Traduction automatique
Contrôle des connaissances – Correction partielle

1 Cours (1pt)

Quelles sont les différences entre systèmes à mémoire de traduction et les systèmes « example-based » ?

- ⇨ Les systèmes à mémoire de traduction sont des outils interactifs de traduction assistée par ordinateur, utilisés surtout par des traducteurs.
- ⇨ Les systèmes EBMT sont des systèmes entièrement automatiques. En particulier la phase de **combinaison** des exemples de traduction qui matchent une partie de la phrase à traduire est faite automatiquement. Autre différence, les exemples sont traités de manière plus sophistiquée, avec possibilité de généraliser des exemples (les rendre plus généraux avec utilisation de classes sémantiques, et ajouts possibles de modifieurs).

Les similarités sont qu'ils réutilisent des traductions humaines existantes, avec une base d'exemples de traductions (phrastiques et sous-phrastiques) dans laquelle la recherche est optimisée.

2 Ambiguïtés (6 pts)

2.1 Qu'est-ce qu'une ambiguïté « syntaxique » (ou « structurelle ») ?

Il s'agit d'une ambiguïté (une même forme pour plusieurs sens), réelle ou artificielle, pour laquelle le niveau d'analyse requis pour lever l'ambiguïté est le niveau syntaxique. On a pour les différentes analyses :

- même représentation morphologique (même séquence de catégories)
- plusieurs représentations syntaxiques différentes (arbres syntagmatiques ou graphes de dépendances différents), correspondant à des sens différents.

Si certains sens sont inacceptables, il s'agit d'une ambiguïté artificielle.

Exemple

Paul commande un canard au fils de la voisine.

« de la voisine » peut s'attacher syntaxiquement à « fils », « canard » ou « commande ». Seul ce dernier rattachement est sémantiquement acceptable.

Ne pas confondre avec les ambiguités dites « morphologiques » : levées lorsque l'on fournit les catégories morpho-syntaxiques (dont les exemples vus en cours sont « osez le savoir », « le boucher sale la tranche »...)

2.2 Qu'est-ce qu'une ambiguïté artificielle ?

On a ambiguïté artificielle pour un énoncé qui n'est pas ambigu pour un humain, mais qui l'est pour un système de TAL : étant données les généralisations habituellement faites en linguistique et en TAL (comme l'utilisation de catégories morpho-syntaxiques, ou de règles de réécriture hors-contexte), on peut prévoir les cas où le système va trouver des analyses concurrentes, alors même que la plupart ne sont pas valides si on prend en compte le sens de l'énoncé.

Rem : cette définition dépasse le cadre de la TA : est valable pour tout système d'analyse automatique.

2.3 Soient les phrases :

(1) Le boucher branche la prise pendant la pause.

(2) Son chiffre d'affaires en fait une filiale convoitée par de nombreuses boîtes.

On considère les niveaux de représentation suivants :

- représentation après reconnaissance des mots
- représentation après analyse morphologique
- représentation après analyse syntaxique (phrase 1 en dépendances, et phrase 2 en constituants)

Pour chaque phrase, donnez pour chaque niveau, la ou les représentations, selon qu'il y a ou pas des ambiguïtés, **réelles ou artificielles**.
Dans le cas de mots composés, justifiez votre analyse.
Pour passer au niveau suivant, vous laisserez de côté les ambiguïtés artificielles.

(Indications : Pour l'analyse syntaxique, pensez aux analyses pouvant être obtenues avec d'autres mots mais la même séquence de catégories. Vous pourrez « factoriser » dans le même dessin plusieurs analyses en donnant avec des couleurs différentes des attachements concurrents).

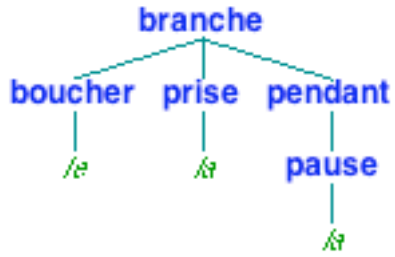
CORRECTION Phrase 1

Segmentation en mots :
Pas de pbs particuliers : *Le / boucher / branche / la / prise / pendant / la / pause*

Analyse morpho (en première colonne la bonne analyse)

Le	Det (défini, masc, sing)	Clitique (acc, masc, sing)
boucher	N(masc, sing)	V(inf)
branche	V(pres,mode=indicatif ou subj, p=1ou3, sing)	N(fem,sing)
la	Det(definit, fem, sing)	Clitique(acc, fem, sing)
prise	N(fem,sing)	V(mode=partpass, fem, sing)
pendant	Prep	V(mode=partpresent) N(masc,sg)
la	Det(definit, fem, sing)	Clitique(acc, fem, sing)
pause	N(fem,sing)	V(pres,mode=ind ou subj, p=1ou3, sing)

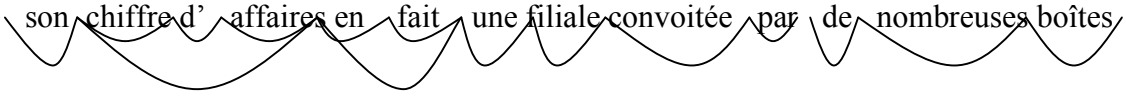
(même si le verbe « pauser » est rare)
NB : on ne donne pas la catégorie « pronom » aux clitiques, cf. ils n'ont pas du tout la même distribution que les pronoms...
Analyse syntaxique en dépendances :L'analyse syntaxique désambiguise totalement les catégories morpho-syntaxiques. Il y a une ambiguité artificielle de rattachement pour la préposition « pendant », sur le nom « prise » (=incorrect) ou le verbe « branche » (=correct). L'arbre de dépendances correct est :



CORRECTION Phrase 2

Segmentation en mots : On a 2 ambiguïtés artificielles de segmentation en mots :
- chiffre d'affaires : 1 seul mot composé (cf. autonome et insécable (*chiffre annuel d'affaires)) versus 3 mots. L'ambiguïté est ici artificielle, et d'ailleurs on ne voit pas de contexte où prévaudrait l'interprétation littérale. Donc « chiffre d'affaires » pourrait être systématiquement reconnu comme un composé, ne donnant plus lieu à ambiguité artificielle.
- en fait : ici on a deux mots (clitique + verbe faire), mais pour la machine, il y a ambiguité avec le mot composé « en fait », comme dans « en fait c'était faux », où on a insécabilité (*en premier fait, *en vrai fait ...).

Pour le clitique, on rappelle que si en linguistique il y a pas mal d'arguments pour considérer que les clitiques ne sont pas des mots (car pas autonomes) mais des affixes du verbe, en TAL, ce n'est pas pratique du tout de considérer un lexique avec les clitiques affixés au verbe. On donne ici une analyse avec les clitiques comme mots autonomes.
Donc la bonne segmentation est
Son / chiffre d'affaires / en / fait / une / filiale / convoitée / par / de / nombreuses / boites
On pouvait donner la représentation de la segmentation ambigue avec un DAG :



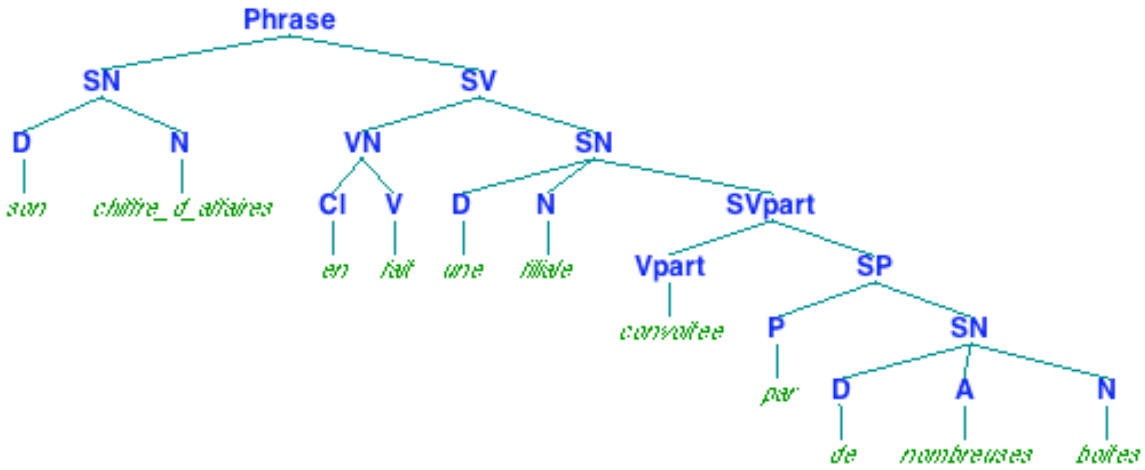
Analyse morpho (en première colonne la bonne analyse)

son	det(poss,masc,sing)	N(masc,sing)
chiffre_d_affaires	N(masc,sing)	
en	clitique(génitif)	prep
fait	V(ind,pres,p3,sing)	N(masc,sing) V(participe passé, masc, sing)
une	det(ind,fem,sing)	pro(fem,sing)
filiale	N(fem,sing)	Adj(fem,sing)
convoitée	V(partpass,fem,sing)	
par	prep	
de	det(ind,plu)	prep
nombreuses	adj(fem,plu)	
boites	N(fem,plu)	V(ind ou subj, p2, pres, sing)

NB : ici le « de » est un déterminant (il commute avec « ces », « les » ...) et pas la préposition !
Pour « convoitée » il n'y a pas d'ambiguité entre Adj et Vpartpass : tous les participes passés ont des emplois adnominaux. Parfois on a des lexèmes différents : un adjectif de sens légèrement distinct de celui du participe passé (exemple *Paul est fatigué* (pas d'agent interprétable) versus *Paul est fatigué par ses voyages bi-hebdomadaires*) mais ça n'est pas le cas pour *convoitée* (c'est toujours le V convoiter qui est sous-jacent (cf. ici « de nombreuses boîtes convoitent une filiale... »)).

Analyse syntaxique syntagmatique On a une ambiguïté artificielle de rattachement pour le SP *par de nombreuses boîtes*, sur *convoitée* ou sur *filiale* ou sur *fait*.
Rem : ce SP ne peut pas grammaticalement se rattacher dans le SN « son chiffre d'affaires » : on ne considère pas par définition de syntagmes discontinus !
Rem : dans le cours on a vu qu'en TAL il est rare de distinguer le rattachement d'un SP à la phrase versus au SV comme reflet de la distinction entre ajout et argument. Cette distinction est problématique dans le cas d'ajouts qui s'intercalent entre des arguments (comme dans *une lettre avait été envoyée la veille aux employés*)
La bonne analyse, en considérant un SN plat (sans niveau Nbarre) et un SV :

Rem : pour le « en », en analyse syntagmatique, on n'a que la solution de l'intégrer au noyau verbal. En analyses en dépendances, on pourrait produire un arbre de dépendances non projectif, avec « en » rattaché à « filiale ».



3 Transfert et Interlingua (4pts)

Soient les deux paires de phrases fr/en en relation de traduction :
La duchesse manquait beaucoup à Paul. / Paul missed the duchess a lot.
La duchesse rendra visite à sa voisine. / The duchess will pay a visit to her neighbor.

3.1 Quelles sont d'après vous les difficultés pour traduire ces phrases françaises en anglais ?
Pour la phrase 1 on a le placement de l'adverbe (rem : « a lot » est un mot composé de catégorie adverbe !), et le passage de l'appariement (correspondance entre rôles sémantiques / fonctions grammaticales) Xsuj, Y-à-obj en français vers l'appariement Y-suj, X-obj.
Pour la phrase 2, c'est la construction à verbe support rendre+visite qui doit être repérée, pour soit utiliser le verbe *to visit*, soit choisir le bon verbe support en anglais (to pay + visit), auquel cas il faut en outre insérer le déterminant en anglais. (NB :on ne parle pas d'expression figée dans ce cas, cf. visite/visit ont bien leur sens habituel.). En outre la traduction du possessif *sa* exige en anglais de connaître le genre du possesseur, ici féminin (la duchesse).

3.2 Proposez un lexique de transfert fonctionnel pour gérer la traduction Fr =>En de ces deux exemples

Comme son nom l'indique, le transfert fonctionnel utilise le concept linguistique de fonction grammaticale :

manquer (SUBJ(X), A-OBJ(Y)) <=> to miss (SUBJ(Y), OBJ(X)

Pour la construction à verbe support, on ne peut pas faire une entrée « rendre visite » cf. ce n'est pas un mot, on peut avoir toute la variation syntaxique (insertion d'adverbes, « visite » est objet direct, qui même s'il n'en a pas toutes les propriétés, en a tout de même certaines (relatives en *que*, clitique accusatif (*cette visite, il l'a rendue de bon cœur)*. Donc il fallait innover, par exemple avec une entrée des 2 verbes, avec l'objet fixé :

rendre (SUBJ(X), OBJ(='visite'), A-OBJ(Y) <=> pay (SUBJ(X), OBJ(='visit'), TO-OBJ(Y))
Rem : il faudrait être plus précis en indiquant qu'en anglais 'visit' doit avoir un déterminant, au contraire du français.

3.3 Proposez une représentation de type Interlingua pour les deux paires de phrases
Proposition de correction :
On demandait la représentation du contenu sémantique des phrases, et pas un lexique interlingua. On se place dans une hypothèse irréaliste où on serait capable d'inférer une telle structure à partir des phrases. Les informations de syntaxe et morphologie ne sont pas représentées : seul le sens est représenté. On pouvait utiliser une représentation par structures de traits, en s'inspirant d'une

représentation à la LFG, avec un système d'identifiants pour représenter la coréférence (pour gérer « sa voisine »). Les arguments sémantiques des prédicats sont indiqués comme valeurs d'attributs de type « rôles thématiques/sémantiques ».
Rem : par définition de la représentation Interlingua, il fallait une seule représentation par couple de phrases….

$$\begin{bmatrix} @MISSING & \begin{bmatrix} experiencer = [@PAUL[id = X]] \\ patient = [@DUCHESS[id = Y]] \\ time = past \\ aspect = imperfective \\ degree = @A\_LOT \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} @VISITING & \begin{bmatrix} time = future \\ agent = [@DUCHESS[id = X]] \\ patient = \begin{bmatrix} @NEIGHBOR & \begin{bmatrix} id = Y \\ neighbor\_of = X \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

4 Commentaires de résultats de traduction auto (5pts)

Correction : voir correction faite en TD

La traduction (1) est par règles, la (2) est statistique.
Deux arguments principaux amènent à cette conlusion :

(i) Globalement la traduction (1) contient des phrases grammaticales, même si parfois lourdes (comme par ex. « systèmes de contrôle de construction japonaise de navigation et de climat »). Au contraire la traduction (2) contient plusieurs agrammaticalités, avec en particulier le fait d'avoir des phrases sans V conjugué (participe seulement : « assis », « fermé »). On peut en effet supposer qu'un traducteur par règles va contrôler la structure globale de la phrase produite….

(ii) Localement, certains termes sont mieux traduits en (2), ce qui indique une traduction guidée par les données (« fabricant de produits de luxe » mieux que « fabricant de luxe de marchandises » ; « réseau électrique » mieux que « grille électrique »)

Exemples de commentaires :
traduction (1)
point négatif : « grille électrique rompue de cette nation » : problème de non reconnaissance d'un terme (electric grid <=> réseau électrique) et donc problème de transfert lexical. « rompue » n'est pas non plus très bon. => « réseau électrique coupé ? de cette nation »
point négatif : le fabricant de luxe de marchandises : visiblement à partir du SN anglais « luxury goods maker », « luxury » a été rattaché à « maker » au lieu de « goods ».
point négatif : « left » traduit par « à gauche » : l'ambiguité de « left » (V to leave versus adjectif) a mal été résolue
point négatif : « wary » traduit par « attentif », mauvais transfert lexical, plus une faute d'accord. => l'analyseur n'a visiblement pas rattaché « wary » au sujet, cf. en effet on a une construction avec syntagme adjectival extraposé.

mieux que (2) : « sit » => « se reposent » : bonne génération des marques flexionnelles de mode, nombre : « sit » a bien été analysé comme V conjugué, dont le sujet est pluriel. Le transfert lexical « to sit / se reposer » n'est pas trop mal, et bien déclenché dans le cas d'un sujet inanimé.

traduction (2)
point négatif : participe passé (« assis », « fermé ») au lieu d'un V conjugué (ou aux conjugué + participe), ce qui rend ces phrases agrammaticales.
mieux que (1) : Ark. bien traduit par Arkansas.
point négatif : pb d'accord entre « signes » et « trouvées » typique d'un système statistique, cf. la distance entre les 2 mots est importante, et donc le modèle de langue ne peut pas gérer l'accord.

point négatif : le SN complexe « japanese-built navigation and climate control systems » est très mal traduit : ne sont respectées ni la portée de la coordination (sans doute [[[navigation and climate] control] systems]) , ni la portée de « japanese-built » qui modifie « control systems ».
mieux que (1) : le terme « electric grid » est localement bien traduit en « réseau électrique »
point négatif : transfert lexical « sit »=> « assis » très mauvais pour un sujet inanimé
mieux que (1) : « northern Japan » => « dans le nord du Japon » est bien sûr mieux que « le Japon du Nord ». Comme c'est un phénomène local, c'est bien capturé par le système statistique. Pour le système par règles, il faudrait coder dans le lexique l'unité « northern Japan »....

5 Apprentissage / tagger de Brill (5pts)

5.1 A quoi sert un tagger ?
Assigner la catégories morpho-syntaxiques de chaque mot, en utilisant le contexte pour désambiguiser les cats ambigus.

5.2 Rappelez l'algorithme d'apprentissage du tagger de Brill, dans sa version supervisée (=version vue en cours).
Voir cours.