Corpus et méthode expérimentale

module X9IT080

Florian Boudin

Département informatique, Université de Nantes

Révision 1 du 19 août 2013

Présentation du module

- notions abordées dans ce module
  - corpus : constitution, analyse et normes d'annotation
  - démarche scientifique : évaluation, correlation, significativité
- volume horaire : 15 séances (20h)
  - 6 CM (8h) - 6 TD (8h) - 3 TP (4h)
- notation
  - projet de création de corpus

Plan

Définitions I

- un **corpus** est un ensemble de documents (textes, images, vidéos) regroupés dans une optique précise
- dans le cadre de ce cours : corpus → textes
- caractéristiques d'un corpus :
  - la taille
  - le langage
  - le temps couvert par les textes du corpus
  - le registre de langage

Définitions II

- le corpus doit atteindre une **taille** critique pour permettre des traitements statistiques fiables
  - 10K+ phrases pour entraîner un *POS-tagger*
  - 20M+ mots pour construire un modèle de langue
- un corpus **monolingue** couvre une seule langue, et une seule déclinaison de cette langue
  - e.g. français de France et français du Québec

Définitions III

- la **période couverte** par les textes doit être courte
  - français d'aujourd'hui $\neq$ français parlé au 19^e siècle
  - français d'aujourd'hui $\neq$ français de 1990 (néologismes)
- les **registres de langage** doivent être séparés
  - un corpus de textes scientifiques ne peut être utilisé pour extraire des informations sur les textes vulgarisés
  - un corpus de textes scientifiques et vulgarisés ne permettra pas de tirer de conclusion sur ces deux registres

Corpus parallèle

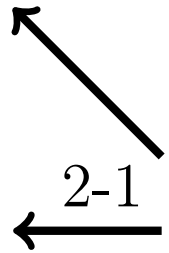
- un **corpus parallèle** est un ensemble de paires de textes tel que, pour une paire, un des textes est la traduction de l'autre
- un **alignement** des unités textuelles est nécessaire :
  - mettre en correspondance des unités textuelles en langue source avec celles de la langue cible
- un alignement peut être manuel ou automatique
- la granularité de l'alignement dépend de l'utilisation
  - traduction automatique, dictionnaires bilingues, etc.

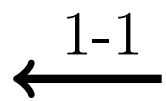
Alignement de phrases

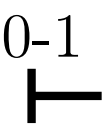
Il s'agit de l'un des sauropodes les plus connus.

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.





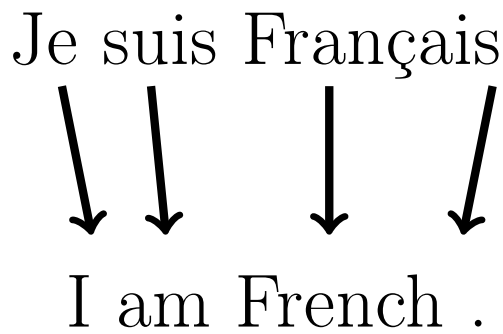


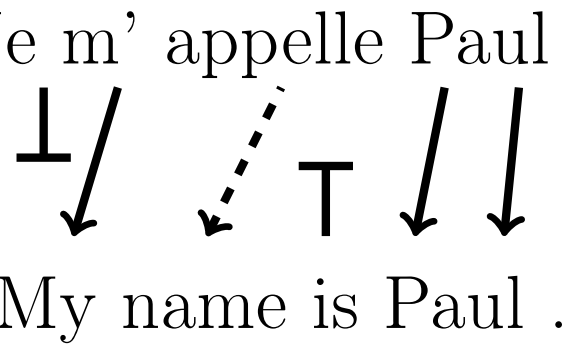
One of the best-known sauropods, Diplodocus was a very large long-necked quadrupedal animal, with a long, whip-like tail.

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.

Alignement de mots





Corpora parallèles disponibles

- Europarl [Koe05]
  - délibérations du Parlement européen disponibles en 21 langues et alignées au niveau de la phrase.
- OPUS [Tie09]
  - *European Medicines Agency documents*, *KDE4 localization files*, Europarl, OpenSubtitles, etc.
  - `http://opus.lingfil.uu.se/`
- livres numériques librement disponibles
  - `http://www.gutenberg.org/`
- Wikipédia : plus comparable que parallèle

Corpus comparable I

- les corpora parallèles sont très couteux et ils ne sont disponibles que dans peu de langues/domaines
- les corpus dits **comparables** sont 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 de corpus comparable :
  - 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
  - Wikipédia, le web multilingue, etc.

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
   - est-ce que le processus peut être automatisé ?
     - e.g. moissonnage à partir du web (*web scraping*)

   → garder en mémoire la méthodologie utilisée

3. annotation du corpus (optionnel)
   - 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

Étude de cas : extraction de termes clés

- un corpus pour entraîner et évaluer un système d'extraction de termes-clés
  - → 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
  - définir un format pour les documents (XML, txt, etc.)
- création d'un référentiel pour l'évaluation
  - tâche subjective → plusieurs annotations par document

Étude de cas : analyse en dépendances

- un corpus de phrases annotées en dépendances entraîner un *parser*
  - → langue(s) des phrases : ____
  - → source(s) des phrases : ____
  - → nombre de phrases : ____
- récupération des phrases
  - → 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
- l'annotation d'un point de vue technique
  - séparation des annotations du texte
  - format (e.g. XML) et encodage standard

L'annotation de corpus II

- critères de sélection des annotateurs
  - spécialistes, 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* !
- variabilité des annotations (*consistency*)
  - résoudre les désaccords par consensus ?
  - considérer plusieurs annotations ?

Étude de cas : annotation de tweets

- détecter la **polarité** d'un tweet
  - annoter un sentiment négatif, neutre ou positif

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

Étude de cas : annotation de tweets

- détecter la **polarité** d'un tweet
  - annoter un sentiment négatif, neutre ou positif

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

Un point sur l'encodage

- la lecture/écriture de fichiers textes 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 textes suppose l'utilisation d'un encodage des caractères
  - anglais : ~~ASCH~~
  - français : ~~ISO-8859-1~~
  - japonais : ~~ISO-2022-JP~~
  - etc.

→ UTF-8

Un point sur les méta-données

- une **méta-donnée** est une donnée servant à définir ou décrire une autre donnée
  - e.g. associer à une donnée la date de création, à une photo les coordonnées GPS du lieu où elle a été prise
- le *Dublin Core* est la principale initiative visant à la convergence des éléments de méta-données à utiliser
- 15 éléments de description :
  - formels (titre, créateur, éditeur)
  - intellectuels (sujet, description, langue, etc.)
  - relatifs à la propriété intellectuelle (droits)

Plan

Les normes de caractérisation et d'annotation

- Partie de Béatrice

Plan

Introduction I

- **démarche expérimentale** → valider une hypothèse par des expériences
  - e.g. expérience du cerf-volant de Benjamin Franklin
- exemple
  - hypothèse → segmenter le chinois avec les CRF
  - expérience → évaluer la précision sur un corpus
- **évaluer** un système de TAL consiste à vérifier qu'il produit le résultat pour lequel il a été conçu

Introduction II

- le TAL englobe un grand nombre de tâches, chacune ayant des critères particuliers quant à son évaluation
- cette partie est basée sur [RL10]
  - `http://www.umiacs.umd.edu/~jbg/teaching/CMSC_773_2012/reading/evaluation.pdf`

Exemple I

- un exemple pour introduire les idées que nous allons voir en détails par la suite
- évaluer un moteur de recherche sémantique
  - requêtes analysées et converties en sujet-relation-objet
- *when was the light bulb patented by Edison ?*

  → [Edison, patented, bulb]
  - permet de retrouver le document « *Thomas Edison's patent of the electric light bulb* »

Exemple II

- comment évaluer le composant d'analyse en sujet-relation-objet ?

→ conduire une évaluation **intrinsèque**

- créer un ensemble de questions analysées manuellement
- évaluer la performance avec des mesures de P-R-F
- comparer deux versions (*formative evaluation*)
- comparer à d'autres méthodes (*summative evaluation*)

Exemple III

- est-ce qu'une analyse précise permet d'améliorer les résultats du moteur de recherche ?

→ conduire une évaluation **extrinsèque**

- évaluer l'impact de l'analyse sur la performance du moteur
- utiliser une collection de test avec des mesures de P-R-F

→ conduire évaluation *in situ*

- proposer le système aux utilisateurs et les observer

Concepts fondamentaux I

- **évaluation manuelle**
  - demander à des sujets humains d'évaluer un système selon des critères pré-définis → souvent la meilleure évaluation !
  - nombreuses limitations : coût important, évaluation lente, résultats inconsistants et non reproductibles
- **évaluation automatique**
  - création d'un *gold standard*, *ground truth*, etc.
  - nécessite une mesure qui « simule » une évaluation manuelle
  - correlation entre les évaluations manuelles et automatiques

Concepts fondamentaux II

- **évaluation intrisèque**
  - la sortie du système est évaluée directement par rapport à des critères pré-définis
- **évaluation extrisèque**
  - la sortie du système est évaluée à travers son impact sur une tâche externe
- exemple du résumé automatique
  - évaluation intrisèque → ROUGE ou évaluation manuelle
  - évaluation extrisèque → les résumés sont-ils utiles pour remplacer les snippets d'un moteur de RI ?

Concepts fondamentaux III

- **accord inter-annotateurs**
  - en TAL, l'évaluation se résume à annoter du texte
  - comparer la performance de plusieurs annotateurs
  - un accord inter-annotateurs faible

    → tâche trop difficile ou mal définie
  - le taux d'accord inter-annotateurs constitue la limite haute (**upper bound**) de ce qu'il est possible d'évaluer
- de nombreuses mesures d'évaluation
  - coefficient kappa de Cohen
  - voir [AP08] pour plus de détails

Le découpage des données

- la plupart des évaluations impliquent un découpage des données en ensembles **disjoints**
  - **training data** (70%)
  - **development data** (20%)
  - **test data** (10%)
- **évaluation croisée**
  - permet d'évaluer sur toutes les données disponibles
  - découper l'ensemble de données en k partitions
    - entraîner sur $k-1$ partitions et tester sur la partition restante
    - calculer la performance moyenne sur les k partitions

Présenter des résultats

- mesures/métriques pour estimer la performance
- toujours inclure au moins une *baseline*

Système	**Mesure 1**	**Mesure 2**	**Mesure combinée**
Baseline 1	M_1^{B1}	M_2^{B1}	M_c^{B1}
Baseline 2	M_1^{B2}	M_2^{B2}	M_c^{B2}
Variation 1	M_1^{V1}	M_2^{V1}	M_c^{V1}
Variation 2	M_1^{V2}	M_2^{V2}	M_c^{V2}
Upper bound	M_1^{U}	M_2^{U}	M_c^{U}

Test de significativité

- considérons les performances (ROUGE-2) de deux systèmes de résumé automatique
  - système 1 : 40% vs. système 2 : 43%

Test de significativité

- considérons les performances (ROUGE-2) de deux systèmes de résumé automatique
  - système 1 : 40% vs. système 2 : 43%

	A	B	C	D	E	Avg.
système 1	41	34	31	35	59	40
système 2	38	29	27	65	56	43

- toujours effectuer un test de significativté
  - e.g. t.test de Student

Plan

Lectures

- corpus parallèle

  [Tie11] Tiedemann, Bitext alignment

- corpus comparable

  [Rap99] Rapp, Automatic identification of word translations from unrelated english and german corpora

- évaluation en TAL

  [RL10] Resnik & Lin, Evaluation of nlp systems

- accord inter-annotateurs

  [AP08] Artstein & Poesio, Inter-coder agreement for computational linguistics

Plan

References I

Ron Artstein and Massimo Poesio.
Inter-coder agreement for computational linguistics.
Computational Linguistics, 34(4) :555–596, 2008.

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

Philip Resnik and Jimmy Lin.
Evaluation of nlp systems.
The handbook of computational linguistics and natural language processing, 57 :271–295, 2010.

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.

References III


Jörg Tiedemann.
Bitext alignment.
Synthesis Lectures on Human Language Technologies, 4(2) :1–165, 2011.