Analyses statistiques sur un corpus parallèle français-chinois

Compte rendu pour le cours de statistiques textuelles
Pan Wenzhu, Paris Nanterre
23 mai, 2019

Présentation du projet et du corpus

Pour mieux traduire, il n'est pas facile de progresser par l'introspection, car on se croit avoir donné déjà de son mieux. Ainsi, par cette petite mise en pratique de statistique textuelle, j'essaie de mener des expériences sur quelques hypothèses que nous avons avancés sur la traduction française-chinoise.

Au début, j'ai voulu travailler sur le corpus constitué de mes traductions, l'idée que je devais abandonner lorsque j'entamais réellement la tâche. Cela pour plusieurs raisons, d'abord le corpus que j'ai pu collectionner n'est pas propre: il y a une vingtaine de traduction, couvrant des domaines très vastes et très variés, et j'avais dû changer le style de traduction pour mieux adapter au genre (il y a 4 livres scientifiques, le reste sont de la fiction d'enfant ou d'adulte). Ensuite, faute de conscience d'archivistes, les traductions que j'ai pu retrouvées ne sont pas celles de publication. Enfin et surtout, le texte français faisant défaut, je ne vois pas comment concrétiser mon objectifs. Si je ne m'appuie que sur mes propres productions, cela revient à l'introspection.

Ainsi, je me suis décidé à faire une analyse avec le texte français et la traduction chinoise. En faisant cela, j'espère, à part de vérifier les hypothèses, je pourrai trouver des caractéristiques dans la paire de langue, au niveau lexique et syntaxique.

Voici mes hypothèses non-vérifiées, que j'avais de l'intuition:

1, au niveau de l'article, la traduction chinoise est moins longue que le texte français (des collègues éditeurs chinois disent que la traduction chinoise représente 80% de la longueur du texte français);
2, au niveau de la phrase, les phrases en français sont plus longues que les phrases en chinoise;
3, au niveau de l'article, nous avons autant de noms, d'adjectifs dans le texte français et sa traduction chinoise;
4, pour tous les articles d'une langue, si nous établissons une liste de POS en nombre d'occurence décroissant, nous allons trouver le même ordre de POS dans la langue traduite;
5, dans les 50 mots les plus utilisés de chaque langue, il existe des paires de traductions.

J'avais commencé avec l'ambition de constituer un corpus de taille importante avec une centaine d'article, mais en réfléchissant bien, surtout en faisant le nettoyage des textes et la traduction, je me suis limité à constituer un corpus de taille petite (10 articles dont la qualité de tra-

duction est garantie), dont 7 ont été crawlé du site de l'ambassade de France en Chine, 3 du site du ministère des affaires étrangères. Pour avoir une cohérence de thème, les articles sont tous du domaine politique international.

Outil linguistique et informatique

Les hypothèses que nous avons énumérées se situent surtout au niveau de la syntaxe et du paragraphe, c'est à dire que nous aurons besoin de segmenteur de phrase. Pour le français, j'ai pris un module de Python, NLTK; pour le chinois, j'ai pris une fonction attestée de l'internet.

Et pour les analyses partie du discours (POS), nous avons besoin de POS-taggeur. Pour le français, j'ai pris treetagger pour la facilité de manipulation de la sortie; du côté chinois, j'ai pris thulac (http://thulac.thunlp.org), fait par Natural Language Processing and Computational Social Science Lab, Tsinghua University.

Les hypothèses et les expériences

1, la traduction chinoise est moins longue que le texte français;

Nous calculons le nombre de caractère des articles chinois, et le nombre de token des articles français, et comparer les deux nombres. Pour le chinois, nous utilisons la fonction len ( ) du python, qui renvoie le nombre de caractères; et pour le français, la fonction word_tokenize de NLTK. Le résultat s'avère tout de suite surprenant : au contraire de l'hypothèse, au niveau de token, le chinois est plus long que le français. Nous avons calculer le nombre de tokens de dix articles, il n' y a qu'un seul article dont la traduction chinoise a moins de token que le texte français.

Ensuite, nous calculons la somme de différence de 10 articles, la somme divisée par 10, nous avons le résultat 264.3. C'est-à-dire, au niveau du nombre de tokens, pour les articles collectionnés, le texte chinois est 264.3 plus que celui en français.

TABLE 1

article	écart de nombre de tokens (nombre token chinois - nombre token français)
1	306 - 259 = 47
2	1258 - 1075 = 183
3	4746 - 3407 = 1339
4	843 - 889 = -46
5	583 - 576 = 7

6	490 - 333 = 157
7	532 - 368 = 164
8	1333 - 1026 = 307
9	536 - 388 = 148
10	1339 - 1002 = 337

A la fin, pour savoir d'où vient cette idée que le chinois est moins long que le français, j'ai copié un texte français et la traduction en chinois dans Word de Microsoft (en chinois, 4746 tokens; en français, 3407 tokens), mise en page pareil, l'article français prend 4 pages alors que l'article chinois prend 3 pages.

En fait, quand les éditeurs disent que le chinois est moins long, c'est par rapport au nombre de pages, ce qui est facile à expliquer: avec le même espacement de ligne, les tokens français occupent plus d'espace par rapport à l'idéogramme de chinois.

2, la phrase française est plus longue que la phrase en chinois;

L'idée vient surtout de la lecture de Marcel Proust, dont les phrases semblent interminables; et le français, avec une multitude de conjonctions et interjections, permet une phrase pleine d'arrêt et de reprise, vrai casse-tête pour les traducteurs. Et pour comparer la longueur de phrase, nous allons employer la même méthode que l'hypothèse une, c'est-à-dire, calculer le nombre de token dans la phrase. Maintenant que l'on a le nombre de token pour chaque article de la première hypothèse, nous n'avons qu'à calculer le nombre de phrase pour chaque article.

TABLE II

article	n de phrase dans l'article français	n de phrase dans la traduction	n de tokens par phrase en français	n de tokens par phrase en chinois
1	8	8	32	38
2	26	30	41	41
3	84	93	40	51
4	22	18	40	46
5	13	9	44	65
6	7	7	47	70
7	11	9	33	59
8	29	35	35	38
9	9	9	43	59

10	33	37	30	36
Moyenne			38	50

De ce tableau, nous allons pouvoir confirmer que, dans notre corpus, en moyenne, la phrase chinoise est plus longue que la phrase française. On rappelle que, par longueur, nous désignons le nombre de tokens. Si on essaie d'en chercher une explication, c'est que les traducteurs ont souvent l'habitude de ne pas couper la phrase tout en essayant de traduire des subordonné, des injections, ils sont donc obligés de répéter, de préciser.

Alors, d'où vient cette impression interminablement longue de la phrase en français. Une comparaison entre un corpus de littérature et un corpus de presse pourra sans doute nous donner des idées.

3 et 4, les nombres de POS + liste de POS en ordre décroissant selon le nombre d'occurence;

On va introduire les POS, un pas plus vers la sémantique. Or, avant d'entrer dans les statistiques, il est nécessaire de parler de la segmentation en mot du chinois. Si les deux expériences nous ont montré que au niveau de token, le chinois est plus nombreux dans une phrase, il est pourtant le contraire pour le nombre de mot.

La notion du token et celle du mot commencent à s'emmêler. Si en français, on peut, sans beaucoup trop de débat, assimiler le mot à un token; en chinois, ce n'est pas du tout le cas, car un mot en chinois n'est pas uni gramme, dans la majorité, un mot a 2 caractère(chose=东西, France=法国), ou 4 (hors du commun=出类拔萃, que les grands rendez-vous soient tenus=如期进行), ou 3(victime de catastrophe=遇难者), ou les mots en 1 uni gramme, le plus souvent des mots grammaticaux (de=的， à ou pour=向). Les segmenteurs en mots sont des modèles d'entraînement. Le nombre de mot en chinois va ainsi largement réduire. De ces 10 articles, on compte 7410 mots, alors que du côté français, 9011.

Comme les POS sont nombreux, nous n'allons voir que les sept POS les plus fréquents de l'article. Entre parenthèse, le POS et son nombre d'occurence.

TABLE III

1	('NOM', 49), ('PRP', 38), ('DET:ART', 24), ('ADJ', 24), ('NAM', 18), ('ADV', 12), ('VER:pres', 11)
	('_v', 42), ('_n', 24), ('_w', 17), ('_u', 14), ('_d', 13), ('_ns', 11), ('_r', 11)
2	('NOM', 185), ('PRP', 99), ('PRO:PER', 99), ('VER:pres', 82), ('PUN', 65), ('KON', 64), ('ADV', 62)
	('_v', 194), ('_n', 123), ('_w', 113), ('_r', 87), ('_u', 60), ('_d', 58), ('_p', 35)

3	('NOM', 661), ('PRP', 418), ('ADJ', 303), ('DET:ART', 279), ('PUN', 217), ('KON', 177), ('VER:pres', 175)
	('_v', 630), ('_n', 518), ('_w', 343), ('_u', 229), ('_r', 189), ('_d', 175), ('_p', 163)
4	('NOM', 219), ('PRP', 111), ('DET:ART', 110), ('ADJ', 80), ('VER:pper', 57), ('PUN', 55), ('PRP:det', 55)
	('_v', 142), ('_n', 123), ('_w', 90), ('_a', 41), ('_u', 20), ('_d', 18), ('_c', 12)
5	('NOM', 145), ('PRP', 84), ('DET:ART', 79), ('VER:infi', 41), ('PUN', 38), ('KON', 33), ('ADJ', 32)
	('_v', 97), ('_n', 86), ('_w', 57), ('_a', 37), ('_u', 23), ('_d', 21), ('_r', 8)
6	('NOM', 65), ('ADJ', 48), ('PRP', 34), ('DET:ART', 27), ('NAM', 23), ('PUN', 22), ('PRP:det', 14)
	('_n', 70), ('_v', 45), ('_w', 41), ('_u', 18), ('_ns', 16), ('_c', 15), ('_p', 9)
7	('NOM', 77), ('PRP', 44), ('ADJ', 40), ('NAM', 32), ('DET:ART', 28), ('PUN', 20), ('PRP:det', 20)
	('_n', 62), ('_v', 55), ('_w', 37), ('_d', 24), ('_ns', 23), ('_u', 18), ('_np', 16)
8	('NOM', 223), ('PRP', 121), ('ADJ', 111), ('DET:ART', 104), ('VER:pres', 49), ('PUN', 46), ('KON', 45)
	('_v', 197), ('_n', 164), ('_w', 100), ('_r', 50), ('_u', 48), ('_p', 45), ('_a', 40)
9	('NOM', 85), ('PRP', 55), ('ADJ', 38), ('DET:ART', 36), ('PRP:det', 22), ('VER:pper', 18), ('KON', 17)
	('_n', 68), ('_v', 67), ('_w', 35), ('_u', 22), ('_p', 21), ('_ns', 15), ('_c', 11)
10	('NOM', 211), ('PRP', 108), ('ADJ', 92), ('DET:ART', 79), ('VER:pres', 63), ('ADV', 51), ('KON', 45)
	('_v', 193), ('_n', 132), ('_w', 94), ('_r', 68), ('_d', 66), ('_a', 60), ('_u', 53)
tous les articles mélangés	[('NOM', 1920), ('PRP', 1112), ('ADJ', 827), ('DET:ART', 823), ('PUN', 533), ('VER:pres', 479), ('KON', 468), ('PRO:PER', 391), ('PRP:det', 361), ('ADV', 358), ('VER:infi', 282), ('NAM', 266), ('VER:pper', 254), ('SENT', 242), ('PRO:DEM', 144), ('DET:POS', 140), ('PRO:REL', 119), ('NUM', 77), ('PRO:IND', 44), ('VER:futu', 40), ('VER:ppre', 32), ('PUN:cit', 20), ('VER:subp', 20), ('VER:impf', 19), ('ABR', 17), ('VER:cond', 9), ('SYM', 4), ('VER:subi', 4), ('VER:simp', 3), ('PRO', 1), ('PRO:POS', 1), ('VER:impe', 1)]
	[('_v', 1662), ('_n', 1370), ('_w', 927), ('_u', 505), ('_r', 447), ('_d', 424), ('_a', 384), ('_p', 357), ('_c', 249), ('_m', 195), ('_ns', 194), ('_j', 120), ('_f', 114), ('_t', 97), ('_q', 87), ('', 80), ('_id', 53), ('_np', 40), ('_g', 34), ('_ni', 24), ('_x', 17), ('_s', 10), ('_i', 7), ('_k', 7), ('_nz', 6)]

* signification des pos-tags du chinois:

n/名词/nom

np/人名/entité nommé, nom de personne

ns/地名/entité nommé, nom du lieu

ni/机构名/entité nommé, nom de l'institution

nz/其它专名/entité nommé, autre que les précédents

m/数词/le nombre

q/量词/l'article

mq/数量词/le nombre et l’article
t/时间词/le mot du temps
f/方位词/le mot de l’espace
s/处所词/le mot de l’endroit
v/动词/verbe
a/形容词/adj
d/副词/adv
h/前接成分/préfixe
k/后接成分/suffixe
i/习语/expression figée
j/简称/abréviation
r/代词/pronom
c/连词/conjonction
p/介词/préposition
u/助词/particules
y/语气助词/particules modals
e/叹词/interjection
o/拟声词/onomatopée
g/语素/morphème
w/标点/punctuation
x/其它/autres

Nous avons l’hypothèse que, les éléments qui véhiculent les objects, les noms, les adjectifs par exemple, ont un nombre d’occurence proche. Dans le tableau ci-dessous, on compare le nombre de noms.

TABLE IV

article	écart de nombre de noms (nc français - nc chinois)
1	49-35=14
2	185-139=46
3	661-619=42
4	219-127=92
5	145-88=57
6	65-93=-28
7	77-102=-25
8	223-180=43
9	223-89=134

10	221-162=59

Nous avons pu voir par ce tableau que l'écart de nombre d'occurence entre texte et sa traduction est très varié, de cela nous ne pouvons pas titrer une conclusion sauf que c'est une question que l'on ne pourra trancher qu'avec un corpus beaucoup plus grand.

Or, revenons à la TABLE III, nous remarquons que, dans tous les articles français, le nombre d'occurence de noms occupent toujours la première place, alors que dans sa traduction en chinois, c'est trois fois sur dix; le reste, c'est les verbes qui viennent en premier. Alors qu'en français, les verbes sont bien derrières. Bien sûr, le calcul est un peu rudimentaire car treetagger a un grand nombre de verbes: ver:pres, ver:subp, ver:condi, ver:futur, etc.

On se souvient encore de la phrase de Victor Hugo, « Le mot, c'est le verbe, et le verbe, c'est Dieu ». On pourrait dire, sans se faire protester, qu'en chinois, « les dieux » sont plus nombreux. Pour tenter d'en trouver la cause, nous allons prendre la première phrase du premier article pour voir en détail l'étiquetage.

« L'approche des fêtes du Nouvel An et du Noël orthodoxe doit constituer pour les parties au conflit dans l'Est de l'Ukraine une occasion de se concentrer sur les besoins des populations civiles, qui souffrent depuis trop longtemps de ce conflit et de ses conséquences. »

Des verbes reconnus par treetagger:

doit	VER:pres	devoir
constituer	VER:infi	constituer
concentrer	VER:infi	concentrer
souffrent	VER:pres	souffrir

新年和东正教圣诞节假期的临近应成为乌克兰东部冲突各方的契机，使其能集中关注因冲突及其所造成的后果而长期遭受苦难的平民的需求。

Des verbes reconnus par THULAC:

教_v (n: religion; v:enseigner)
临近_v (n: approche; v: approcher)
应_v (v:devoir)
成为_v (v:constituer)
冲突_v (n:conflit; v:affronter)
使_v (v:permettre)
能_v (v:être capable de)
集中_v (v:se concentrer)
关注_v (v:faire attention à)
冲突_v (n:conflit; v:affronter)

造成_v (v:résulter)
遭受_v (v:souffrir)

D’abord, dans les 12 mots chinois étiquetés comme « verbe », 4 sont faux (mis en rouge, première étiquette entre parenthèse étant la bonne). Les verbes nombreux sont dûs à l’usage fréquent de verbe comme « devoir », « être capable de », « résulter » (ils sont tous uni gramme), pour couper des constituants trop longs de la phrase en français, mais aussi pour avoir des pauses prosodiques à une longueur convenable.

5, les mots les plus utilisés dans les deux langues;

Nous allons faire un pas de plus vers la sémantique: par les mots les plus fréquents, nous espérons voir, à part les mots grammaticaux qui figureront sûrement nombreux, des paires de mots qui sont la traduction l’une de l’autre.

TABLE V

français	('le', 693), ('<unknown>', 557), (',', 483), ('de', 466), ('et', 309), ('du', 286), ('.', 241), ('à', 169), ('être', 149), ('un', 148), ('avoir', 142), ('nous', 129), ('en', 123), ('ce', 115), ('que', 107), ('sur', 94), ('notre', 77), ('dans', 76), ('au', 75), ('pour', 69), ('qui', 57), ('je', 54), ('il', 54), ('@card@', 49), ('devoir', 41), ('tout', 37), ('avec', 37), ('faire', 37), ('pouvoir', 35), ('son', 34), ('vous', 33), ('mais', 31), ('par', 30), ('pas', 27), ('France', 27), ('plus', 26), ('y', 26), ('Chine', 25), ('comme', 24), ('ne', 23), ('européen', 23), ('se', 22), ('ensemble', 22), ('aussi', 22), ('sécurité', 22), (':', 20), ('dire', 20), ('parti', 19), ('climatique', 16), ('politique', 15) * les mots ici sont des lemmes
chinois	(', ', 437), ('的', 400) , ('。', 174), ('在', 119), ('、', 114), ('我们', 105), ('和', 92), ('是', 92), ('我', 66), ('这', 66), ('不', 51), ('中', 49), ('与', 49), ('对', 47), ('国', 47), ('一', 46), ('了', 45), (‘法国’[France], 43), ('家', 42), ('来', 35), ('将', 32), ('两', 31), ('法', 31), ('并', 30), ('就', 30), ('共同', 30), ('向', 29), ('为', 28), (‘中国’[Chine], 27), ('能', 26), ('一个', 26), (‘关系’[relation], 26), (‘气候’[climat], 24), (‘伙伴’[compagnon], 24), (‘领域’[domaine], 23), (‘联合国’[Nation unie], 22), ('其', 21), ('也', 21), ('上', 21), ('党', 21), ('要', 20), ('及', 20), ('因为', 20), (‘国际’[international], 20), ('你们', 19), ('您', 18), ('保持', 18), (‘安全’[sécurité], 17), (‘经济’[économique], 17), (‘方面’[aspect], 17).

* Nous soulignons les noms et les adjectifs en rouge.

parmi les 50 mots les plus fréquents, il y a effectivement des paries de traductions, dont le nombre d’occurence n’est pas équivalent.

Bien sûr, il manque de la cohérence dans cette expérience, car au lieu de traiter les articles un à un, thème par thème, nous avons mélangé tous les articles, ce qui pourra fortement perturber

le nombre d'occurence de certains mots. Nous avons choisi cette méthode car, d'une part, certains articles sont trop courts, d'autre part, le corpus étant un corpus parallèle, le nombre d'occurence de mot et celui de la traduction doivent être à peu près similaire.

Conclusion

Nous avons pu voir, à travers toutes ces expériences, dans notre corpus parallèle français-chinois, que :

1, généralement, le nombre de tokens des textes chinois est plus grand que celui de français; or, au niveau de la mise en page, c'est le texte chinois qui prend moins de page que le français (c'est dans ce sens-là que les gens disent que le chinois est moins « long »).

2, ce n'est pas une bonne idée de comparer des POS directement entre deux langues aussi éloignées, surtout avant une analyse profonde de chaque POS. Pourtant, nous avons pu constater des caractéristiques de traduction, par exemple, l'usage fréquent des verbes en chinois pour introduire des liens sémantiques entre des éléments de la phrase.

3, de l'expérience V, nous avons pu voir que la traduction n'est pas de copier-coller. Nous pouvons retrouver des paires de mots, avec un nombre d'occurence différent dans chaque langue. Et à partir de cela, nous pourrons dire que la sémantique s'adapte à la syntaxique.

Nous savons, et nous regrettons que le corpus soit d'une taille minime, chaque analyse que nous avons pu avancée semble dépourvu de base solide. Dans les prochaines études, il est indispensable d'enrichir le corpus. En même temps, d'autres pistes sont possibles: compter le nombre de pronom pour étudier le phénomène de référence dans chaque langue, compter le nombre de virgule pour étudier la prosodie ou la longueur de chunk.