

Conception et Réalisation d'un Système de Traduction Automatique à base d'Exemples

Bessou Sadik, Benzdira Sana
Université Ferhat Abbas- Sétif

Résumé

Dans ce travail, nous présentons la conception d'un system de traduction automatique à base d'exemple de l'anglais à l'arabe. Nous proposons une méthode qui traduit la phrase à partir des mots constituants qui se trouvent dans d'autres phrases dans la base de données, l'avantage de la méthode est la désambiguïsation des mots car ces derniers se trouvent dans des phrases c'est-à-dire dans un contexte.

1. Introduction

Le but de ce travail est de réaliser un système de traduction automatique Anglais Arabe par l'utilisation des phrases alignées. Autrement dit à l'aide d'un ensemble prédéfini de phrases, on traduira un nombre supérieure en combinant les mots constituants ces phrases pour obtenir de nouvelles phrases.

2. Les outils utilisés

Dans ce travail on a utilisé deux outils :

Part of speech Pos-tagger et Microsoft Office Access_

2.1Part of speech Pos-tagger

Part of speech tagging

Dans la linguistique des corpus, part-of-speech tagging (pos tagging ou post) appel aussi grammatical tagging ou désambiguïsation de la catégorie des mots, est le processus de marquage des mots dans un texte (corpus), basée sur sa définition, ainsi que son contexte.

Autrement dit c'est une forme simplifiée de ce qui est généralement enseignée aux enfants d'âge scolaire, à l'identification des mots comme les noms , verbes , adjectifs et adverbes .

Pour quoi utiliser part of speech pos tagger ?

Dans les langues naturelles, des bons pourcentages de mots sont ambigus, en considérant qu'il y a 9 catégories de mots (part of speech) en anglais (dans notre cas on parle de l'anglais parce que c'est la langue dont on démarre):

nom , verbe , article ,adjectif , préposition , pronom ,adverbe ,conjonction et interjection . Cependant, il y a évidemment bien plus de catégories et sous-catégories. Pour

- le nom: il y a le pluriel et le singulier.
- le verbe: il y a le temps du verbe.
- l'article: il y a le genre masculin et le genre féminin.

Dans part of speech tagging par ordinateur, il est courant de distinguer les 50 à 150 parties distinctes de la parole pour l'anglais, par exemple, NN pour le singulier des noms communs, NNS pour le pluriel des noms communs, NNP pour le singulier des noms propres, NNPS pour le pluriel des noms propres.

Exemple

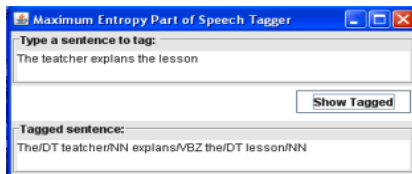


Figure.1 : Quelques mots en anglais et leurs Pos-tags.

Dans le tableau ci-dessous quelques catégories utilisés dans le Pos-Tagger

Tableau 1 : Quelques catégories utilisées par Pos-Tagger.

Catégorie	Signification	Exemple
DT	déterminer	The, a
NN	Singular noun	teacher
NNS	Plural Noun	teachers
RBR	comparatif adverb	more
CC	coordinating conjunction	And, or
IN	Preposition/subordinating conjunction	In, by, for
CD	cardinal numeral	one, two, 31
VBZ	verb, 3rd. singular present	Eats,plays,makes
JJ	adjective	Franch,big
VBD	verb, past tense	Condemned
MD	modal auxiliary	can, should, will
PP\$	possessive personal pronoun	my, our

2.2 Microsoft Office Access

Pour traduire une phrase il faut sauvegarder les mots constituant cette phrase dans une base de données.

Conception et Réalisation d'un Système de Traduction Automatique à base d'Exemples

Ainsi il faut déterminer la signification des mots en langue source (anglais) pour pouvoir traduire cette phrase en langue cible (arabe) en respectant les différents niveaux :

● Morphologique (qui est la mise en évidence des signes minimaux et l'analyse de la façon dont ces signes minimaux peuvent constituer des mots).

● Syntaxique (l'ordre des mots dans la phrase et les relations qui existent entre les éléments qui la composent).

● Sémantique (c'est le sens des mots constituant la phrase et le sens de la phrase elle-même).

Les tables de la base sont :

➤ *La table des phrases:*

Numero	PhraseSource	PhraseCible
1	Ebay condemned to compensate an injured purchaser	تدان ايباي بتعويض مشتري متضرر
2	The French soldiers kill hundreds in Afghanistan	يقتل الجنود الفرنسيون مئات في أفغانستان
3	The China denies the impact of its dams on drought of the Mekong	تنكر الصين تأثير سدودها على جفاف المكونغ
4	The Sudan defer the elections	تؤجل السودان الانتخابات
5	Four wounded by balls in NewYork	اصابة اربعة برصاصين بانيويورك
6	The Polish president killed in plane crash	قتل الرئيس البولندي في تحطم طائرة
7	Israel and Hamas reject the peace	ترفض إسرائيل و حماس السلام
8	The Indian navy sinks Somali pirate ship	البحرية الهندية سفينة قرصان سومالية
9	Kosovo declares the independence	تعلن كوسوفو الاستقلال
10	France prepare for war with Iran	تخطر فرنسا الحرب على إيران
11	Osama Bin Laden releases a new video	ينشر أسامة بن لادن فيديو جديد
12	new threats to kill Al Jazira journalist	تهديدات جديدة بقتل صحفي الجزيرة
13	North Korea tests seven missiles	تختبر شمال كوريا سبعة صواريخ
14	Myanmar moves its capital	تنقل ميانمار عاصمتها
15	Earthquake kills 3000 in South Asia	يقتل زلزال 3000 في جنوب اسيا
16	The hurricane Katrina kills hundreds	يقتل الإعصار كاترينا مئات
17	Iran resumes the nuclear work	تستأنف إيران العمل النووي
18	Al Qaeda warns of more attacks	تحذر القاعدة بهجمات أكثر
19	The China closes 1250 sites for healthier Web	تغلق الصين 1250 موقع لويب اصح
20	The Japanese car stills most reliable	تبقى السيارة اليابانية موثوقة أكثر
21	He reads a book	يقرأ كتاب
22	The boy plays with a car	يلعب الولد بسيارة
23	The teacher explains the lesson	يشرح المعلم الدرس
24	The colonization torture the combaters	يعذب الاستعمار المقاومين

Figure.2 : La table des phrases.

➤ *La table de grammaire des phrases en anglais :*

N°	Categ1	Categ2	Categ3	Categ4	Categ5	Categ6	Categ7	Cate
1	NNP	VBD	TO	VB	DT	JJ	NN	
2	DT	JJ	NN5	VBP	NNS	IN	NNP	
3	DT	NNP	VBZ	DT	NN	IN	PRP\$	NN5
4	DT	NNP	VB	DT	NNS			
5	CD	VBN	IN	NNS	IN	NNP	NNP	
6	DT	JJ	NN	VBN	IN	NN	NN	
7	NNP	CC	NNP	VBP	DT	NN		
8	DT	JJ	NN	VBZ	JJ	NN	NN	
9	NNP	VBZ	DT	NN				
10	NNP	VB	IN	NN	IN	NNP		
11	NNP	NNP	NNP	VBZ	DT	JJ	NN	
12	JJ	NNS	TO	VB	NNP	NNP	NN	
13	NNP	NNP	VBZ	CD	NNS			
14	NNP	VBZ	PRP\$	NN				
15	NNP	VBZ	CD	IN	NNP	NNP		
16	DT	NN	NNP	VBZ	NNS			
17	NNP	VBZ	DT	JJ	NN			
18	NNP	NNP	VBZ	IN	JJR	NNS		
19	DT	NNP	VBZ	CD	NNS	IN	RBR	NN
20	DT	JJ	NN	VBZ	RBS	JJ		
21	PRP	VBZ	DT	NN				
22	DT	NN	VBZ	IN	DT	NN		
23	DT	NN	VBZ	DT	NN			
24	DT	NN	VB	DT	NNS			

Figure.3 : La table de grammaire d'anglais.

➤ *La table de grammaire des phrases en arabe*

N°	MotSource	MotCible	Categorie	Genre	SourceVB	Racine
175	calls	ندعو	VBZ		دعوى	دعوة
135	campaign	حملة	NN	F		
14	capital	عاصمة/مها	NN	F		
104	car	سيارة	NN	F		
30	China	صين	NNP	F		
118	closes	تغلق	VBZ		علق	علق
99	colleague	زميل	NN	M		
122	colonization	استعمار	NN	M		
124	combaters	مقاتلين	NNS	M		
176	commerce	تجارة	NN	F		
53	compensate	تعوّض	VB		عوض	تعوّض
52	condemned	كأن	VBD		دان	ادانة
153	countries	بلدان	NNS	F		
209	country	بلد	NN	M		
159	country	بلد	NN			
184	crash	حادثة	NN	M		
65	dams	سدود	NNS	F		
163	deal	صفقة	NN	F		
156	deal	صفقة	NN			
79	declares	أعلن	VBZ		أعلن	إعلان
112	defer	تؤجل	VB		أجل	تأجيل
62	denies	نكر	VBZ		نكر	إنكار
203	Dollars	دولار	NNS	M		
67	drought	جفاف	NN	M		

Figure.4 : La table de grammaire d'arabe.

➤ *Une table pour l'aide à la conjugaison des verbes:*

N°	Categ1	Categ2	Categ3	Categ4	Categ5	Categ6	Categ7	Cate8
1	NNP	VBD	TO	VB	DT	JJ	NN	
2	DT	JJ	NNS	VBP	NNS	IN	NNP	
3	DT	NNP	VBZ	DT	NN	IN	PRP\$	NNS
4	DT	NNP	VB	DT	NNS			
5	CD	VBN	IN	NNS	IN	NNP	NNP	
6	DT	JJ	NN	VBN	IN	NN	NN	
7	NNP	CC	NNP	VBP	DT	NN		
8	DT	JJ	NN	VBZ	JJ	NN	NN	
9	NNP	VBZ	DT	NN				
10	NNP	VB	IN	NN	IN	NNP		
11	NNP	NNP	NNP	VBZ	DT	JJ	NN	
12	JJ	NNS	TO	VB	NNP	NNP	NN	
13	NNP	NNP	VBZ	CD	NNS			
14	NNP	VBZ	PRP\$	NN				
15	NNP	VBZ	CD	IN	NNP	NNP		
16	DT	NN	NNP	VBZ	NNS			
17	NNP	VBZ	DT	JJ	NN			
18	NNP	NNP	VBZ	IN	JJR	NNS		
19	DT	NNP	VBZ	CD	NNS	IN	RBR	NN
20	DT	JJ	NN	VBZ	RBS	JJ		
21	PRP	VBZ	DT	NN				
22	DT	NN	VBZ	IN	DT	NN		
23	DT	NN	VBZ	DT	NN			
24	DT	NN	VB	DT	NNS			

Figure.5 : Une table pour l'aide à la conjugaison.

➤ *La table d'alignement :*

NUM	position1	position2	position3	position4	position5	position6	position7	position8
1		0	2	3	-1	5	4	
2	3	4	2	0	5	6	7	
3	1	2	0	-1	3	-1	-1	
4	1	2	0	3	4			
5	1	0	2	3	4	5	6	
6	3	4	2	0	5	7	6	
7	1	2	3	0	4	5		
8	3	4	2	0	7	6	5	
9	1	0	2	3				
10	1	0	2	3	4	5		
11	1	2	3	0	-1	5	4	
12	1	0	2	3	5	6	4	
13	1	2	0	3	4			
14	1	0	-1	2				
15	1	0	2	3	4	5		
16	1	2	3	0	4			
17	1	0	4	5	3			
18	1	2	0	3	4	5		
19	1	2	0	3	4	5	7	
20	3	4	2	0	6	5		
21	1	0	-1	-1	0	-1	1	
22	1	2	0	3	-1	4		
23	1	2	0	3	4			
24	1	2	0	3	4			

Figure.6 : Une table pour l'alignement.

Sachant que chaque enregistrement de la phrase en anglais correspond un enregistrement de phrase en arabe.

3. La conception du système

Le but de notre travail est de: «concevoir un système de traduction automatique anglais arabe dans un domaine restreint par les corpus parallèles» .en passant par les trois étapes de la traduction automatique :

3.1. La phase d'analyse :

On peut la résumer dans le schéma suivant:

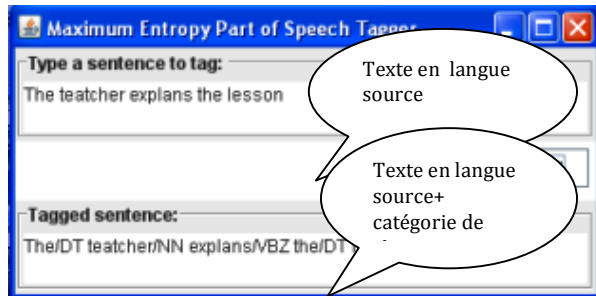


Figure.7 : Analyse avec POS Tagger

Donner en entrée une phrase en langue source (l'anglais), puis la faire passer au Part of speech pos-tagger, et on aura en sortie chaque mot et sa catégorie.

3.2. La phase de transfert

Dans cette phase on aura comme entrée le résultat du pos-tagger, c'est-à-dire les mots en langue source (anglais) et leurs catégories et en sortie des mots en langue cible (arabe).

Plus précisément en a deux cas:

- Si la phrase donnée en entrée existe dans le corpus (nos phrases), on affiche sa traduction.
- Si une partie de la phrase ou une phrase composée de deux parties existe alors on doit la faire subir à des traitements pour la traduire :

3.3. La phase de génération

Dans cette partie on aura en entrée la sortie de la phase précédente, et en suivant un certain nombre de règles de transformations en générera une phrase cohérente en langue cible (arabe).

4. *Les étapes de traduction*

Dans le cas où la phrase n'est pas prédéfinie (elle n'existe pas dans le corpus), on peut à partir des mots prédéfinis former de nouvelles phrases.

On prend la phrase en entrée et on sépare les deux parties.

- * On trouve les deux phrases entières en arabe puis on extrait l'équivalent de chaque partie.
- * On découpe chaque partie en arabe en mots, c'est une forme de lemmatisation.
- * A l'aide d'alignement on extrait les deux parties équivalentes en arabe.
- * On forme la phrase traduite en arabe.

Exemple :

1. "Israel warns Palestine to kill hundreds."
2. Trouver les deux parties:
 - "Israel warns Palestine to"
 - "kill hundreds."
 -
3. Trouver l'équivalent en arabe dans des phrases complètes:
 - "تحذر إسرائيل فلسطين بتوقع هجمات أكثر"
 - "يقتل الجنود الفرنسيون مئات في أفغانستان"
4. Découper chaque phrase en mots pour pouvoir correspondre chaque mot en anglais à son équivalent en arabe:

- تحذر إسرائيل فلسطين ب توقع هجومات أكثر
- يقتل ال جنود ال فرنسيون مئات في أفغانستان

5. Trouver l'équivalent des deux parties en arabe à l'aide des règles de grammaire et d'alignement:

- "تحذر إسرائيل فلسطين ب"
- "يقتل مئات"

6. Former des deux parties une phrase cohérente:

"تحذر إسرائيل فلسطين بقتل مئات"

Dans cette dernière phase il y a beaucoup de problèmes pour attacher les deux parties, parmi ces problèmes:

- Si la première partie se termine avec "ب" ou "ل" il faut concaténer ces derniers à la deuxième partie.
- S'il y a "ال" avant un mot il faut ajouter ce dernier au mot.

Par exemple: الجنود —→ الجنود mais si la phrase se termine par "ال" il faut l'attacher au premier mot de la deuxième partie.

- Le temps du verbe : il faut conjuguer le verbe selon le nouveau sujet puis réorganiser la phrase.
- L'adjectif : il faut que l'adjectif suit le nom qui vient après lui en genre et en nombre, donc il faut changer le genre et le nombre de l'adjectif selon le nom qui vient après puis réorganiser la phrase.

Exemple

The **French** soldiers: l'adjectif (French) suit le nom « soldats » en genre (masculin) et en nombre (pluriel), et c'est le cas en arabe: الجنود الفرنسيون

Par contre la phrase: « The **French** president » donne:

الرئيس الفرنسي

5. L'architecture du système

Le schéma suivant résume les trois étapes précédentes

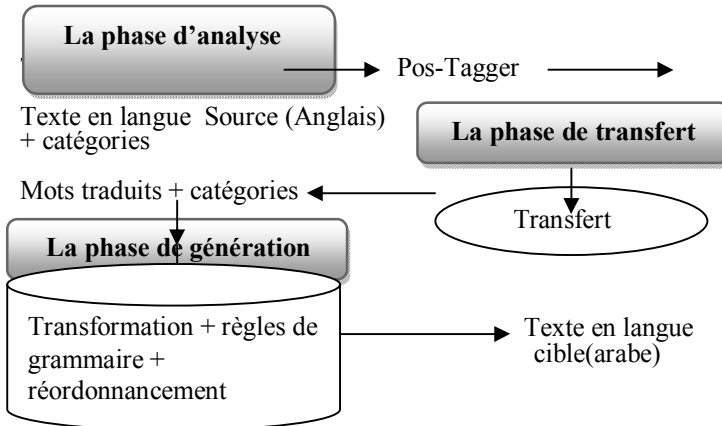


Figure.8 : Architecture du système de traduction automatique.

6. Conclusion

Nous avons présenté notre méthode pour concevoir et réaliser un système de traduction automatique à base d'exemple (Example based), en se basant sur la décomposition des phrases en deux parties seulement. La perspective du travail est la généralisation du travail à n parties pour traduire plus de phrases.

Références

- [1] Bessou Sadik.(2010) Morphological analysis and Generation for Machine Translation From and To Arabic, Colloque international sur la traductologie, université d'Oran Algérie.
- [2] Laurence Danlos. (2001) La compréhension en traitement automatique de la langue et la traduction automatique La Tribune Internationale des Langues vivantes.

- [3]: http://fr.wikipedia.org/wiki/Traduction_automatique . consulté Avril 2010.
- [4]: Tellier I. (2004) Introduction au TALN et à l'ingénierie linguistique université de Lille3
- [5]: Elaine Ménard, Approches et techniques de recherche d'information multilingue.
- [6]: Yves Roberge. Une brève introduction aux concepts de la syntaxe, l'Université de Toronto avec la collaboration de Henriette Gezundhajt
- [7]: Christian Jacquemin et Pierre Zweigenbaum, traitement automatique des langues par chapitre4
- [8]: Jérôme Goulian. (2000) Analyse linguistique détaillée pour la Compréhension Automatique de la Parole spontanée.
- [9]: Nanterre la Défense. (2009) « L'ambiguïté dans les sciences du langage », Université de Paris Ouest
- [10]: Philippe Muller, Traitement de Langages et Dialogues Naturels par Ordinateur
- [11]: Marie-Paule Muller. (2005) LANGAGES - GRAMMAIRES – AUTOMATES, 14 juillet 2005
- [12]: Christian Retoré. (2006) Logique, Linguistique et Informatique, Université Bordeaux 1.
- [13]: Yayoi NAKAMURA-DELLOYE. (2007) Introduction à la syntaxe.

