Mémoire

Florian Constant et Aubry Gudel

22 février 2014

Table des matières

Chapitre 1

Introduction à la problématique

1.1 Le scénario

Avant toute chose, nous allons commencer par définir le scénario, sans quoi, il serait difficile de se poser des limites dans nos réflexions.

Notre scénario Dans un cadre professionnel, dix personnes de nationalités différentes et donc ne parlant pas la même langue se joignent autour dune table afin davoir une réunion. Chaque personne parlera dans sa langue maternelle. Afin que tout le monde puissent comprendre ce que les autres disent et que la réunion soit dynamique, un système sera mis en place afin de traduire chaque personne et de retranscrire vocalement la traduction. Ce système sera relativement poussé puisquil traduira en temps réel, cest-à-dire quil fera la traduction alors même que linterlocuteur na pas fini sa phrase. De plus, ce système sera capable danalyser les émotions dans la voix et de la restituer dans la synthèse vocale.

Nous allons devoir identifier chaque composante de ce système dans notre scénario et affiner celui-ci.

1.2 Les composantes

Le contexte Premièrement, nous allons devoir déterminer le contexte de la future discutions pour notre système. En effet, que le système sache dans quelle contexte se situe la discussion est très important car il permettra daméliorer grandement la fiabilité de la traduction. Lors de la traduction le niveau de langage approprié pourra être choisi, ce qui rendra la traduction plus fidèle.

On pourrait imaginer que le système analyse la discussion afin de déterminer ce contexte mais le système serait moins efficace le temps de cette analyse. Cest pourquoi nous allons partir du principe que quelquun initialise le contexte avant la réunion (via une interface par exemple). Le contexte choisi par cette personne sera donc « Professionnel ».

Identification et langues Une fois le contexte choisi, la réunion peut commencer. Les personnes commenceront alors à parler. Ici nous trouvons alors un autre problème : Comment détecter qui parle ? Et dans quelle langue ?

Afin de répondre à la première question, nous avons pensé à un système qui analyse les fréquences des voix ainsi que leurs signatures afin didentifier chaque personne si le système navait quun micro central. Cela fonctionnerait certainement mais dans ce cas, comment restituer dans la langue de la personne la traduction sans avoir une cacophonique de traduction dans la salle ? Il nous paraît alors évident quil soit nécessaire que chaque participant ai une oreillette dans laquelle on lui traduit les discussions.

Afin déviter un processus didentification des propriétaires par rapport à leur voix afin de ne pas leur traduire leur propre parole, il nous a paru plus simple de supposer que les oreillettes contiennent un micro dans lequel parle linterlocuteur. De cette manière, chaque personne sera identifiée de manière simple et sûre.

Pour ce qui est de la seconde question, cest-à-dire de comment connaître la langue, il sagira de la détecter en fonction de ce que la personne dira. Un peu à la manière dun Google Translate avec lécrit. Mais nous reviendront sur la réalisation de cette tâche plus tard.

Reconnaissance automatique de la parole A ce stade, il nous faut encore réussir à mettre sous format textuel ce quune personne dit afin de le traduire par la suite. Il existe des logiciels spécialisé dans ce travail, on les appelle des logiciels de Speech To Text, ou en français : « Reconnaissance automatique de la parole ». Il existe une multitude de logiciels comme ceci, cest pourquoi nous nallons pas plus en parler pour le moment et estimé que ce travail est fait.

Traduction Maintenant que lon dispose du texte, nous pouvons enfin le traduire dans toutes les autres langues présentes dans la salle. Pour ce faire nous allons devoir utiliser un traducteur, seulement, pour parfaire la traduction, le traducteur devra prendre en compte le contexte de la discussion afin de choisir les traductions les plus adaptées.

Synthèse vocale Il est désormais temps de restituer la traduction à chaque interlocuteur. Pour réaliser cela, nous allons devoir faire appel à un système de synthèse vocale. Une synthèse vocale a pour but dexprimer un texte par la parole, de la manière la plus naturelle et humaine possible dans notre cas. De nos jours, il existe une multitude de logiciels de synthèse vocale, nous en parlerons plus en détails dans une autre partie.

Émotions Il temps de parler d'une composante qui est transverse à plusieurs autres, citées plus haut. Il s'agit de l'émotion, nous souhaitons que la traduction permette à celui qui écoute de ressentir l'émotion transmise par la personne qui parle. Pour se faire, il vas falloir analyser la voix (Speech To Text), en analyser les émotions puis la retranscrire dans le choix de certains mots peut-être, durant la traduction. Puis il faudra la retranscrire via la synthèse vocale de la manière la plus fidèle possible.

La composantes de l'émotion est très intéressante car encore absente des solutions temps réel actuel, sa détection et synthétisation en temps réel est un problème toujours non résolu alors que son apport en terme de dynamisme

comme d'aide à la compréhension -qui sont deux facteurs très important pour un système temps réel. C'est pourquoi nous nous pencherons plus particulièrement sur ce facteur et détaillerons comment la détecter, comment la comprendre et comment la retransmettre dans les limites imposés de notre scénario.

Chapitre 2

Résolution du système

2.1 Reconnaissance automatique de la parole

Introduction La reconnaissance automatique de la parole (que nous abrégerons ci-après en RAP) est un système qui permet d'analyser la parole d'une personne pour la retranscrire sous la forme d'un texte.

Cette technique est fréquemment utilisé de nos jours. Que ce soit dans les serveurs vocaux (messageries, assistance...), sur un ordinateur avec la dictée vocale ou dans les assistants personnels (Google Now, Siri).

2.1.1 Fonctionnement générale

Petit historique Les premiers systèmes de RAP numériques datent d'il y'a plus de cinquante ans, ces premières solutions était limité à la reconnaissance de mots isolé dans un vocabulaire très limité (cardinal d'une dizaine de mots). Nous nous intéresserons ici uniquement aux techniques modernes permettant de couvrir une langue complète et de reconnaître de la parole continue, soit des phrases et discours complets ayant un sens.

Parole continue Les système de reconnaissance de parole continue reçoivent en entrée un signal (analogique ou numérique) correspondant à un message orale à retranscrire en texte, ce qui pose un certain nombre de difficultés successive. La continuité de texte impose au système de :

- redécouper le signal en mots, tâche compliqué par le fait qu'à l'oral il n'y a pas de réel séparation (pause) entre les mots, si ce n'est pour ponctuer, et que dans certain cas des liaisons sont faite entre plusieurs mots une syllabe pouvant ainsi se retrouver *à cheval* entre deux mots
- rassembler les mots en phrase en suivant des problématiques de syntaxe
- de vérifier que le résultat obtenue est sémantiquement correcte, chaque phrases doit avoir un sens

Le modèle le plus courant actuellement consiste à séparer les deux types principaux de difficultés, d'une part le signal acoustique est décodé vers une information phonétique puis les informations phonétiques sont traité par des algorithmes de modélisation du langage.

Tout d'abord le décodage phonétique s'opère en analysant des fenêtrant le signal d'origine et comparant ses composés à la bibliothèque d'unités connue utilisée par le système (mot, syllabe, diphone, phonème, etc.). Dans un second temps les informations décodé doivent être traité pour obtenir des phrases syntaxiquement et sémantiquement correcte, pour ce faire les systèmes actuels reposent sur des procédés statistiques étudiant la probabilité d'une suite de mots et ajustant ainsi les séquences trouvé.

Étude du signal (décodage acoustico-phonétique)

L'étude du signal s'opère le plus le plus souvent via une analyse paramétrique de ce dernier et permets ainsi de retrouver l'élément acoustique le plus probable correspondant au signale. Lors de l'acquisition des données acoustiques des pertes liées au matériel (plage de fréquence du microphone) sont inévitable ainsi que des perturbations liée à lenvironnement (réverbération de la salle, bruit ambiant). Nous nétudierons pas ici les différentes solutions qui existent permettant damoindrir l'impacte des deux facteurs précédemment énoncé et passerons directement à l'étude du signale nettoyé. L'étude se fait alors par fenêtrage, le signal de la parole n'évoluant que peu sur des durées de quelques millisecondes (stationnarité locale). Sur chaque fenêtre une analyse spectrale est alors effectuée, par exemple grâce à la décomposition du signal sur la fenêtre considéré via une transformée de fournier puis en analysant le résultat avec un ensemble de filtres passe-bande permettant le représentation de léchantillon par un sonagramme. Il Suffirait alors d'effectuer une reconnaissance de forme en comparant le résultat à notre bibliothèque et en conservant l'élément le plus probable. Malheureusement il n'est pas possible d'effectuer directement une tel opération, la durée de chaque unité à reconnaître n'étant pas stable il faut pouvoir faire coïncider le signal relevé avec les données de notre bibliothèque via des procédés de normalisation temporel. Pour ce faire il existe plusieurs modèles notamment les modèles Markoviens Cachées, les modèles neuronimétriques ou encore les algorithme de comparaison dynamique.

La technique la plus couramment utilisé aujourd'hui se fait par modèles de markov caché. Il est alors possible de retrouver l'ensemble des états probablement parcouru à partir des observations faites en isolent le chemin le plus probable.

Reconstruction sémantique

Une fois les unités phonétiques de bases isolées il reste alors à identifier les formes lexicales correspondantes. Cette étape est appelé le décodage. Les techniques de décodages reposent sur des graphes on retrouve alors des résolutions se basant sur l'algorithme A*. On génère alors un graphe où les noeuds comportent tous les mots probables composés par les unités phonétique préalablement isolés.

2.1.2 Et pour nous ?

Dans le cadre de notre scénario, plusieurs des difficultés peuvent être purement ignorées. Ainsi l'étape de soustraction du bruit ambiant n'est pas nécessaire, les réunions se déroulant typiquement dans des lieux éloignés des perturbations sonores.

Pour ce qui est du problème de parole concurrente (deux interlocuteurs se mettent à parler en même temps), celui ci pose deux problèmes : séparer ce que disent les deux intervenants pour permettre d'identifier ce que chacun dit, puis retranscrire le résultat à l'auditoire. Si il est certainement possible de séparer ce que dit chaque intervenant cela demanderait des traitement supplémentaire qui ne sont d'après nous pas justifiable : comment retranscrire ensuite à l'auditoire ce qui est dit ? Le résultat de chaque traduction ne peut pas être rendu en même temps, aucun ne serait alors intelligible et si l'on décale la synthétisation de l'une des traduction quand la passer ? Elle se retrouverait déplacé à plus tard indéfiniment à mesure que d'autres intervenants prendraient la parole. Nous avons donc trouvé préférable d'ignorer ce cas en nous reposant sur la discipline des participants.

Quand aux techniques de reconnaissances se posent trois questions : l'unité à utiliser, le modèle de décodage acoustico-phonétique et l'algorithme de reconnaissance lexical/grammatical ; ce dernier étant directement liée au choix du modèle de décodage acoustico-phonétique. Dans le cas de la parole continue il a été identifié que l'unité la plus adapté était le phonème : en effet l'ordre de grandeur des phonèmes nécessaire à la modélisation d'un langage est très faible, même lors de l'utilisation de phonème contextualisé (prenant en compte de possibles liaisons) et permet ainsi une relative légereté du dictionnaire comportant les mots reconnaissable. Nous utiliserons donc le phonème. Les autres paramètres nayant pas dimpact particulière sur le cur de notre recherche nous sélectionnerons le système le plus utilisé, les Modèles de Markov cachés adjoint à , ceux là ayant déjà prouvé leurs efficacités.

2.2 Traduction

2.2.1 Présentation et but

La traduction est le principe de faire passer un texte d'une langue à une autre. Elle sert à représenter un texte dans une autre langue, elle en garde donc le même sens et les deux textes comportes donc beaucoup de similitudes.

Historiquement, la traduction à d'abord été un travail d'humain. Puis par la suite, c'est l'informatique qui s'en est chargé. Lorsqu'un système informatique traduit un texte, nous appelons cela de la "traduction automatique". Dans notre scénario, c'est cette traduction automatique dont nous allons parler.

La traduction ne prend pas seulement en compte le fait de traduire une suite de mot d'une langue à une autre. Il faut préserver le sens et être le plus fidèle possible. Pour cela, le traducteur (ou système de traduction) doit connaître le contexte du texte, la grammaire des deux langues, ainsi que leur cultures. En effet, la culture compte pour beaucoup car certains mots dans une langue n'ont pas la même signification dans une autre, il faut alors transformer les mots afin de refléter l'idée du texte original. C'est pourquoi les traduction les plus fiables et fidèles sont faites par des humains spécialisés dans la traduction.

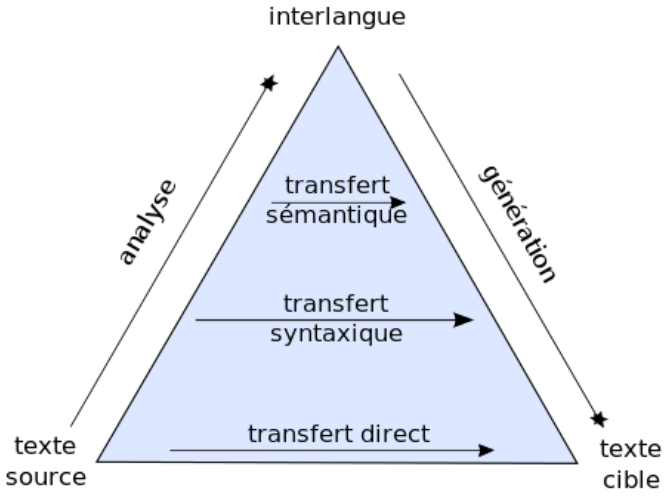

FIGURE 2.1 – Triangle de Vauquois

2.2.2 Solutions

Tout d'abord, il faut savoir que dans la traduction non informatisée il y a trois phases durant la traduction :

- La compréhension : comprendre le sens du texte d'origine
- La dé-verbalisation : garder le sens du texte sans les mots
- La ré-expression : formulation du sens du texte dans la langue d'arrivée

Dans la traduction informatique, la compréhension sera appelée **analyse**, la dé-verbalisation **transfert** et la ré-expression **génération**. Ces trois phases sont représentées dans le triangle de Vauquois ci-dessous.

Ce schéma est représentatif des différents chemin possibles afin de traduire un texte source dans une langue cible. Ces différents chemins visibles dans le schéma sont les différentes manières de faire actuellement utilisées. Nous pouvons interpréter cela comme "'Plus l'analyse est longue, plus le transfert est court"'.

Il existe quatre possibilités :

- **Le transfert direct** : il n'y a pas presque pas d'analyse, toute la traduction se joue au niveau du transfert. Les méthodes de traductions utilisant ce type de transfert sont la traduction par l'exemple et la traduction statistique. Dans cette possibilité, la traduction n'est qu'un processus de décodage.
- **Le transfert syntaxique** : ici, le transfert est syntaxique, c'est à dire que l'on va s'appuyer sur les arbres syntaxique des langues afin de construire des phrases de même sens. La méthode de traductions utilisant ce type de transfert est la traduction automatique à base de règles.

– **Le transfert sémantique** : le transfert est sémantique, c'est à dire que l'on traduit à partir du sens du texte d'origine. Il s'agit de la méthode de traduction que nous utilisons (les humains). Il n'y a que très peu de traducteur automatique se basant sur le transfert sémantique, de part sa complexité à modéliser les sémantiques ainsi que part sa difficulté à la mettre en place.
– **L'interlangue** : cette possibilité supprime le transfert. La traduction devient alors universelle. Et il ne reste plus que les phases d'analyse et de génération. On désigne également l'interlangue sous le nom de "langue pivot". Cette méthode n'a pas eu de succès chez ceux qui l'ont essayé.

De nos jour, les moteurs de traduction utilisent pour la plupart la traduction par règles ou la traduction statistiques. Mais il existe une approche utilisant ces deux systèmes de traductions qu'utilise les leaders du marché que sont Systran, Google Translate et Bing Translator.

Traduction statistique

La traduction automatique statistique est basée sur l'utilisation de « modèles statistiques » auto construits à partir de corpus monolingues et bilingues. La construction des « modèles statistiques » est rapide mais requiert d'avoir à disposition des volumes importants de textes traduits. Généralement, un modèle bilingue nécessite au minimum 2 millions de mots pour la traduction dans un domaine spécifique, mais il en nécessite beaucoup plus pour le domaine général. La traduction automatique statistique requiert donc des configurations matérielles lourdes afin d'utiliser les modèles de traduction tout en fournissant des performances normales.

Les modèles statistiques sont basés sur des corpus parallèles, c'est à dire que le même corpus est traduit dans une (ou plusieurs) autre(s) langue(s). Les alignements entre les textes peuvent être réalisés à différents niveaux : paragraphes, phrases, expressions et mots. Cependant, l'alignement mot-à-mot reste celui qui nous offre le plus d'information. Pour observer comment cet alignement est réalisé, veuillez vous référer au schéma 4.1 en annexe.

Un dictionnaire bilingue est alors construit à partir de ces bi-phrases et les probalité de correspondances sont calculées. Le tableau ci-dessous représente ce dictionnaire dans exemple du français vers l'anglais.

Chaque ligne du tableau représente une traduction et chaque élément de cette traduction est séparée par le symbole « ||| ». Dans la première colonne apparaît le mot ou la séquence de mots d'une langue source, dans la seconde colonne apparaît le mot ou la séquence de mots en langue cible. Puis, la dernière colonne représente la probabilité de correspondance de ce mot lors d'une traduction. Dans la réalité, la dernière colonne contient une série de scores du modèle statistique. Mais nous avons choisi de n'afficher que le score du modèle de traduction afin de simplifier la compréhension. Le score du modèle de traduction est utilisé pour évaluer les différentes traductions possibles pour un même mot ou séquence de mot en langue

la	|||	the	|||	0.6
la	|||	this	|||	0.4
femme	|||	woman	|||	0.5
femme	|||	wife	|||	0.5
la femme	|||	the woman	|||	0.6
a femme	|||	this woman	|||	0.4
de	|||	of	|||	0.6
ménage	|||	household	|||	0.8
femme de ménage	|||	maid	|||	1.0
a sauté	|||	jumped	|||	0.85
a sauté	|||	skipped	|||	0.15
un repas	|||	a meal	|||	1.0
sauté un repas	|||	skipped a meal	|||	1.0

FIGURE 2.2 – Dictionnaire bilingue avec probabilités

source. Plus le score est proche de 1, plus la traduction a de chance d'être la bonne. Par exemple, la traduction de « la » en « the » a une probabilité de 0.6 (60%), c'est à dire que dans la majorité des cas, il faudra utiliser cette traduction et dans 40% des cas, on devra utiliser la traduction en « this ».

Lors d'une traduction, le système de traduction automatique statistique va commencer par segmenter le texte en mots, séquences de mots ou signes de ponctuations. Puis, il assemble un ensemble d'hypothèses de traduction utilisant les traductions que nous avons montrés dans la figure 4.1. Comme il existe de nombreuses façons de découper le texte et de traduire les séquences, la création des hypothèses produit une longue liste d'hypothèses de traduction. A chaque hypothèse on associe un score calculé à partir du modèle statistique. L'hypothèse ayant le plus gros score est alors choisi pour être la traduction que l'on va retourner.

Traduction par règles

TODO

Traduction par l'exemple

La traduction par l'exemple, également appelé traduction par analogie, réside dans le principe de prendre des petites phrases courtes qui sont traduites dans plusieurs langues et d'identifier ces phrases dans le texte à traduire. Cela permet une meilleure préservation du sens des phrases. Le résultat est donc plutôt satisfaisant.

La traduction par lexemple s'introduit entre la traduction par règles et la traduction statistique. En effet, beaucoup dapproches de la traduction intègrent des règles et des techniques statistiques. Cependant, certaines

caractéristiques dissocient la traduction par l'exemple de la traduction par règle et de la traduction statistique. Dans la traduction par l'exemple, on isole le texte par phrase.
Voici les principales étapes de la traduction par l'exemple :
- **Décomposition** de la phrase en séquences correspondant le mieux aux exemples en base de données
- **Traduction** des séquences dans la langue cible par analogie avec les exemples en base de données
- **Recomposition** des séquences afin de former une phrase dans la langue d'origine

La traduction par l'exemple est particulièrement performante avec les verbes à particules - verbes qui changent de sens selon la particule (adverbe, adposition, nom ou adjectif) comme "'mettre bas"' ou "'passer outre"' en français - qui sont très présent dans les langues germaniques. Ce type de traduction est particulièrement performant avec ces verbes car il sappuie sur des séquences de phrases, le contexte est donc conservé et la génération d'un verbe à particule est donc beaucoup plus aisé. Cependant, cette traduction nécessite une base de données très fournie pour fonctionner correctement. Cette base de données devra être alimenté par des textes de références traduit dans chaque langue. Il s'agit d'un lourd investissement.

2.2.3 Notre choix

TODO

2.3 Synthèse vocale

2.3.1 Présentation et but

TODO

2.3.2 Solutions

TODO

2.4 Émotions

2.4.1 Présentation et but

TODO

2.4.2 Travaux sur le sujet

TODO

2.4.3 Solutions

TODO

Chapitre 3

Aller plus loin

TODO

Chapitre 4

Annexes

FIGURE 4.1 – Corpus Parallèle