Corpus et méthode expérimentale

module X9IT080

Florian Boudin

Département informatique, Université de Nantes

Révision 1 du 30 juillet 2012

Présentation du module

- Notions abordées dans ce module
  - ...
- Volume horaire : 15 séances (20h)
  - 6 CM (8h) - 6 TD (8h) - 3 TP (4h)
- Notation

Plan

Définitions I

Corpus

Un **corpus** est un ensemble de documents (textes, images, vidéos, etc.) regroupés dans une optique précise.

Dans le cadre de ce cours : corpus → corpus de textes

Plusieurs caractéristiques sont à prendre en compte pour la création d'un corpus bien formé :

- la taille ;
- le langage ;
- le temps couvert par les textes du corpus ;
- le registre de langage.

Source : `http://fr.wikipedia.org/wiki/Corpus`

Définitions II

Taille

Le corpus doit évidemment atteindre une taille critique pour permettre des traitements statistiques fiables.

- Si un corpus est utilisé pour construire des modèles de langue, quelle doit être sa taille minimum ? 1M mots, 10M mots, etc.

Langage

Un corpus **monolingue** bien formé doit nécessairement couvrir une seule langue, et une seule déclinaison de cette langue (e.g. français de France et français du Québec).

Définitions III

Période couverte

Le temps joue un rôle important dans l'évolution du langage : le français parlé aujourd'hui ne ressemble pas au français parlé il y a 200 ans ni, de façon plus subtile, au français parlé il y a 10 ans, à cause notamment des néologismes.

Registre de langage

Un corpus construit à partir de textes scientifiques ne peut être utilisé pour extraire des informations sur les textes vulgarisés, et un corpus mélangeant des textes scientifiques et vulgarisés ne permettra pas de tirer de conclusion sur ces deux registres.

Les différents types de corpora

La notion de corpus

- Corpus parallèle
- Corpus comparable
- La constitution de corpus
- L'annotation de corpus

Corpus parallèle

Un corpus parallèle est un ensemble de paires de textes tel que, pour une paire, un des textes est la traduction de l'autre.

Construire un corpus parallèle nécessite un **alignement** des unités textuelles :

- mettre en correspondance des unités textuelles en langue source avec celles de la langue cible.

L'alignement des unités textuelles peut être manuel ou automatique.

La granularité de l'alignement (documents, phrases, mots) dépend de l'utilisation du corpus :

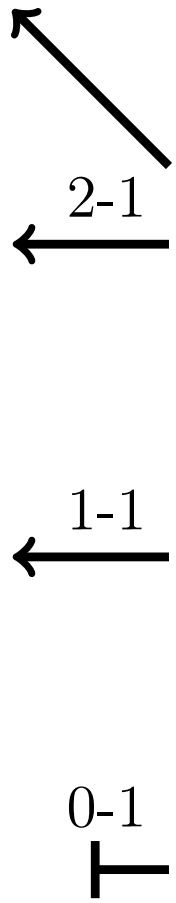
- traduction automatique, génération de paraphrases, construction dictionnaires bilingues, etc.

Alignement de phrases

Il s'agit de l'un des sauropodes les plus connus.

C'était un très grand quadrupède au long cou, avec une longue queue en forme de fouet.

Ses membres antérieurs étaient légèrement plus courts que ses membres postérieurs, ce qui lui donnait une posture horizontale.



One of the best-known sauropods, Diplodocus was a very large long-necked quadrupedal animal, with a long, whip-like tail.

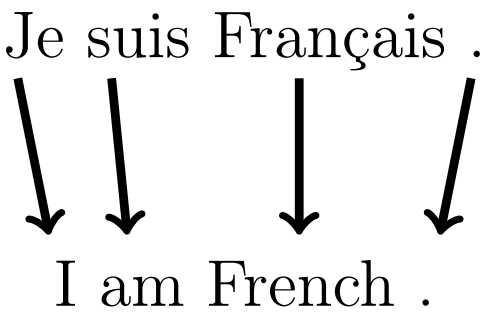
Its forelimbs were slightly shorter than its hind limbs, resulting in a largely horizontal posture.

It is the longest dinosaur known from a complete skeleton.

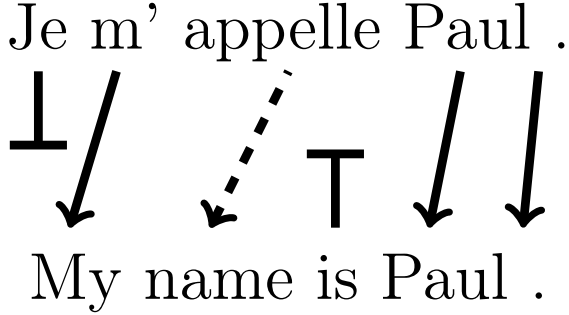
Source : `http://en.wikipedia.org/wiki/Diplodocus`

Alignement de mots

Exemple d'alignment simple.



Exemple d'alignment plus compliqué.



Corpora parallèles disponibles

- Europarl [Koe05]
  - Délibérations du Parlement européen disponibles en 21 langues et alignées au niveau de la phrase.
- OPUS [Tie09]
  - Collection de corpora parallèles issus de sources variées : EMEA (European Medicines Agency documents), KDE4 (KDE4 localization files), Europarl, OpenSubtitles, etc.
  - `http://opus.lingfil.uu.se/`
- Les livres numériques librement disponibles.
  - `http://www.gutenberg.org/`
- Wikipédia : plus comparable que parallèle

Corpus comparable I

- Les corpora parallèles sont très couteux à produire et ils ne sont disponibles que dans un nombre de langues/domaines réduit.
- Les corpus dits **comparables** sont largement plus répandus.
- D'après Déjean & Gaussier [DG02] :
  - Deux corpus de deux langues l_1 et l_2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue l_1, respectivement l_2, dont la traduction se trouve dans le corpus de langue l_2, respectivement l_1.

Corpus comparable II

- Exemple : un ensemble d'articles de journaux dans différentes langues, traitant d'une même actualité et à la même époque.
- Applications :
  - Extraction de phrases parallèles [SQT10]
  - Constitution de dictionnaires bilingues [Rap99]
- Peu (ou pas ?) de corpora comparables disponibles.
  - Wikipedia

La constitution de corpus

1. Définir les caractéristiques du corpus :
   - Quels sont les phénomènes que l'on souhaite observer ?
   - Quelle est la tâche que l'on souhaite réaliser ?
   - Le corpus doit-il pouvoir être distribué ?

2. Assembler les unités textuelles (documents, phrases, etc.)
   - Est-ce que le processus peut être automatisé ?
     - e.g. moissonnage à partir du web (*web scraping*).
   - Une sélection manuelle est-elle nécessaire (et possible) ?

   → Garder en mémoire la méthodologie utilisée (e.g. README).

Opt. Annotation du corpus
   - Annoter les unités textuelles (e.g. *Part-Of-Speech*).
   - Créer un référenciel pour l'évaluation (e.g. termes-clés).

   → Définir des *guidelines* à joindre au corpus.

Exemple 1

Corpus pour évaluer un système d'extraction de termes-clés

- Caractéristiques du corpus
  - On souhaite évaluer un système d'extraction de termes-clés et pouvoir distribuer le corpus pour des raisons de comparaison.

  → Choix de la nature des documents : articles, blogs, tweets, etc.

  → Langue(s) des documents : anglais, français, etc.

  → Source(s) des documents : lemonde.fr, wikinews, etc.

  → Nombre de documents : 10, 20, 50, 100, 1000, etc.

- Récupération des documents
  - Automatiser le processus à partir d'un *dump* de wikinews

  → Filtrage (manuel) des documents trop courts
  - Définir un format pour les documents (XML, HTML, txt, etc.)

- Création d'un référentiel pour l'évaluation
  - Tâche subjective → plusieurs annotations par documents

Exemple 2

Corpus de phrases analysées en dépendances

- Caractéristiques du corpus
  - On souhaite créer un corpus composé de phrases analysées en dépendances afin d'étudier des phénomènes linguistiques et d'entraîner un outil d'analyse en dépendances.

  → Choix de la nature des phrases : ____

  → Langue(s) des phrases : ____

  → Source(s) des phrases : ____

  → Nombre de phrases : ____

- Récupération des phrases

  → Récupération automatisée ou manuelle ?

- Annotation des phrases
  - Recruter des spécialistes ? définir des *guidelines* ?

L'annotation de corpus I

- Différents niveaux d'annotation :
  - collection de documents, e.g. regroupement par sujets
  - document, e.g. jugement de pertinence en RI
  - paragraphe, e.g. découpage thématique
  - phrases, e.g. segmentation en phrases
  - multi-mots, e.g. détection d'entités nommées
  - mots, e.g. tokenization
  - caractères, e.g. analyse morphologique
- Critères de sélection des annotateurs
  - Spécialistes du domaine, e.g. médecin pour de la RI médicale
  - Natif de la langue, e.g. évaluation de la grammaticalité
  - Extérieur au projet, e.g. annotation par les auteurs...
  - Attention au *crowdsourcing* !

L'annotation de corpus II

- L'annotation d'un point de vue technique
  - Séparation entre les unités et les annotations
  - Format (e.g. XML) et encodage standard (e.g. UTF-8)
- Variabilité des annotations (*consistency*)
  - Plusieurs annotateurs → annotations différentes
    - Résoudre les désaccords par consensus ?
    - Considérer plusieurs annotations comme vraies ?

Exemple d'annotation de corpus

Détection de sentiment dans les tweets

- Détecter la **polarité** d'un tweet
  - Annoter un sentiment négatif, neutre ou positif

1. @pcinpact
   Décès de Whitney Houston : la vilaine bourde de Sony Music
2. @PoufyGB
   Interview des #DaftPunk sur France Inter 1 mois après la sortie de l'album
3. @Zzzzzzzz
   Seuls 126% des marseillais exagèrent.

Exemple d'annotation de corpus

Détection de sentiment dans les tweets

- Détecter la **polarité** d'un tweet
  - Annoter un sentiment négatif, neutre ou positif

1. @pcinpact → négatif
   Décès de Whitney Houston : la vilaine bourde de Sony Music
2. @PoufyGB → neutre
   Interview des #DaftPunk sur France Inter 1 mois après la sortie de l'album
3. @Zzzzzzzz → ?
   Seuls 126% des marseillais exagèrent.

Un point sur l'encodage

- La lecture/écriture de fichiers contenant du texte suppose l'utilisation d'un encodage des caractères.
  - Anglais : ASCII
  - Français : ISO-8859-1
  - Japonais : ISO-2022-JP
  - etc.

Un point sur l'encodage

- La lecture/écriture de fichiers contenant du texte suppose l'utilisation d'un encodage des caractères.
  - Anglais : ~~ASCH~~
  - Français : ~~ISO-8859-1~~
  - Japonais : ~~ISO-2022-JP~~
  - etc.

$\rightarrow$ **UTF-8**

Un point sur les méta-données

Plan

Les normes de caractérisation et d'annotation

- Partie de Béatrice

Plan

Démarche expérimentale

- Comment évaluer (de manière rigoureuse) les performances d'un outil ?

Plan

References I

Hervé Déjean and Eric Gaussier.
Une nouvelle approchea l'extraction de lexiques bilinguesa partir de corpus comparables.
Lexicometrica, Alignement lexical dans les corpus multilingues, pages 1–22, 2002.

Philipp Koehn.
Europarl : A parallel corpus for statistical machine translation.
In MT summit, volume 5, 2005.

Reinhard Rapp.
Automatic identification of word translations from unrelated english and german corpora.
In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pages 519–526. Association for Computational Linguistics, 1999.

References II

Jason R Smith, Chris Quirk, and Kristina Toutanova.
Extracting parallel sentences from comparable corpora using document level alignment.
In Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 403–411. Association for Computational Linguistics, 2010.

Jörg Tiedemann.
News from opus-a collection of multilingual parallel corpora with tools and interfaces.
In Recent Advances in Natural Language Processing, volume 5, pages 237–248, 2009.