

Kedi
Alice Wallard

Sommaire

Connaissances

Au cours de ce semestre nous avons acquis des connaissances nous permettant de réaliser deux corpus parallèles dans l'objectif d'entraîner un modèle de traduction automatique au deuxième semestre.

Il nous a fallu réunir des textes qui sont des traductions l'un de l'autre (peu importe le sens) dans un corpus généraliste, qui regroupe des textes issus d'un discours non spécialisé (presse, sites internet non spécialisés, communiqués de presse, sites des organisations internationale, etc.) et dans un corpus spécialisé, qui regroupe des textes issus d'un discours spécialisé.

Nos langues de travail sont : le polonais, le coréen, le chinois. La langue pivot est l'anglais pour des raisons de commodité : trouver des traductions étant un des enjeux principaux de ce travail.

Le choix du domaine de spécialité

À l'origine nous avons choisi de travailler sur un aspect du domaine de l'écologie : la surconsommation, mais le sujet était trop peu délimité et nous avons du mal à savoir où chercher. Assez rapidement nous avons décidé de changer de domaine de spécialité.

Pourquoi la conquête spatiale?

Nous avons choisi de faire un corpus de spécialité sur le domaine la conquête spatiale. On entend par là : les politiques gouvernementales et les avancées technologiques des agences spatiales, ainsi que l'exploration spatiale rendue possible par ces avancées. En effet, le sujet a le vent en poupe. Sur le plan international, la Chine affirme son ambition d'être le grand rival des États-Unis dans ce domaine stratégique, l'Inde émerge comme puissance spatiale, l'Europe continue ses programmes et sa collaboration avec la NASA et les questions écologiques s'appliquent désormais aussi à ce domaine.

L'Agence spatiale européenne présente un site multilingue, la Chine une multitude de sites bilingues et la Corée du sud a également une agence spatiale qui propose un site Internet bilingue anglais-coréen.

Concernant les genres textuels, le domaine spatial permet de regrouper des sous-corpus de plusieurs genres : scientifique, journalistique, juridique. Nous avons surtout obtenu des données textuelles de l'organisation internationale UNOOSA.

Compte tenu de la difficulté de trouver des traductions dans des domaines de spécialité nous avons d'abord cherché à rassembler des textes sur le domaine de spécialité.

Cette étape a été très longue.

Le corpus de spécialité

Les nombreux sites bilingues de République populaire de Chine ne publient pas des traductions des articles malgré leur aspect de site bilingue. Il nous a été impossible de récupérer du contenu bilingue des sites cités en Annexe I, pourtant très documentés dans les deux langues. Nous avons essayé en partant des urls, du code source, en effectuant des recherches sur le site, pour voir si il y avait un semblant de correspondance entre les versions en chinois et en anglais des sites mais rien n'y a fait, il semble que les articles en anglais ne sont pas des traductions du chinois mais des articles rédigés en anglais. Les publications en langues étrangères dans les canaux de communication officiels sont une pratique courante depuis très longtemps en Chine.

Nous avons pensé au site de l'ONU qui regroupe des traductions de qualité mais il fallait un site davantage spécialisé dans le spatial que le site de l'ONU. En cherchant sur le site de l'ONU Vienne avec le mot clé "space" on s'est aperçu qu'il existait une organisation internationale onusienne l' UNOOSA (United Nations Office for Outer Space Affairs). Son rôle est de gérer " [...] *le programme relatif aux utilisations pacifiques de l'espace extra-atmosphérique, qui vise à renforcer la coopération internationale dans le domaine des activités spatiales et de l'utilisation des sciences et techniques spatiales* [...] ". Sur son site on trouve les traductions des nombreux rapports officiels.

C'est là que nous avons trouvé la majeure partie de notre corpus de spécialité.

Nous avons recueilli les données du site de Safran qui a une section Espace¹.

Nous avons recueilli des phrases bilingues et de la terminologie relevant de notre domaine de spécialité dans un gros répertoire personnel qui regroupe des textes bilingues (français-anglais, anglais-chinois, chinois-français) de plusieurs types et genres.

Nous avons pensé à parcourir un site bilingue de recherche académique en physique, dans l'espoir de trouver des abstracts bilingues d'articles d'astrophysique, par recherche de mot clé. En cherchant un mot clé anglais spécialisé dans le site chinois et vice-versa. en vain. Il faudrait pour cela faire une recherche plus poussée sur le site de l'Académie des sciences mais nous n'en avons pas encore eu le temps.

Méthodes d'acquisition des données

Corpus de spécialité

a. Méthode automatique

récupération de contenu sur Internet

La récupération de contenu sur Internet est cruciale dans le cadre d'un traitement quantitatif. Elle pose plusieurs difficultés. La difficulté majeure est que chaque site est différent. Il est donc difficile de faire un script qui pourra être utilisé sur plusieurs sites. Un script est réalisé pour un site.

Quand bien même il s'agit d'adapter un script existant, le travail d'adaptation prend presque autant de temps que de programmer un nouveau script.

1

Nous avons pu faire un script pour récupérer le contenu du site de Safran. Nous devons l'adapter pour le site de la nasa bilingue découvert tardivement, mais qui présente en fait des sources idéales pour notre travail.

Types de sites :

- sites bilingues : la référence à la langue est dans le code source introduite par une balise < hreflang = nom_de_la_langue>, ou lang="zh-CN". Dans ce cas, on passe d'une page dans une langue A vers la page traduite en langue B par un lien dans une balise href. Cela suppose d'avoir dans notre script une fonction qui permet de rebondir d'une page à l'autre en suivant ces balises.
- site qui présentent une traduction sur la même page. C'est le cas des pages en chinois du site "nasachina.cn" qui nous intéresse particulièrement.²

b. Méthode manuelle

Mon indexeur (personnel) regroupe 7 346 fichiers de textes bilingues (français-anglais, anglais-chinois, chinois-français) au sein de répertoires, accumulés au fil des années en français, anglais et chinois. En faisant une recherche parallèle par mots clés de spécialité (ex. space et 空间, cosmic et 宇宙), on trouve des textes bilingues ou des listes terminologiques. On trouve notamment dans ces répertoires des guides des interprètes de l'ONU au format word (tableaux de terminologie). On extrait le vocabulaire en faisant attention à ce qu'il s'agisse bien de mots du domaine car on y trouve de tout. Nous n'avons pas eu le temps d'ajouter à nos corpus les terminologies des interprètes de l'ONU.

Méthode :

- si un mot à plusieurs traductions (ex : space disposal, 空间处置 ou 宇宙处置), on l'écrit deux fois à la ligne pour qu'il ait dans le fichier parallèle sa deuxième traduction, en vue d'un alignement des fichiers.
- si un mot est une expression avec un nom propre, ex : Hubble Space Telescope, 哈勃太空望远镜, on le décompose pour avoir aussi les alignements des mots qui le composent, lorsque ceux-ci sont utiles (space telescope, 太空望远镜, telescope, 望远镜).

Anticiper les problèmes d'alignement : problème de l'adaptation de certaines traductions. Par exemple dans le cadre d'une traduction ponctuelle de l'anglais vers le chinois sur un site spécialisé³, la traduction est adaptée et elle présente des ajouts pour expliciter le contexte de l'article de l'article d'origine. Dans ce cas, on peut anticiper un décalage de l'alignement mais il faut donc lire lire l'article en travers pour repérer les ajouts et les enlever.

Observations des problèmes de conventions typographiques à prendre en compte pour le traitement automatique du nettoyage du texte en vue de la segmentation:

- en anglais:

Le point en anglais pose problème s'agissant de la segmentation en phrases. Le point final de la phrase se retrouve également dans d'autres contextes : dans les nombres décimaux (ex. "22.5"), dans les citations d'articles scientifiques (ex. "*Yan et al.*").

- en chinois : Le point final est une typographie spécifique ce qui nous aidera.

² <https://www.nasachina.cn/>

³ <https://www.nasa.gov/missions/webb/nasas-webb-hubble-combine-to-create-most-colorful-view-of-universe/>

Le traitement des pages pdf : récolte et nettoyage

Nous avons récolté les documents de l'UNOOSA manuellement, en téléchargeant les documents pdf depuis leur site. Il nous fallait récupérer le texte. Nous avons utilisé l'outil PyMuPdf⁴ qui permet de récupérer le texte du pdf. cf. script pdftotext.py

Il a fallu nettoyer les pages récoltés. La numérotation systématique des paragraphes des documents onusien présente un avantage pour le traitement automatique. Chaque paragraphe est précédé du n° du paragraphe suivi d'un point puis d'une espace, ex : "15. ". Cette particularité nous a permis de récupérer le texte constituant le paragraphe, en suivant ce schéma, grâce à des expressions régulières en étant assuré du parallélisme entre les documents dans les deux langues (cf. script).

Nous avons donc deux répertoires : un pour les textes en chinois et un pour les textes en anglais, qui contiennent chacun les textes au format .txt, numérotés de manière parallèle, ex : "ZH_18.txt" dans le répertoire chinois correspond à "EN_18.txt" dans l'autre répertoire.

La manipulation des expressions régulières est délicate, elle nécessite de nombreux tests et une logique implacable. On peut mettre au point une regex qui ne fonctionnera pas si on applique avant une autre regex mise au point pour un autre but (interférence). La mise au point du script de nettoyage a pris du temps.

Corpus généraliste

Nous avons cherché nos corpus sur le site Opus. Les corpus généralistes que nous avons sélectionné sur le site Opus sont des corpus de communiqués de presse, les corpus News Commentary v16 et v11.⁵

- le corpus News Commentary v16:

En regardant le contenu des fichiers, on s'aperçoit que :

1. dans le fichier en anglais : toutes les phrases semblent être bien sur une ligne chacune, mais certaines phrases sont coupées par un retour à la ligne.
2. dans le fichier en chinois : il y a des phrases en anglais et les phrases sont regroupées en paragraphes : plusieurs phrases se suivent sans retour à la ligne.

Si nous regardons le nombre de lignes dans les corpus généraliste avec la commande wc -l, nous avons :

648886 EN_1.txt

175229 ZH_1.txt

après retrait des phrases en anglais dans le chinois (cf. script) :

648886 EN_1.txt

170294 ZH_1.txt

soit un nombre de phrases beaucoup plus important en anglais.

cela est lié

après application des fonctions cf. script (nettoyage_corpus_general_lang.py):

⁴ <https://pymupdf.readthedocs.io/en/latest/recipes-text.html>

⁵ <https://opus.nlpl.eu/News-Commentary-v16.php>

<https://opus.nlpl.eu/News-Commentary-v11.php>

648886 EN_1.txt

463319 ZH_1_retour_ligne.txt

On se rapproche du même nombre de lignes malgré un écart encore important. Cependant, après avoir parcouru le corpus on s'aperçoit qu'il manque des parties de l'anglais dans le texte chinois. Il y a un décalage.

ligne 462887 : 中国的支持至关重要。
作为低收入国家毫无争议的双边最大债主[...]

ligne 555760 : China's support matters immensely.
As the low-income countries' largest bilateral creditor by far, China accounts for roughly 20% of the to[...]

- le corpus News Commentary v11 :

Ce corpus parallèle présente le même nombre de lignes dans les deux langues, et après un copu d'oeil on voit que les fichiers sont alignés. Cependant, en parcourant le fichier en anglais on s'aperçoit que plusieurs communiqués sont en devanagari. En parcourant le chinois on s'aperçoit que plusieurs communiqués sont en anglais et en français. On est dans une problématique de nettoyage opposée à la précédente. Il va falloir repérer des lignes et les enlever parallèlement dans les deux fichiers.

On se pose la question de savoir si on ne laisse pas tout en l'état et que l'on pourra filtrer après l'alignement mais cela ne semble pas être une bonne idée car cela prendra autant de temps et surtout on craint que cela ne fausse notre alignement.

On a fait le nettoyage à la main car à cette étape nous manquions de temps mais nous projetons de mettre au point un script (bash : sed -n ou python : enumerate + Counter()) pour repérer et enlever les lignes dans le fichier en anglais tout en conservant les n° de lignes pour faire la même opération dans le fichier en chinois, ou faire la même opération en parallèle. Nous n'avons pas eu le temps mais le ferons car ce problème est sans doute récurrent dans les corpus téléchargés.

Au final pour les corpus de la version11, on se retrouve avec :

68383 lignes v11.en

68398 lignes v11.zh

Soit un décalage de 15 lignes de plus en chinois.

Nous avons pourtant veillé lorsque nous retirions la correspondance en devanagari dans le fichier en chinois, à ce que les phrases terminant l'article précédent et commençant le suivant soient les mêmes dans les deux langues. Malgré cela il y a un décalage.

En faisant la commande uniq vers un autre fichier, on trouve le même nombre de lignes dans le fichier dédoublonné donc pas de doublons en chinois qui pourraient expliquer ce surplus de lignes par rapport à l'anglais.

- le corpus News Commentary v14 :

Le corpus semble propre à vue d'œil.

Après avoir repéré des doublons avec le corpus v11 et les avoir enlevés, nous avons toujours le même nombre de lignes (125863).

Après avoir retiré les sauts de lignes systématiques entre les communiqués des corpus v14, nous décidons de concaténer le corpus v11 au corpus v14 pour faire notre corpus de base.

```
sed /^$/d v14.en > v14_compact.en  
sed /^$/d v14.zh > v14_compact.zh
```

Nous avons bien toujours le même nombre de lignes.

116130 v14_compact.en

116130 v14_compact.zh

Concaténation:

cat v11.en v14_compact.en > corpus_gen_en.txt

cat v11.zh v14_compact.zh > corpus_gen_zh.txt

On retrouve le décalage de 15 lignes en surplus dans le fichier chinois :

184455 corpus_gen_en.txt

184470 corpus_gen_zh.txt

Il y a bien correspondance au début et à la fin des fichiers mais il est difficile de remonter à l'endroit du fichier où commencerait un décalage. Le temps étant limité, nous décidons de garder ces fichiers en l'état.

Taille et proportionnalité des corpus

Finalement, nous aboutissons aux corpus suivants :

- un corpus généraliste de 110 Mo (52 + 58 Mo).
- un corpus de spécialité de 9,3 Mo.

On pourrait augmenter la taille de notre corpus généraliste mais nous avons décidé que, dans un premier temps, nous garderions une proportion de 8% entre le corpus généraliste et le corpus spécialisé car nous prévoyons d'abord d'augmenter la taille de ce dernier, notamment grâce au site de la nasa. Et nous manquons là encore de temps.

Statistiques avant tokenisation :

le corpus_gen_en.txt contient 8 471 312 mots

le corpus_gen_zh.txt contient 1 657 518 mots

Conclusion sur le nettoyage

Nous devons améliorer notre bibliothèque de fonctions de nettoyage car nous avons dû faire certaines opérations à la main.

Aussi, les documents de l'ONU sont truffés de références à d'autres documents internes, il reste des éléments à nettoyer comme par exemple la chaîne : "A/AC.105/C.1/L.406/Add.6 V.23-02608 4/5", c'est un schéma que nous n'avons pas pris en compte. Nous devons revoir nos regex.

Lors du nettoyage on envisage le corpus comme une masse de données sans regarder leur sens. Il faut être en quelque sorte "impitoyable" et ne pas se laisser attendrir par la sémantique, l'aspect purement linguistique. Cela peut être déconcertant lorsque l'on est en recherche d'informations car on tendance à vouloir la préserver au maximum.

C'est une sorte de règle générale à priori dans le cadre du traitement automatique : il faut se mettre à la place de la machine et revoir notre façon d'aborder la langue, s'en détacher, l'efficacité du traitement et la quantité de données étant les objectifs principaux.

Mais nous souhaitons aller au-delà de ces premiers objectifs quantitatifs incontournables, pour pouvoir affiner notre recherche et valoriser nos corpus sur le plan linguistique. Notamment, en prenant le temps de mieux les sélectionner et les filtrer. Car nous pensons que la qualité linguistique du corpus (terminologie, niveau de langue, respect des genres textuels) est déterminante pour la suite du traitement, le résultat final de la traduction automatique et, de ce fait, pour son utilité.

Tokenisation des corpus

Pour la tokenisation des corpus nous utilisons la bibliothèque Spacy.

Le modèle ne peut pas prendre plus de 1 000 000 de mots à chaque entrée.

Nous pensons donc à scinder nos corpus. Le corpus en anglais compte plus de 8 millions de mots nous le divisons donc par 9, et nous divisons le corpus chinois en 2, le tout manuellement en calculant les coupures au niveau du nombre de lignes.

Lors de cette manipulation, en parcourant à nouveau le corpus anglais nous apercevons des communiqués en indonésien qui sont immédiatement écartés avec leur correspondance en chinois. Ce qui réduit encore notre corpus.

Par contre, cette opération nous a permis de récupérer un nombre de lignes par document par langue plus cohérent :

20492 ./EN/corpus_gen_en_1.txt	20492 ./ZH/corpus_gen_zh_1.txt
20481 ./EN/corpus_gen_en_2.txt	20482 ./ZH/corpus_gen_zh_2.txt
20458 ./EN/corpus_gen_en_3.txt	20472 ./ZH/corpus_gen_zh_3.txt
20486 ./EN/corpus_gen_en_4.txt	20487 ./ZH/corpus_gen_zh_4.txt
20474 ./EN/corpus_gen_en_5.txt	20474 ./ZH/corpus_gen_zh_5.txt
20467 ./EN/corpus_gen_en_6.txt	20467 ./ZH/corpus_gen_zh_6.txt
20490 ./EN/corpus_gen_en_7.txt	20490 ./ZH/corpus_gen_zh_7.txt
20440 ./EN/corpus_gen_en_8.txt	20440 ./ZH/corpus_gen_zh_8.txt
20294 ./EN/corpus_gen_en_9.txt	20294 ./ZH/corpus_gen_zh_9.txt

En réalité, nous n'allons pas donner au modèle Spacy un texte en entier mais des phrases. Mais nous préférons avoir plusieurs fichiers textes plutôt qu'un gros fichier

Nous segmentons les corpus en phrases : cf. script `segmentation_phrases_en()` et `segmentation_phrases_zh()`.

Le point final des phrases en anglais est problématique, nous prenons garde de préciser qu'il doit être précédé d'une lettre minuscule ou d'un guillemet ou d'une parenthèse fermante.

Malgré cela nous ne trouvons pas exactement le même nombre de phrases dans les deux langues par exemple sur les fichiers du corpus spécialisé n° 17, on trouve 36 phrases en anglais et 42 phrases en chinois.

L'écart de nombre de phrases est encore plus grand en testant sur les 3 premiers communiqués de presse du corpus généraliste. 33 phrases pour l'anglais et 104 pour le chinois. Là encore il y a des ajustements à faire.

Il se creuse encore si l'on prend le corpus 1 généraliste qui a 20 492 lignes dans chacune des 2 langues. La segmentation donne : 37 439 phrases en anglais pour 52 962 phrases en chinois.

Ensuite nous tokenisons ces phrases avec Spacy : `tokenize_en()` et `tokenize_zh()`.

L'idée est de faire une chaîne de traitement et d'envoyer le doc Spacy dans l'aligneur Parallel Sentence qui prend en entrée une liste de chaînes.

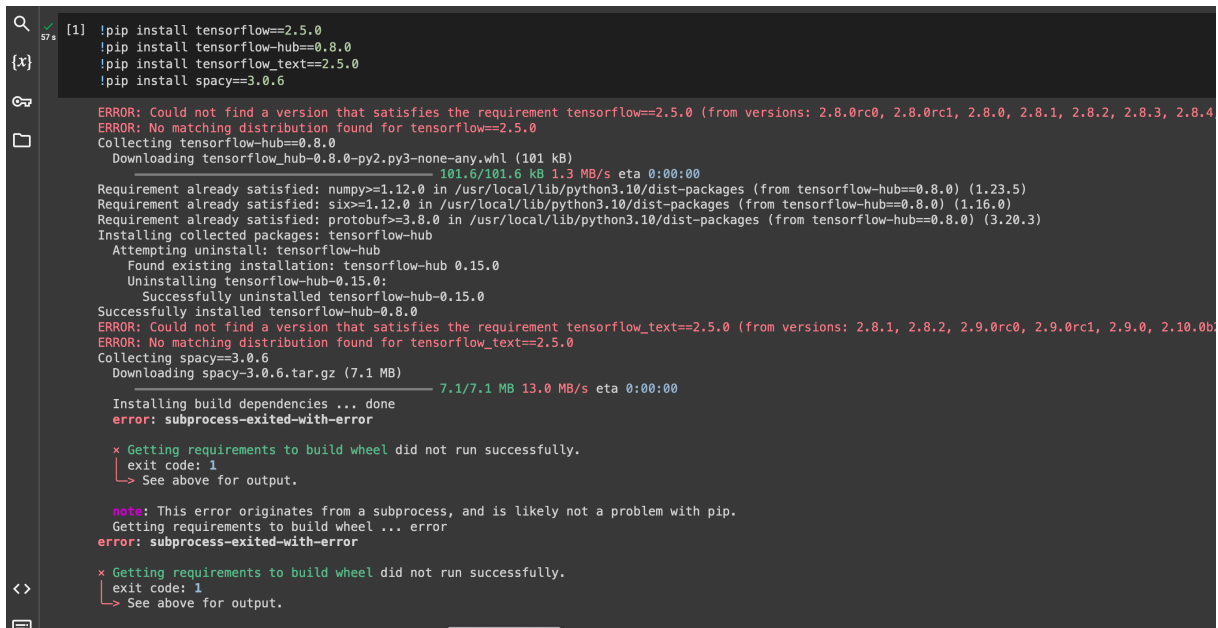
la fonction `tokenize()` retourne une liste de liste chacune étant une phrase tokénisée :

Les 3 premières phrases du corpus généraliste en anglais puis en chinois :

```
[[ '1929', 'or', '1989', '?', '!', '\n'], [ 'PARIS', '-', 'As', 'the', 'economic', 'crisis', 'deepens', 'and',  
'widens', '!', 'the', 'world', 'has', 'been', 'searching', 'for', 'historical', 'analogies', 'to', 'help', 'us',  
'understand', 'what', 'has', 'been', 'happening', '!', '\n'], [ 'At', 'the', 'start', 'of', 'the', 'crisis', '!',  
'many', 'people', 'likened', 'it', 'to', '1982', 'or', '1973', '!', 'which', 'was', 'reassuring', '!', 'because',  
'both', 'dates', 'refer', 'to', 'classical', 'cyclical', 'downturns', '!', '\n'],
```

```
[[ '1929 年', '还是', '1989 年', '?', '\n', '巴黎', '-', '随着', '经济', '危机', '不断', '加深', '和', '蔓延',  
'!', '!', '整个', '世界', '一直', '在', '寻找', '历史', '上', '的', '!', '类似', '事件', '希望', '有助于', '我们',  
'了解', '目前', '正在', '发生', '的', '!', '情况', '!', '!', '\n'], [ '一开始', '!', '!', '很多', '人', '把', '这次', '危机',  
'!', '比作', '1982 年', '或', '1973 年', '所', '发生', '的', '!', '!', '情况', '!', '!', '!', '这样', '得', '!', '类比', '是', '!', '令', '人',  
'宽心', '的', '!', '!', '!', '因为', '这', '两', '段', '时期', '意味着', '!', '典型', '的', '!', '周期性', '衰退', '!', '!', '\n'], [ '如今', '人们', '的', '!', '心情', '却', '是', '!', '沉重', '多', '了', '!', '!', '!', '许多', '人', '开始', '把', '这次', '危机',  
'!', '与', '1929 年', '和', '1931 年', '相比', '!', '!', '!', '即使', '一些', '国家', '政府', '的', '!', '表现', '仍然', '似乎',  
'!', '把', '视', '!', '目前', '的', '!', '情况', '!', '为', '!', '是', '!', '典型', '的', '!', '而', '看见', '的', '!', '衰退', '!', '!', '\n'],
```

Nous n'avons pas pu installer tensor Flow sur Google Colab, alors que nous voulions tester Parallel Sentence pour l'alignement.



```
[1] !pip install tensorflow==2.5.0
!pip install tensorflow-hub==0.8.0
!pip install tensorflow-text==2.5.0
!pip install spacy==3.0.6

ERROR: Could not find a version that satisfies the requirement tensorflow==2.5.0 (from versions: 2.8.0rc0, 2.8.0rc1, 2.8.0, 2.8.1, 2.8.2, 2.8.3, 2.8.4)
ERROR: No matching distribution found for tensorflow==2.5.0
Collecting tensorflow-hub==0.8.0
  Downloading tensorflow_hub-0.8.0-py2.py3-none-any.whl (101 kB)
    101.6/101.6 kB 1.3 MB/s eta 0:00:00
Requirement already satisfied: numpy>=1.12.0 in /usr/local/lib/python3.10/dist-packages (from tensorflow-hub==0.8.0) (1.23.5)
Requirement already satisfied: six>=1.12.0 in /usr/local/lib/python3.10/dist-packages (from tensorflow-hub==0.8.0) (1.16.0)
Requirement already satisfied: protobuf>=3.8.0 in /usr/local/lib/python3.10/dist-packages (from tensorflow-hub==0.8.0) (3.20.3)
Installing collected packages: tensorflow-hub
  Attempting uninstall: tensorflow-hub
    Found existing installation: tensorflow-hub 0.15.0
    Uninstalling tensorflow-hub-0.15.0:
      Successfully uninstalled tensorflow-hub-0.15.0
  Successfully installed tensorflow-hub-0.8.0
ERROR: Could not find a version that satisfies the requirement tensorflow-text==2.5.0 (from versions: 2.8.1, 2.8.2, 2.9.0rc0, 2.9.0rc1, 2.9.0, 2.10.0b)
ERROR: No matching distribution found for tensorflow-text==2.5.0
Collecting spacy==3.0.6
  Downloading spacy-3.0.6.tar.gz (7.1 MB)
    7.1/7.1 MB 13.0 MB/s eta 0:00:00
Installing build dependencies ... done
error: subprocess-exited-with-error

x Getting requirements to build wheel did not run successfully.
  exit code: 1
  -> See above for output.

note: This error originates from a subprocess, and is likely not a problem with pip.
Getting requirements to build wheel ... error
error: subprocess-exited-with-error

x Getting requirements to build wheel did not run successfully.
  exit code: 1
  -> See above for output.

note: This error originates from a subprocess, and is likely not a problem with pip.
```

C'est dommage, mais nous allons continuer à tester les outils et à avancer dans notre chaîne de traitement.

Nous avons essayé de télécharger le corpus généraliste 1 sur Sketch Engine pour évaluer l'alignement et la segmentation de nos phrases. Mais la quantité de données était trop importante pour l'espace alloué avec notre compte.

Conclusion :

Ce travail nous a beaucoup appris. Nous avons réalisé la complexité de la construction de gros corpus exploitables.

Le facteur quantitatif change notre façon de travailler à toutes les étapes par rapport à ce que l'on a l'habitude de faire en cours.

Pour la réalisation du corpus de spécialité, il nous a fallu trouver des quantités de données mais aussi il a fallu également arriver à les nettoyer au mieux et les règles par expression régulières ne sont pas faciles à déterminer et à appliquer sur de gros volumes de textes sans qu'il y ait des oublis ou pire des interférences qui nous feraient perdre du texte. On s'aperçoit que l'on ne peut pas travailler dans le détail mais qu'il faut bien entendu travailler quand même au niveau de la phrase pour définir les règles de nettoyage.

Pour la mise au point des scripts, le fait de travailler sur de gros volumes de données nous oblige à adapter nos scripts et à utiliser des plateformes comme Google Colab, mais nous avons eu des problèmes avec l'installation de la version demandée de Tensor Flow et le temps nous manquait.

Nous n'avons pas pu aller jusqu'au bout de l'évaluation de notre corpus, mais nous savons comment continuer le travail et comptons le mener à bien.

Annexe I : les sites bilingues chinois-anglais liés au domaine de l'espace

Les sites bilingues exploitables :

- site du Bureau des affaires spatiales des Nations Unies (UNOOSA), Organization of the Office for Outer Space Affairs, 外层空间事务厅 :

- en anglais : <https://www.unoosa.org/oosa/en/aboutus/contact.html>
- page des documents officiels traduits dans les langues de l'ONU dont le chinois : https://www.unoosa.org/oosa/documents-and-resolutions/search.jsp?lf_id=

- site de la nasa, bilingue sur une même page :

- <https://www.nasachina.cn/>

- site de l'Académie des sciences, Chinese Academy of Sciences, 中国科学院 :

- en anglais : <https://english.cas.cn/>
- en chinois : <https://www.cas.cn/>

- site de la société Safran, partie Espace :

- en anglais : <https://www.safran-group.com/group/profile/space>
- en chinois : <https://www.safran-group.com/cn/qun-ti/jieshao/kongjian>

Les sites plus adaptés à des corpus comparables :

- site de l'agence spatiale chinoise :

- en anglais : <https://www.cnsa.gov.cn/english/>
- en chinois : <https://www.cnsa.gov.cn/>

- site du Centre national des sciences spatiales (National Space Science Center), NSSC, 国家空间科学中心 :

- en anglais : <http://english.nssc.cas.cn/>
- en chinois : <http://www.nssc.cas.cn/>

- site de la société Beijing Interstellar Glory Space Technology Co., Ltd., 北京星际荣耀空间科技股份有限公司 :

- en anglais : <http://www.i-space.com.cn/>
- en chinois : <http://www.i-space.com.cn/>

- site de la China Aerospace Science and Technology Corporation (CASC):

- en anglais : <http://english.spacechina.com/>
- en chinois : <https://www.spacechina.com/n25/index.html>