'autre, notamment lors de la phase de *désambiguation*. C'est le cas de l'équipe du professeur Tanaka.

2.3.1 Un analyseur morpho-syntaxique basé sur l'algorithme LR à Tokodai

L'équipe du professeur Tanaka, au Tokyo Institute of Technology, a proposé une méthode d'analyse combinant l'analyse de la morphologie et de la syntaxe en une seule étape, mais en conservant des règles morphologiques et syntaxiques distinctes. Ce point est crucial, tant il est vrai que de nombreuses théories ont déjà été étudiées dans chacun de ces deux domaines.

L'analyse morphologique "classique" utilise un dictionnaire, qui spécifie pour chaque mot ou groupe de caractères sa catégorie morphologique *mcat*, et une matrice de connectivité autorisant ou non une suite de deux *mcat*. Beaucoup d'ambiguïtés subsistent si cette méthode est utilisée seule.

Pour palier à cet inconvénient, l'équipe du professeur Tanaka a proposé de combiner cette approche pour la morphologie avec la méthode d'analyse syntaxique LR, qui part des règles d'une grammaire CFG pour en déduire une matrice LR permettant à un automate d'analyser la phrase de gauche à droite. Pour cela, les règles de la grammaire CFG (qui concernent des catégories syntaxiques *cat*), sont augmentées de règles reliant les *mcat* aux *cat*, grâce au dictionnaire dans lequel chaque mot possède une *cat* et une *mcat*. En général, une *cat* est associée à plusieurs *mcat*. Cette grammaire étendue de manière automatique, considérée comme une grammaire CFG, engendre une matrice LR ayant pour base non plus des *cat*, mais l'ensemble des *cat* et des *mcat*.

La matrice de connectivité est ensuite utilisée pour éliminer les actions de réduction illégales de manière automatique, dans la matrice LR généralisée, ce qui engendre une nouvelle table LR modifiée, travaillant sur les *mcat* et les *cat*, et contenant les contraintes morphologiques de la matrice de connectivité. Un algorithme LR légèrement modifié, pour travailler sur les caractères et non les mots, permet de construire un arbre résumant la structure morphologique et syntaxique de la phrase.

Chapitre 3

Analyse et représentation sémantique

L'analyse sémantique d'un mot, d'une phrase ou d'un texte, consiste à en déterminer le sens. La notion de sens est cependant ambiguë, et doit être comprise au sein d'un formalisme donné, que l'on peut également appeler *représentation sémantique*. Ce sont les différents formalismes utilisés par des laboratoires japonais, souvent pour des applications très concrètes, que le présent chapitre tente de présenter.

Deux grandes familles de représentations sémantiques se dégagent, qui parfois se recoupent :

- une organisation *structurée* des différentes unités sémantiques, appelées *concepts*, souvent sous forme de graphes ;
- une *liste non ordonnée* de concepts, définissant l'espace des connaissances comme un espace vectoriel de grande dimension et permettant d'utiliser différents outils de géométrie euclidienne (produit scalaire, projection orthogonale...)

3.1 L'organisation structurée des concepts

3.1.1 Un dictionnaire de concepts à EDR

Le dictionnaire électronique créé par l'entreprise EDR (Japan Electronic Dictionnary Research Institute, Ltd.) comprend plusieurs dictionnaires, dont un dictionnaire de concepts. Ce dictionnaire décrit les relations entre les 400.000 concepts, ou unités sémantiques de base, qui ont été introduits en tant que sens des mots dans le dictionnaire des mots. Concrètement, un concept est représenté par une suite de 6 caractères alphanumériques, et à chacun des mots des dictionnaires japonais ou anglais sont attachés un ou plusieurs concepts.

Le dictionnaire des concepts est constitué de :

- la liste des concepts, avec leur identification alphanumérique, une illustration sous forme de phrase et une explication de leur sens ;
- la classification des concepts, qui décrit la liste des relations binaires "super-/sub-" existant entre concepts ;
- la description des concepts, qui décrit les relations binaires sémantiques existant entre concepts, telles que agent-action, objet-action, matériaux-objet etc...

L'utilisation de ces dictionnaires en TALN passe par des outils de navigation au sein de graphes, afin de comprendre le sens des phrases et les relations entre les différents mots qui la constituent. La *désambiguation* sémantique, par exemple, peut être effectuée en recherchant les concepts liés aux mots à *désambiguer*, puis en comparant les positions de ces concepts au sein du dictionnaire de concepts.

3.1.2 Des graphes de co-occurence à ETL

Le Docteur Tanaka-Ishii, du laboratoire ETL à Tsukuba (Electrotechnical Laboratory), a travaillé sur la construction de graphes de co-occurence entre mots, afin d'obtenir une représentation sous forme de graphes des concepts représentés par les mots. Cette approche a été utilisée dans le cadre de l'étude de corpus multi-lingues non alignés. En effet, en utilisant un dictionnaire bilingue, et les graphes de co-occurence obtenus avec deux corpus des deux langues considérées, il est possible de rechercher une fonction reliant les deux graphes ; cette fonction peut ensuite être utilisée pour la *désambiguation* des mots nécessaire à la traduction d'une langue à l'autre.

La structure créée dans ce cadre est de nouveau un graphe, mais construit automatiquement à partir d'un corpus. De même que pour le graphe d'EDR, il est supposé représenter des relations valables quelle que soit la langue étudiée.

3.1.3 Regroupement hiérarchique de mots, à ATR

Partant d'une base de donnée textuelle, constituée d'archives du *Wall Street Journal*, des chercheurs de l'institut ATR (Advanced Telecommunication Research Institute), près de Kyoto, ont créé une classification automatique des 70.000 mots les plus employés au sein de classes ordonnées hiérarchiquement. Le résultat de cette classification est un arbre binaire, où les 70.000 feuilles terminales représentent chacun des 70.000 mots, et dans lequel chaque nœud représente une classe de mots (englobant les classes des deux nœuds dérivés).

L'arbre a été construit automatiquement, en partant de 70.000 feuilles isolées, et en regroupant de manière itérative les classes le plus souvent utilisées dans des situations similaires. Chaque nœud de l'arbre binaire peut ensuite être représenté par une suite de bits, codant par la même des concepts

plus ou moins généraux.

Cette construction, de nouveau entièrement automatique, propose ainsi une organisation des concepts sous forme d'arbre binaire ; la navigation au sein de cet arbre ouvre de nouvelles possibilités d'études sémantiques.

3.2 Représentations sémantiques non ordonnées

3.2.1 Une approche vectorielle pour l'extraction d'information à NTT

Le laboratoire "Human Interface" de NTT, situé à Yokasuka, entre Tokyo et Yokohama, a un projet en cours concernant l'extraction automatique du sujet des informations à partir de journaux radiodiffusés en anglais. Il utilise pour ce faire une liste de 70.000 mots décrivant des sujets possibles, cette liste pouvant être considérée comme une base de concepts pour un univers sémantique restreint.

Utilisant des archives d'information, une matrice de distance entre un mot de cette liste et n'importe quel mot anglais fut créée, en comptant les co-occurence entre un mot de la liste apparaissant dans la "headline" et les mots apparaissant dans la dépêche. Cette matrice est alors utilisée pour créer des modèles d'information mutuelle ou de χ^2, afin de définir un score entre chacun des 70.000 mots de la liste et n'importe quel mot anglais. Finalement, le score d'un mot de la liste par rapport à une dépêche quelconque est défini comme la somme (normalisée) des scores entre le mot-clé et les mots composant la dépêche.

Il est ensuite possible de montrer les mots-clés ayant les scores les plus importants par rapport à une dépêche donnée, ceux-ci étant censés représenter le sujet de la dépêche. Le modèle peut également être utilisé dans le sens inverse, afin de faire de la recherche de documents correspondant à un mot-clé donné (en recherchant dans la base de données d'articles les dépêches ayant les scores les plus élevés par rapport au mot-clé choisi)

Sous couvert de régression, afin d'estimer la fonction qui à un article associe son sujet, l'approche décrite ici utilise implicitement une représentation de l'espace des sujets, comparable à l'espace des concepts, comme un espace vectoriel à 70.000 dimensions, dans lequel chaque mot anglais a une représentation (ses coordonnées étant ses "scores"). Des outils de géométrie apparaissent naturellement, comme la définition du vecteur d'une dépêche comme l'équibarycentre (c'est-à-dire la moyenne) des vecteurs correspondant aux mots qui la composent.

3.2.2 Des vecteurs sémantiques à l'université de Shinshu

Le laboratoire des professeurs Nakano, Okamoto et Maruyama, de l'université de Shinshu à Nagano, travaille sur des outils de représentation 3D

pour la recherche de documents sur Internet. L'approche utilise des vecteurs sémantiques, définis comme des vecteurs normés dont les coordonnées sont les poids des mots-clés associés aux vecteurs de base de l'espace euclidien à grande dimension représentant l'espace sémantique. Ces poids eux-mêmes sont calculés à partir des fréquences absolues et relatives d'apparition des mots-clés dans le document considéré, par rapport aux autres documents de la base de données.

Ces vecteurs sémantiques sont utilisés pour représenter, entre autre :
- le contenu (sémantique) d'un document ;
- le point de vue (un centre d'intérêt) de l'utilisateur.

Ils servent également à définir simplement la similarité entre deux documents, comme le produit scalaire de leurs vecteurs sémantiques.

Encore une fois, le passage par des représentations vectorielles permet l'utilisation de techniques simples (projections, produits scalaires...) auxquelles il est possible de donner un sens sémantique.

3.2.3 Une représentation sémantique utilisant des réseaux de neurones à ETL

Le docteur Takahashi, des laboratoires ETL, a proposé une méthode pour représenter le contenu sémantique des mots ou des phrases japonaises par des vecteurs réels, grâce à l'utilisation de réseaux de neurones.

Les vecteurs représentant le sens des mots ou des phrases, appelés *vecteurs de représentation sémantique (VRS)*, ont tous une taille fixée à l'avance et sont obtenus à l'aide de réseaux de neurones du type RAAM (Recursive Auto-Associative Memory) entraînés sur un corpus, de telle manière que des VRS similaires soient donnés à des mots ou phrases similaires. De plus, la transformation d'une phrase en VRS est réversible, ce qui assure que des mots ou phrases distincts ont des VRS distincts.

La représentation sémantique est ainsi obtenue dans un espace euclidien, de dimension fixée à l'avance, ce qui permet d'utiliser les outils classiques de calcul vectoriel.

3.3 Approches mixtes

3.3.1 La classification automatique de documents, à l'université de Tokushima

Le laboratoire du professeur Aoe, de l'université de Tokushima (située sur l'île de Shikoku), a développé un système, bientôt commercialisé, de classification automatique de textes en japonais situés dans un répertoire informatique, en fonction de leur contenu. Cet outil utilise une approche en deux étapes :
- recherche des *mots-clés* de l'article ;

– classification des textes grâce à une hiérarchie de concepts.

La recherche des mots-clés d'un texte s'effectue en comptant les fréquences absolues et relatives d'une liste de 40.000 bigrammes (suite de deux caractères japonais), permettant d'extraire ceux qui caractérisent le plus le texte considéré. Cette première étape est donc caractéristique des représentations sémantiques vectorielles non ordonnées.

La seconde phase, par contre, utilise une hiérarchie sémantique préexistante sur les mots-clés, afin de classer les textes eux-mêmes selon une hiérarchie sémantique. Cet phase s'apparente donc aux approches évoquées précédemment, utilisant des représentations sémantiques ordonnées.

La combinaison des deux types d'approches, dans le but ici de classer une base de données textuelles en fonction de critères sémantiques, présente l'intérêt de combiner les avantages calculatoires de la méthode par représentation vectorielle non ordonnée aux avantages sémantiques - notamment pour les relations entre les mots-clés - des approches utilisant des structures de représentation sémantique ordonnée.

3.3.2 Mesure de similarité entre mots au laboratoire CS de NTT

Le laboratoire CS (Communication Science Laboratory) de NTT, situé près de l'ancienne capitale impériale Nara, a développé une hiérarchie d'environ 3.000 concepts sémantiques reliés par des relations "has-a" et "is-a". Si elle apparaît beaucoup plus petite que le dictionnaire de 400.000 concepts développé par EDR, cette base de données hiérarchisée se veut également plus robuste que sa grande sœur.

Cette base de concepts, appelée "base de connaissances", est utilisée comme base d'un espace euclidien à 3.000 dimensions dans lequel est représenté l'ensemble des mots japonais, réduit à un ensemble de 40.000 mots-concepts par standardisation, grâce notamment au thésaurus de NTT qui utilise lui-même la classification hiérarchique des 3.000 concepts de base. Le résultat de ces opérations est une matrice de 3.000 concepts sur 40.000 mots-concepts, où les coordonnées des mots-concepts sont normalisées.

Cette construction permet de calculer la similarité entre deux mots selon un certain point de vue. Par exemple, selon le point de vue "animal", le mot "cheval" sera plus près de "lapin" que de "voiture", mais le résultat sera contraire selon le point de vue "moyen de transport". Pour inclure le point de vue dans la mesure de similarité, des opérations vectorielles sont introduites, principalement des projections (pour projeter selon un point de vue) et des calculs de produits scalaires (pour mesurer des similarités).

Même si cette réalisation s'apparente aux approches purement vectorielles, il convient de remarquer que la base de l'espace des connaissances est constituée de concepts placés précisément dans un graphe hiérarchique, ce qui ouvre la voie à d'éventuelles utilisations de cette base comme moyen

d'analyse plus poussée.

Chapitre 4

Applications

Les analyses morphologiques, syntaxiques et sémantiques des textes, qui ont été présentées dans les parties précédentes, ne constituent en général pas des fins en elles-mêmes, mais plutôt des préalables nécessaires à diverses applications. L'objet de ce chapitre est de présenter un tour d'horizon des applications du TALN telles qu'elles sont apparues lors de ce voyage d'étude.

Les équipes japonaises se penchent globalement sur les mêmes problèmes que leurs homologues américaines ou européennes, à savoir les sujets qui permettront aux ordinateurs de réellement apporter une plus-value par rapport aux humains dans le traitement et la manipulation de documents textuels. Parmi ces sujets, au centre de formidables enjeux économiques, on retrouve :

- le recherche et l'extraction d'information dans des grosses bases de données ;
- la classification automatique de documents ;
- l'alignement automatique entre diverses sources (journaux, émissions télévisées ou radiodiffusées...) ;
- la *désambiguation* pour la reconnaissance de la parole ou de l'écriture ;
- le résumé automatique de texte ;
- la traduction automatique ;
- les interfaces homme/machine adaptées à l'homme.

4.1 Recherche et extraction d'information

La recherche d'information (en anglais Information Retrieval) dans des grosses bases de données est une discipline dans laquelle la compétition est rude et les progrès rapides. Cette discipline a connu un tournant avec la création, sous la tutelle de la DARPA (Defense Advanced Research Projects Agency) et du NIST (National Institute of Standards and Technology) aux États-Unis, de la compétition annuelle TREC (Text REtrieval Conference) à laquelle ont participé 51 groupes de recherche de 12 pays différents en 1997. Un seul groupe japonais a participé à cette compétition (NEC), mais des

chercheurs de divers instituts ont participé aux travaux de cette équipe. Il est prévu qu'une compétition similaire soit organisée au Japon à partir de 1999, avec des recherches sur des documents en langue japonaise.

4.1.1 Expansion de requête à Tokodai

Le laboratoire du professeur Tokunaga, au Tokyo Institute of Technology (Tokodai), travaille sur une méthode d'expansion de requête. Lorsqu'un utilisateur recherche des documents dans une base de donnée, il exprime son désir sous forme de requête, c'est-à-dire d'une suite de mots. L'expansion de requête consiste à utiliser des moteurs de recherche existant, mais à compléter la requête de l'utilisateur avec de nouveaux mots avant d'envoyer cette requête au moteur de recherche.

La méthode proposée par le laboratoire utilise la base de donnée *Wordnet*, développée à partir de 1985 à Princeton University sous la direction du professeur George Miller. Cette source de connaissance lexicale est utilisée dans beaucoup d'applications du TALN, mais ne s'est jamais avérée très efficace pour la recherche d'information. La méthode consiste à rechercher des mots similaires aux mots de la requête, la similarité étant détectée à partir de trois sources d'informations :

- un thésaurus, construit automatiquement à partir des co-occurences de mots ;
- un thésaurus, construit automatiquement à partir de données du type "prédicat - argument" observées sur un corpus syntaxiquement décomposé ;
- la base de donnée Wordnet.

Une expérience d'utilisation de cette méthode en combinaison avec le moteur de recherche *SMART* de Cornell University a démontré la complémentarité de ces trois sources lexicales, et l'intérêt de les combiner.

4.1.2 Organisation géographique des connaissances à NTT

Le laboratoire NTT Software Laboratory, situé à Musashino, dans la banlieue est de Tokyo, met en place un outil de recherche d'information en fonction de la position géographique de l'utilisateur. Cet outil est accessible à l'URL http ://www.kokono.net et prétend fournir un guide des institutions, hôtels, restaurants, boutiques etc... situés à proximité de l'utilisateur, en utilisant les informations disponibles sur Internet. Pour ce faire, le moteur de recherche doit d'une part repérer la position géographique de l'utilisateur, par exemple grâce à la localisation satellite des téléphones portables, et d'autre part rechercher sur Internet les informations géographiquement proches de l'utilisateur.

La méthode choisie consiste à organiser géographiquement l'information sur Internet, soit en utilisant des bases de données comme les pages jaunes,

pour lesquelles les informations géographiques sont facilement accessibles, soit en recherchant des informations de localisation au sein des archives textuelles d'Internet, telles des adresses ou des ZIP codes.

Cette recherche d'information, ici limitée aux informations géographiques, permet ainsi la réorganisation d'une grosse base de donnée textuelle (Internet dans le cas présent) selon un critère original : la situation géographique. Le site http ://www.kokono.net, toujours expérimental en juillet 1998, devrait prochainement devenir commercial.

4.1.3 Une classification 5W1H à NEC

Le groupe de *Pattern Analysis and Human Language Technology* de l'entreprise NEC, géant de l'informatique et de la communication, a développé un moteur de navigation dans des bases de données textuelles à partir de requête sous la forme des six questions élémentaires 5W1H (who, when, where, what, why, how). Dans la phase d'organisation de l'information au sein de la base de données, le programme extrait un vecteur à six dimensions, contenant des informations relatives aux six questions 5W1H, pour chaque phrase. Cette extraction d'information utilise des techniques de TALN et de pattern matching, afin d'identifier les informations intéressantes.

La phase de navigation découle naturellement de l'organisation des informations, en ce sens que l'utilisateur doit remplir un ou plusieurs champs du questionnaire 5W1H, puis le navigateur recherche les documents répondant aux critères sélectionnés. Les démonstrations des applications actuelles concernent des dépêches économiques, pour lesquelles le formalisme 5W1H est particulièrement bien adapté.

4.1.4 Extraction d'information utilisant des *template matching* à Kyutech

Le laboratoire du professeur Nomura, au Kyushu Institute of Technology, a développé un outil permettant l'extraction automatique d'information à partir de dépêches diverses. Une démonstration disponible sur Internet en juillet 1998 utilisait par exemple une base de données de 2.000 dépêches concernant la commercialisation de différents types de biens, pour lesquels des informations pouvaient être extraites automatiquement : le type de bien, son nom, le nom du fabricant, le prix, la date de sortie etc...

L'extraction des informations commence par l'analyse morphologique des dépêches à l'aide de l'analyseur JUMAN de l'université de Kyoto (voir 2.1.1). Les informations de surface données par cette analyse sont ensuite utilisées par le moteur de recherche d'information à l'aide de *template matching*. En d'autres termes, chaque type d'information recherché est caractérisé par un ensemble de propriétés (les *templates*) portant sur les environnements linguistiques de l'information (par exemple, le type de particule qui précède une

date) et sur l'information elle-même. Le moteur de recherche repère les candidats à chaque type d'information par une procédure classique de *template matching*. Le système applique pour cela chaque template à chaque phrase de la dépêche étudiée. Chaque phrase produit alors des candidats pour l'information recherchée, et le système choisit les candidats les plus probables, en prenant en compte le nombre de *templates* qui ont extrait chaque candidat, et le poids à donner à chaque candidats (dépendant du nombre de phrases qui l'ont extrait).

Un ensemble de 3.840 *templates* fut initialement créé à la main, afin de caractériser chaque information à extraire. Grâce à une procédure de réduction automatique à partir de l'entraînement du système sur un corpus, ce nombre est ensuite passé à 1.403. La précision des résultats est supérieure à 90% sur la base de données proposées en démonstration, pour les différentes informations à extraire.

4.1.5 Utilisation de la base de concept à NTT

Dans le Communication Science Laboratory de NTT, un projet de recherche d'information est en cours, en collaboration avec l'université de Stanford et le Stanford Japan Center. Ce système est basé sur la base de concept (voir 3.3.2) qui permet de mesurer la similarité entre deux mots, et de regrouper les textes selon leur contenu.

La base de concept, qui donne des informations sur le sens des mots, contient environ 20.000 concepts. Elle permet de représenter un ensemble de mots (par exemple une requête) par un vecteur dans l'espace des concepts à 20.000 dimensions. La similarité entre deux mots est définie comme le produit scalaire des vecteurs les représentant dans l'espace des concepts, et le vecteur d'un ensemble de mots est défini comme la moyenne entre les vecteurs des différents mots. Inversement, les mots caractéristiques d'un vecteur sont définis comme ceux dont la représentation vectorielle est proche (au sens de la norme euclidienne) du vecteur considéré.

Ces définitions permettent par exemple d'étendre une requête en ajoutant les mots caractéristiques de sa représentation vectorielle. D'autre part, les articles ayant des représentations vectorielles proches peuvent être regroupés, et les mots caractéristiques peuvent en être extraits. Cette méthode permet en particulier de retrouver des textes ne contenant pas explicitement les termes de la requête initiale, mais conceptuellement proches.

Alors que seule une version japonaise était disponible lors de la visite au laboratoire de NTT, une version anglaise a été développée à l'université de Stanford.

4.2 Classification de documents

La classification automatique de documents vise à regrouper les documents d'une base de données textuelles quelconque en fonction de leur contenu.

4.2.1 Recherche de kanjis à l'université de Kyoto

Le laboratoire du docteur Kurohashi, de l'université de Kyoto, a proposé une méthode de classification de documents japonais sans passer par l'analyse morphologique des documents, mais en observant directement les caractères constituant les phrases[1].

Pour ce faire, une base de données de textes classés par thèmes (philosophie, botanique etc...) est utilisée pour extraire les *kanjis* caractéristiques de chaque thème par une méthode du χ^2. Ces kanjis caractéristiques sont ensuite utilisés pour classer automatiquement tout nouveau texte en fonction des caractères utilisés.

4.2.2 Classification automatique de documents à l'université de Tokushima

Le laboratoire du professeur Aoe, à l'université de Tokushima, a développé un système de classification automatique des documents présents dans un répertoire informatique. Ce système, encore en développement en juillet 1998 mais promis à une commercialisation prochaine, est censé être utilisable pour l'organisation automatique des répertoires sur les ordinateurs personnels ou sur Internet, supporte une utilisation bilingue anglais-japonais, et fournit de plus une phrase explicative pour chaque classe de document, pouvant être utilisée comme base de résumé automatique.

Ce système passe par la recherche de mot-clés (voir 3.3.1) et par l'utilisation de la hiérarchie de concepts associée, afin de classer les textes selon leur contenu.

4.3 Alignement de différents médias à l'université de Kyoto

Une application multi-média du TALN au sens étymologique du terme, c'est-à-dire lorsque plusieurs médias interagissent, consiste à aligner automatiquement les émissions de différents médias en fonction de leur contenu. Il s'agit par exemple de repérer quand une information particulière est traitée à la télévision, à la radio et dans les journaux.

[1] Les caractères japonais appelés *kanji* conservent souvent un sens intrinsèque, même lorsqu'ils sont assemblés pour former un mot. L'examen des caractères constituant un texte donne donc des informations sémantiques

Le laboratoire du docteur Kurohashi, à l'université de Kyoto, a développé un système d'alignement automatique entre les informations télévisées et les journaux écrits. Pour ce faire, le système détecte les mots qui apparaissent au sein des deux médias, et multiplie leurs fréquences par un poids qui dépend de l'endroit où ils apparaissent (dans le titre, le sous-titre, le premier paragraphe etc...). L'alignement s'opère alors entre les articles et parties de programmes ayant le plus d'affinités selon cette méthode.

4.4 *Désambiguation* de caractères

La *désambiguation* de caractères est grandement utilisée dans les logiciels de reconnaissance automatique de l'écriture (O.C.R.), et consiste à choisir le candidat le plus probable parmi une liste de caractères pouvant correspondre à un caractère écrit donné. La méthode la plus répandue pour effectuer cette tâche consiste à construire un modèle simplifié du langage, fournissant la probabilité d'apparition d'un caractère conditionnellement aux caractères qui le précèdent, en approximant de telles probabilités par des modèles de n-grammes, où n est généralement égal à 2 ou 3.

4.4.1 Un modèle de n-grammes performant à l'université de Kyoto

Pour estimer les probabilités découlant d'un modèle de n-gramme à partir d'un corpus, où n est quelconque, un comptage des fréquences d'apparition de chaque n-gramme dans le corpus est nécessaire. Le laboratoire du docteur Kurohashi, à l'université de Kyoto, utilise une méthode simple pour obtenir ces estimations rapidement pour n quelconque.

La méthode consiste à utiliser des pointeurs qui pointent sur les caractères du corpus. Il y a donc autant de pointeurs que de caractères dans le corpus (pour obtenir un modèle de n-grammes sur les caractères). Cet ensemble de pointeurs est ensuite classé selon un ordre lexicographique, en utilisant la suite de caractères commençant à l'endroit pointé pour caractériser chaque pointeur. Une fois ce classement effectué, il est aisé de compter le nombre de fois qu'une suite de n caractères apparaît dans le corpus, en repérant cette suite de caractères dans la liste ordonnée des pointeurs, et en comptant simplement combien de pointeurs commencent par cette suite de caractères.

4.4.2 Un mélange de n-grammes à l'université de Tohoku

Le laboratoire du professeur Aso, de l'université de Tohoku dans la ville de Sendai, s'est spécialisé dans la digitalisation intelligente de documents papier. En particulier, il a mis au point un système de reconnaissance de caractères, incluant une partie de *désambiguation*.

Cette méthode de *désambiguation* utilise des n-grammes comme approximation de modèles du langage, mais a l'originalité de combiner des n-grammes pour $n = 0, 1, 2, 3$. Le modèle final apparaît comme une combinaison linéaire de ces différents modèles, où les poids de chaque modèle prennent en compte la taille du corpus qui a servi à les entraîner. Cette approche permet d'obtenir des approximations même quand le corpus d'entraînement est relativement petit, car les poids sont estimés afin d'optimiser la consistance du modèle final.

4.5 Résumé automatique

Le résumé automatique de texte est une application typique du TALN qui permettrait d'augmenter considérablement la productivité du travail humain grâce aux machines, en remplaçant la lecture de quantités de documents par la lecture de leurs résumés.

4.5.1 Le système TESS de Kyutech

Le laboratoire du professeur Nomura, au Kyushu Institute of Technology, a développé un logiciel de résumé automatique baptisé TESS (TExt Summarization System). La stratégie déployée pour résumer un texte est composée des étapes suivantes :

- coller une étiquette à chaque phrase ;
- évaluer l'importance de chaque phrase ;
- éliminer les phrases sans importance pour le résumé ;
- produire le résumé.

Les étiquettes qui sont collées aux phrases, lors de la première phase, contiennent des informations linguistiques (question, opinion, expression d'un désir, d'un jugement etc...) qui sont en grande partie déterminées grâce à la structure des fins de phrases en japonais, ainsi que des informations sur les relations entre phrases (addition, parallélisme, contradiction etc...) grâce à diverses informations linguistiques (anaphores, conjonctions etc...). L'importance des phrases est quantifiée grâce à ces étiquettes, et le système conserve les phrases jugées importantes. Les phrases ainsi sélectionnées sont finalement réarrangées afin d'obtenir un texte résumé en langage correct (réarrangement des pronoms, division des phrases complexes etc...).

4.6 Traduction automatique

La traduction automatique est une des premières tâches qui ont été assignées comme application du TALN dès les années 1940, et s'avère être une de celles qui résistent le plus aux avancées de la discipline. Les résultats de

traduction automatique restent aujourd'hui très pauvres par rapport aux traductions humaines, et souvent inutilisables. En même temps, cette discipline a des enjeux économiques colossaux, surtout dans les pays non-anglophones à l'ère du développement d'Internet. Les investissements japonais pour le développement de cette discipline sont très importants, et les logiciels de traduction automatique produits par tous les grands constructeurs d'ordinateurs se succèdent à un rythme élevé dans les rayons des boutiques spécialisées du quartier d'Akihabara à Tokyo.

4.6.1 La méthode PIVOT de NEC

En tant que géant de l'informatique, NEC développe bien entendu son propre système de traduction, basé sur un algorithme baptisé PIVOT. Commercialisé sous le nom de *Honyaku Adaptor II*, la version grand public du système de traduction de NEC est également basée sur la méthode du pivot, qui consiste à utiliser *Interlingua*, un formalisme de représentation sémantique universelle.

Chaque phrase à traduire est analysée morphologiquement, puis syntaxiquement, puis enfin sémantiquement, afin d'aboutir à une représentation de la phrase sous forme *Interlingua*[2]. Une fois mise sous forme de représentation *Interlingua*, le processus est réitéré dans l'ordre inverse, mais vers une autre langue.

Face à l'imperfection des résultats globaux de traduction, l'accent est plutôt mis sur l'assistance à l'écriture en anglais (pour les japonais) sous forme d'aide interactive, ainsi que sur l'ergonomie pour les utilisations pratiques de la traduction assistée.

4.6.2 Le système ALT-JE de NTT

Le laboratoire CS de NTT (Communication Science Laboratory) a également développé un système de traduction automatique, baptisé ALT-JE (Automatic Language Translation - Japanese to English), encore à l'état expérimental en juillet 1998.

La méthode employée est une traduction multi-niveaux, basée sur une analyse morphologique poussée du japonais. La phrase à traduire est séparée entre la partie *subjective* (information temporelle et modale) et la partie *objective* (le noyau de la phrase). La partie *objective* est traduite par l'intermédiaire de la méthode à plusieurs niveaux, où les règles larges sont appliquées en premier (comme le transfert des arbres de structure morphologique), suivies des expressions idiomatiques, des structures trouvées dans le dictionnaire de structures sémantiques, puis finalement des règles générales par défaut.

[2]Interlingua est une représentation des concepts reliés par 49 relations possibles, du genre agent-instrument, etc...

Le système ALT-JE utilise la hiérarchie de concepts évoquée dans la partie 3.3.2, ainsi qu'un dictionnaire de mots, une liste des structures communes sémantiques et des idiomes au sein d'un dictionnaire pour le transfert des structures d'une langue à l'autre, et un dictionnaire bilingue japonais-anglais décrivant la syntaxe et la sémantique de chaque mot.

Ce logiciel de traduction devrait être commercialisé au grand public dès l'automne 1998.

4.6.3 Le système ATR MATRIX

Le centre de recherche ATR (Advanced Telecommunications Research Institute) a lui aussi développé un système de traduction automatique, dénommé MATRIX (Multilingual Automatic TRanslation system for Information eXchange). La particularité de ce système est qu'il est conçu pour la traduction orale du langage oral.

Il inclut de ce fait un module de reconnaissance vocale (faisant de la reconnaissance en temps réel), un module de traduction, et un module de synthèse vocale (incluant la possibilité de différentes personnalités). Le système expérimental reconnaît environ 3,000 mots, et peut en traduire 10.000. La phase de traduction mélange une approche par règles et par exemples, et utilise un dictionnaire pour le transfert des structures entre langues. Elle s'appuie également sur une série d'exemples de traduction, qui sont utilisés par similarité.

4.7 Interface homme/machine

Un des défis des décennies à venir est sans conteste la création d'interfaces homme-machine conviviales. Le TALN a certainement un grand rôle à jouer, dans la mesure où la communication en langage naturel entre l'homme et la machine permettrait de franchir un grand pas dans cette recherche de convivialité. Diverses expériences sont d'ores et déjà menées pour tenter de rendre la machine plus "humaine".

4.7.1 Un système de dialogue homme/machine à NTT

Le laboratoire de recherche fondamentale de NTT (Basic Research Laboratory), situé à Atsugi, dans la banlieue de Tokyo, travaille sur le développement de systèmes de dialogue entre l'homme et la machine, incluant la représentation d'un visage sur l'écran de l'ordinateur pour le rendre plus "humain". L'innovation principale du système en cours de développement est qu'il utilise ce que dit l'utilisateur humain avant que celui-ci ne termine de parler, pour émettre des sons, parler ou modifier l'expression de son visage. La communication homme-machine se rapproche alors d'un véritable

dialogue, où il est courant que les interlocuteurs se coupent la parole mutuellement.

Le modèle linguistique utilisé pour ces expérimentations reste simple, puisque le vocabulaire est limité à une cinquantaine de mots et la grammaire est simplifiée, du moins pendant la phase de mise au point.

D'autres sujets de recherche, également étudiés au BRL de NTT, vont dans la direction d'interfaces nouvelles, comme par exemple le contrôle de l'ordinateur avec les yeux, le curseur se plaçant à l'endroit visé par l'oeil de l'utilisateur.

4.7.2 La compétition DIALEAGUE organisée par ETL

L'évaluation de systèmes de dialogue automatique est très difficile, car il ne s'agit pas d'un problème possédant une unique bonne réponse. Le groupe du docteur Hasida d'ETL (Electrotechnical Laboratories) organise une compétition entre les différents systèmes de dialogue[3]. Cette épreuve consiste à faire dialoguer des systèmes de dialogue automatique avec des volontaires (participant via Internet) afin de leur faire trouver un chemin à suivre sur un plan de métro dont chacun ne possède qu'une partie. Le critère jugeant les bons dialogues est le nombre de mots utilisés pour parvenir à la solution.

Plus de 700 internautes ont participé à la dernière édition de ce tournoi, en 1997, ce qui a permis d'une part d'obtenir une classification des systèmes de dialogue automatique, et d'autre part de repérer les utilisateurs (humains) capables le mieux de tester les systèmes informatiques.

4.7.3 Le système IRENA à Kyutech

Développé en 1997 pour la réservation automatique de tickets, IRENA est un système de dialogue créé par le laboratoire du professeur Nomura, au Kyushu Institute of Technology (Kyutech). Ce programme tente d'analyser les phrases parlées émises par l'utilisateur, et de gérer les réponses appropriées dans le cadre de la réservation de tickets. Cet objectif nécessite entre autre de gérer la navigation dans le dialogue, de comprendre les requêtes, et de rechercher les réponses dans une base de données.

La navigation dans la dialogue est gérée grâce à l'utilisation de *dialog frames*. Pour permettre la *désambiguation* des requêtes, le logiciel travaille en logique floue, aussi bien pour les ambiguïtés de langage que pour les ambiguïtés de logique. Ceci lui permet de générer des questions permettant d'affiner la requête. Pour ce faire, une fonction floue est attribuée à chaque expression linguistique ("environ", "à partir de" etc...), et une intégration floue permet de représenter la requête totale.

[3]Cette compétition est décrite en japonais à l'URL http ://www.etl.go.jp/etl/nl/dialeague

L'ordre des expressions, les choses dites spontanément, les refus, et les expressions linguistiques diverses, sont autant de critères qui modifient la fonction floue finale.

Chapitre 5

Ressources linguistiques

Cette dernière partie recense les principales ressources linguistiques (dictionnaires et corpus) développées dans les différents laboratoires et largement évoquées dans les parties précédentes.

5.1 Les productions d'EDR

En avril 1986, le Japan Electronic Dictionnary Research Institute (E.D.R.) fut créé, dans le but de réaliser un dictionnaire électronique utilisable pour la recherche avancée en TALN. Pour mener à bien ses recherches, cette entreprise reçut des fonds du Japan Key Technology Center et de huit entreprises productrices d'ordinateurs : Fujitsu, NEC, Hitachi, Sharp, Toshiba, Oki Electric, Mitsubishi Electric, Matsushita Electric. Le projet, s'étalant sur une période de 9 ans, entre 1986 et 1994, permit de créer un ensemble de cinq dictionnaires, utilisables indépendemment :

- *Dictionnaire japonais* Il s'agit d'un dictionnaire de 250.000 mots, contenant pour chaque mot des informations morphologiques (prononciation, accent, etc...), des informations syntaxiques (caractérisation grammaticale, aspect, ...) et des informations sémantiques (explication du sens et lien avec tous les concepts concernés).
- *Dictionnaire anglais* Reprenant la philosophie du dictionnaire japonais, il contient 190.000 mots, et définit pour chacun d'eux les concepts qu'on peut lui attribuer ainsi que des informations morphologiques (inflection, adjacence, prononciation, accent), syntaxiques (POS, dénombrabilité...) et sémantiques.
- *Dictionnaire technique* Spécialisé en traitement de l'information, il contient 120.000 mots japonais et 90.000 mots anglais.
- *Dictionnaire de concepts* Ce dictionnaire original décrit et classe l'ensemble des 400.000 concepts qui ont été définis pour comprendre le sens de chaque mot. La classification utilise des relations super/supra. La description contient des relations sémantiques binaires entre concepts,

telles que agent/action, objet/action etc...
- *Dictionnaire bilingue*
- *Co-occurence* Cette table contient des informations sur l'acceptabilité ou non de combinaisons de mots dans les phrases, et sur les collocations binaires de concepts.
- *Corpus japonais et anglais* Ce corpus contient 220.000 phrases en japonais et 160.000 phrases en anglais. Pour chacune de ces phrases, les informations morphologiques, syntaxiques et sémantiques sont précisés.

Ces dictionnaires ont par exemple été utilisés dans la mise au point du système d'analyse morphologique JUMAN, développé à l'université de Kyoto et faisant aujourd'hui référence.

Depuis 1996, EDR a rejoint le ANSI Ad-Hoc Group for Ontology Standards, et travaille dans le but de relier EDR et Worldnet.

En juillet 1998, le prix du dictionnaire était de 100.000 JPY (environ 5.000 FF) pour les universités et 1.200.000 JPY (environ 60.000 FF) pour les entreprises.

5.2 Les productions de NTT

Le géant des télécommunications a produit des dictionnaires et des corpus pour ses recherches en NLP, et plus particulièrement en traduction automatique. Ces ressources, produites entièrement à la main, ont ensuite été mises à la disposition de nombreux centres de recherche.

Le dictionnaire proposé par NTT contient 400.000 mots. Pour chacun d'eux, la prononciation, la forme canonique, ainsi que des informations syntaxiques et sémantiques sont fournies. Les informations sémantiques utilisent un graphe hiérarchique contenant 3.000 attributs sémantiques, classés grâce à des relations du type "is a" ou "has a". Pour chaque mot du dictionnaire, les attributs sémantiques correspondants sont précisés.

Parallèlement à ce dictionnaire japonais, NTT a développé un dictionnaire bilingue japonais/anglais de structures classiques et d'idiomes, contenant 17.000 entrées dont 6.000 verbes ambigus. Ce dictionnaire contient les équivalences entre structures japonaises et anglaises.

Enfin, un dictionnaire japonais/anglais, contenant les informations syntaxiques et sémantiques, a été développé.

5.3 Le corpus de l'université de Kyoto

Un projet de réalisation de corpus est en œuvre à l'université de Kyoto. Le but de ce projet est de créer de manière semi-automatique un corpus de textes japonais analysés grammaticalement, tout en améliorant les outils d'analyse automatique.

Le corpus était composé de 20.000 phrases, en juillet 1998. Les phrases sont décomposées grammaticalement de manière automatique, en utilisant l'analyseur morphologique JUMAN et le logiciel de décomposition grammaticale KNP. Chaque phrase est vérifiée et éventuellement modifiée par l'homme, les erreurs détectées étant utilisées pour l'amélioration des algorithmes d'analyse. Le corpus augmente à une vitesse d'environ 40 phrases par heure et par personne.

Ce corpus peut être téléchargé par Internet sur le site de l'université de Kyoto, mais il faut en plus acheter le CD-Rom du journal utilisé.

5.4 Les ressources d'ATR

Les laboratoires d'ATR ont développé deux types de ressources pour l'utilisation en recherche sur le traitement automatique du langage. Il est à noter que ces ressources sont en anglais.

- *Regroupement hiérarchique de mots* Partant d'une base de données en anglais (des archives du Wall Street Journal), les chercheurs d'ATR ont créé une classification des 70.000 mots les plus employés au sein de classes ordonnées hiérarchiquement. Ils ont obtenu un arbre binaire, où les 70.000 feuilles finales représentent ces 70.000 mots, et où chaque nœud représente une classe. L'arbre a été construit automatiquement, principalement en regroupant les classes qui sont le plus souvent utilisées dans des situations similaires, de manière itérative. L'algorithme utilise en particulier la notion d'information mutuelle, définie par :

$$MI(a,b) = P(a,b)\log\frac{P(a,b)}{P(a)P(b)}$$

  Cet arbre binaire, où chaque classe peut être représentée sous forme de bits, constitue une organisation sémantique utilisable pour diverses applications.
- *ATR/Lancaster Treebank* Ce corpus est constitué de 730.000 mots, organisés en 950 documents ayant des longueurs comprises entre 30 et 950 mots. Un degré maximum de variation a été voulu dans les longueurs des textes, leurs sujets ou leurs auteurs. Ces textes sont tous annotés avec le système d'annotation utilisé par ATR. Les phrases sont décomposées selon la grammaire anglaise d'ATR, qui est une "feature-based context-free phrase-structure grammar". Elle contient 67 caractéristiques et 1100 règles.

5.5 Le projet GDA à ETL

Le centre de recherche d'ETL à Tsukuba, qui a développé l'environnement multilingue MULE disponible sur les versions 20 de GNU Emacs, tente de

promouvoir un standard d'annotation pour les documents HTML publiés sur Internet : le système Global Document Annotation (GDA). Il permettrait aux machines de reconnaître automatiquement les structures sémantiques et pragmatiques du document. Les initiateurs du projet espèrent qu'une quantité importante de données annotées vont peu à peu apparaître, pouvant en particulier servir de corpus linguistique. Pour promouvoir ce nouveau standard, les initiateurs du projet ont proposé une collection d'annotations permettant aux ordinateurs de deviner les structures du document, et ont développé des applications censées favoriser l'emploi de ce système, comme de la traduction automatique, du data mining, des résumés automatiques ou des présentations automatiques à partir d'un unique document.

Chapitre 6

Les laboratoires

Le présent chapitre détaille la liste des laboratoires qui ont accepté de contribuer à ce rapport en ouvrant leurs portes et en présentant certains de leurs thèmes de recherche.

6.1 Advanced Telecommunications Research Institute International (ATR)

Coordonnées du laboratoire
ATR Interpreting Telecommunications Research Laboratories

Site Internet

`http://www.itl.atr.co.jp/`

Quelques thèmes de recherche
Traduction automatique, analyseur syntaxique, corpus arboré.

6.2 Electrotechnical Laboratory (ETL)

Coordonnées du laboratoire
Ministry of International Trade and Industry (MITI)
Electrotechnical Laboratory
Natural Language Learning Laboratory

Site Internet

`http://www.etl.go.jp`

Quelques thèmes de recherche
Systèmes de dialogue, représentation sémantique, standard d'annotation GDA.

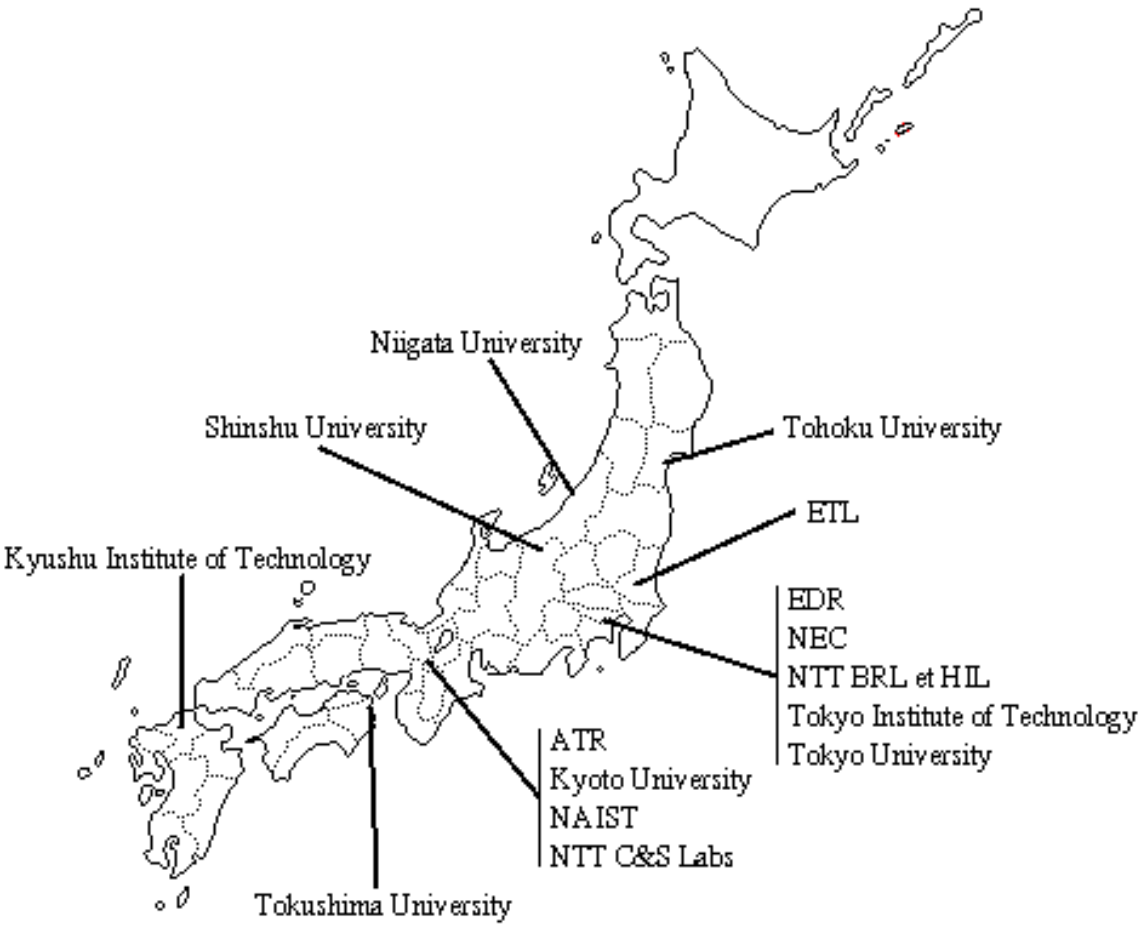

FIG. 6.1: Localisation des laboratoires

6.3 Japan Electronic Dictionnary Research Institute, Ltd (EDR)

Coordonnées du laboratoire

Japan Electronic Dictionnary Research Institute, Ltd

Site Internet

`http://www.iijnet.or.jp/edr/`

Quelques thèmes de recherche

Dictionnaires, représentation sémantique et corpus.

6.4 Kyoto University

Coordonnées du laboratoire

Language Media Laboratory (Nagao/Kurohashi Lab)
School of Electrical and Electronic Engineering
Faculty of Engineering
Kyoto University

Site Internet

`http://www-nagao.kuee.kyoto-u.ac.jp/index-e.html`

Quelques thèmes de recherche

Analyse morpho-syntaxique, corpus, génération de texte, statistiques pour analyse de texte et recherche d'information.

6.5 Kyushu Institute of Technology (Kyutech)

Coordonnées du laboratoire

Nomura and Nakamura Laboratory
Department of Artificial Intelligence
Faculty of Information Engineering
Kyushu Institute of Technology

Site Internet

`http://www.dumbo.ai.kyutech.ac.jp/htdocs/nomura-ken/nomura-ken-e.html`

Quelques thèmes de recherche

Extraction d'information, résumé, système de dialogue

6.6 Nara Institute of Science and Technology (NAIST)

Coordonnées du laboratoire

Matsumoto Laboratory (Computational Linguistic Laboratory)
Graduate School of Information Science
Nara Institute of Science and Technology

Site Internet

`http://cactus.aist-nara.ac.jp/lab-english/home-e.html`

Quelques thèmes de recherche

Syntaxe et sémantique des langages naturels, discours.

6.7 NEC

Coordonnées du laboratoire

Pattern Analysis Technology Group
C&C Media Research Laboratories

Site Internet

`http://www.nec.co.jp`

Quelques thèmes de recherche

Traduction automatique, recherche d'information.

6.8 Niigata University

Coordonnées du laboratoire

Miyazaki Laboratory (NLP Laboratory)
Department of Information Engineering
Faculty of Engineering
Niigata University

Site Internet

`http://www.nlp.info.eng.niigata-u.ac.jp/nlp/index.html`

Quelques thèmes de recherche

Analyseur morpho-syntaxique.

6.9 NTT Basic Research Laboratories (BRL)

Coordonnées du laboratoire

Information Science Research Laboratory
NTT Basic Research Laboratories

Site Internet

`http://www.brl.ntt.co.jp/info/index.html`

Quelques thèmes de recherche

Dialogue, interfaces homme-machine

6.10 NTT Communication Science Laboratories

Coordonnées du laboratoire
NTT Communication Science Laboratories

Site Internet
`http://www.kecl.ntt.co.jp`

Quelques thèmes de recherche
Traduction automatique, recherche d'information, représentation sémantique, dictionnaire.

6.11 NTT Human Interface Laboratories

Coordonnées du laboratoire
NTT Human Interface Laboratories

Site Internet
`http://www.hil.ntt.co.jp/top/index_e.html`

Quelques thèmes de recherche
Dialogue, extraction d'information.

6.12 NTT Software Laboratories

Coordonnées du laboratoire
Global Computing Laboratory
NTT Software Laboratories

Site Internet
`http://www.kokono.net`

Quelques thèmes de recherche
Recherche d'information mobile..

6.13 Shinshu University

Coordonnées du laboratoire
Nakano and Murayama Laboratory
Department of Information Engineering
Faculty of Engineering
Shinshu University

Site Internet
`http://sunak2.cs.shinshu-u.ac.jp/index.html`

Quelques thèmes de recherche
Analyse et reconnaissance de documents écrits.

6.14 Tohoku University

Coordonnées du laboratoire

Aso Laboratory
Department of Electrical and Communication Engineering
Graduate School of Engineering
Tohoku University

Site Internet

`http://www.aso.ecei.tohoku.ac.jp/index-e.html`

Quelques thèmes de recherche

Analyse de documents écrits, reconnaissance de caractères.

6.15 Tokushima University

Coordonnées du laboratoire

Aoe Laboratory
Department of Information Science and Intelligent Systems
Faculty of Engineering
Tokushima University

Site Internet

`http://www-b3.is.tokushima-u.ac.jp/aoe/index.html`

Quelques thèmes de recherche

Classification automatique, recherche d'information.

6.16 Tokyo Institute of Technology

Coordonnées du laboratoire

Tanaka and Tokunaga Laboratory
Department of Computer Science
Graduate School of Science and Engineering
Tokyo Institute of Technology

Site Internet

`http://tanaka-www.cs.titech.ac.jp/tanaka-home-e.html`

Quelques thèmes de recherche

Analyseur morphologique et syntaxique, recherche d'information.

6.17 Tokyo University

Coordonnées du laboratoire

Tsujii Laboratory
Department of Information Science

Faculty of Science
Tokyo University

Site Internet

`http://www.is.s.u-tokyo.ac.jp/~tsujiilab/`

Quelques thèmes de recherche

Analyseur morphologique et syntaxique.