

Méthodes Empiriques et Langages de Script

Paola Merlo

University of Geneva

année académique 2008-2009

Objectifs du cours

- Apprentissage de UNIX/LINUX et du langage de programmation Perl
- Introduction à l'utilisation d'un corpus
- Introduction aux méthodes d'apprentissage automatique et statistique en TALN

Evaluation

- ATTESTATION : obtentions d'une note suffisante au TP
- EXAMEN ÉCRIT :
attestation requise avant l'examen pour se présenter à l'examen
théorique et pratique : il faut 4 dans les deux parties pour passer

Cours

- ENSEIGNANTS :
Cours : Paola Merlo (Réception : sur rendez-vous)
TP : Gabriele Musillo (Réception : voir page web)

- HORAIRES ET SALLES :
Cours : Mercredi 12h-14h, L 208
TP : Mercredi 14h-16h, B315
Page web du cours :

Admission

- Sont admis au cours les étudiants de BA en Lettres et BA/MA en Sciences.
- Ceci n'est pas un cours d'introduction à la programmation. Si vous n'avez pas de bonnes bases en programmation structurée, vous ne réussirez pas à suivre ce cours.
- Les auditeurs doivent obtenir ma permission pour assister au cours.

Le Plagiat

- Le plagiat – la copie du travail autrui sans citation des sources – est interdit et passible de sanctions.
Le plagiat aux TPs entraîne la note 0 pour tous les TPs.
- Ceci s'applique tant aux textes écrits que aux programmes
- Le plagiat à l'examen entraîne la note 0 et l'annulation de la session d'examen.

Références

Perl

- L. Wall et R. Schwartz, *Programming Perl*, O'Reilly Associates
- E. Quigley, *Perl by example*, Prentice Hall
- J. Friedl, *Mastering Regular Expression*, O'Reilly Associates

Approche Corpus

- B. Habert, A. Nazarenko, et A. Salem, *Les linguistiques de corpus*, Armand Colin
- T. Mc Enery et A. Wilson, *Corpus Linguistics*, Edinburgh Press

Méthodes empiriques et statistiques

- C. Manning et H. Schuetze, *Foundations of Statistical Natural Language Processing*, MIT Press
- D. Jurafsky et J. Martin, *Speech and Language Processing*, Prentice Hall

Introduction à la linguistique

Jacques Moescheler, Antoine Auchlin, *Introduction à la linguistique Contemporaine*, Colin, 2005
Victoria Fromkin editor *Linguistics : an Introduction to Linguistic Theory*, Blackwell, Part 2, chapter 3.

Références sur le web

Cours Perl

<http://www.med.univ-rennes1.fr/poulique/cours/perl/>, Cours sans exercices, avec quelques exemples. Assez clair.

<http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/perl/index.htm>, avec des exercices. Niveau très basique.

<http://www.ftls.org/fr/initiation/perl/>, avec quelques exercices et des exemples. Assez clair.

Documentation Perl

<http://www.enstimac.fr/Perl/>, en français

<http://Perldoc.com>, référence officielle en anglais

Cours Perl et Unix

<http://www.esil.univ-mrs.fr/dgaut/Cours/sommaire-unixperl.html>

Cours avancé Unix

http://www.iie.cnam.fr/Berthelot/Tfse/unix_trsp/unix_trsp1.html

Programme

- INTRODUCTION : le TALN, les approches basées sur les corpus, quelques exemples des problèmes et solutions avec ces approches.
- MÉTHODES QUANTITATIVES : données qualitatives et quantitatives. La notion de distribution de fréquence, la distinction entre type et token, la loi de Zipf, les n-grammes.
- DONNÉES TEXTUELLES : qu'est-ce qu'un corpus, corpus balisé, exemples de corpus textuels : la Penn Treebank, le British National Corpus, le NEGR@ corpus et autres.

Programme

- PERL : un langage de programmation pour les données textuelles :
Les bases : variables, structures de données, gestion du contrôle.
Les expressions rationnelles (régulières) : théorie et pratique.
Les tableaux associatifs (*Hash Tables*).
- LES DONNÉES TEXTUELLES COMME BASE POUR L'ÉVALUATION : les mesures de précision, de rappel et d'exactitude.

Programme

- INTRODUCTION À LA THÉORIE DES PROBABILITÉS :
probabilité simple et probabilité conditionnelle ; le théorème de Bayes ; variables aléatoires et fonctions de probabilité.
- INTRODUCTION À L'APPRENTISSAGE AUTOMATIQUE : le concept d'apprentissage supervisé
- MODÈLE PROBABILISTE POUR L'APPRENTISSAGE AUTOMATIQUE –APPLICATIONS AU TALN :
 - classification de textes
 - l'apprentissage automatique des rôles thématiques

le TALN

Qu'est-ce que l'analyse du langage naturel ?

L'analyse du langage naturel tente de donner à un ordinateur la faculté de comprendre des langues naturelles comme l'anglais, le français ou le japonais.

Par « comprendre », nous ne voulons pas faire croire que l'ordinateur acquiert un mode de pensée, des sensations et des connaissances humaines. Nous voulons seulement dire que l'ordinateur peut reconnaître et utiliser des informations exprimées à l'aide d'une langue naturelle.

Applications du TAL

- L'anglais comme langage de commande – c'est-à-dire l'usage d'une langue naturelle en lieu et place d'un langage artificiel comme c'est encore le cas dans les langages de commande des ordinateurs.
- Les banques de données et les environnements d'aide peuvent accepter des requêtes en anglais.
- La traduction assistée par ordinateur de documents scientifiques et techniques ou bien d'informations commerciales d'une langue naturelle vers une autre.
- La génération automatique de banques de données à partir de documents techniques, tels que des rapports de pannes ou des rapports médicaux.

Applications du TAL

- Aide à la Rédaction
 - correction des textes
 - génération de textes
- Recherche documentaire
- Filtrage/classification d'information
- Résumé automatique, pour un seul document, pour plusieurs documents sur le même sujet

Les propriétés des applications en TAL

Les contextes d'application imposent plusieurs contraintes :

TRAITEMENT RAPIDE : Nécessitent des algorithmes de complexité polynomiale (et non exponentielle).

BONNE COUVERTURE DE LA LANGUE UTILISÉE Nécessitent des ressources linguistiques représentatives en quantité suffisante.

L'approche à base de corpus

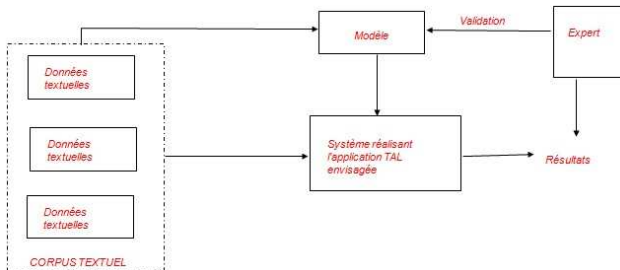
Les ressources linguistiques représentatives en quantité suffisante sont très difficiles et chères à construire. On ne cherche plus à reproduire la compétence à l'aide de modèles formalisant notre compréhension du langage mais à reproduire, pour une classe d'applications TAL donnée,

la performance linguistique associée

et ce, à l'aide de modèles automatiquement extraits de volumes importants de données textuelles caractéristiques de la classe d'application envisagée.

L'approche à base de corpus

La validation des modèles obtenus n'est pas liée à leur capacité explicative du fonctionnement de la langue mais repose sur l'évaluation de l'amélioration des performances que permettent ces modèles pour l'application TALN envisagée



Exemple : L'attachement au PP

Je mange la pizza avec le couteau

Je mange la pizza avec le fromage

Avant : modélisation des connaissances linguistiques et extra-linguistiques nécessaires à enlever l'ambiguïté.

Par exemple, sémantique du verbe et du syntagme prépositionnel (SP) :

verbe d'action ou verbe d'état ?

SP instrumental ou de manière ? ou spécification ?

Connaissance du monde : est-ce qu'on mange des couteaux et on mange avec du fromage ?

Exemple : L'attachement au PP

Je mange la pizza avec le couteau
Je mange la pizza avec le fromage

Méthode à l'aide de corpus.

P(mange, avec, fromage) vs (soupe, avec, fromage)

P(mange, avec, couteau) vs (soupe, avec, couteau)

Avantages

- Acquisition : identification et encodage automatique des connaissances nécessaires.
- Couverture : on couvre automatiquement tous les phénomènes linguistiques dans le domaine d'application donné.
- Robustesse : on s'adapte facilement au bruit et aux données imprévues.
- Portabilité : en principe, assez facile à étendre vers une nouvelle langue.
- Évaluation : on arrive à évaluer de façon expérimentale des systèmes pratiques et des hypothèses scientifiques.

Les Méthodes Quantitatives

- Données qualitatives et quantitatives
- La notion de distribution de fréquence
- La distinction entre type et token
- Les n-grammes

Les Données Qualitatives

Exemple : le jugement de grammaticalité des phrases

- Je mange la pizza avec le fromage.
- Je mange la pizza au fromage.

Pas de fréquences, toutes les données ont la même importance.

- Exemples se trouvent dans un corpus ou sont obtenus de façon naturelle.
- Beaucoup d'attention aux détails
- Les conclusions tirées sur la base d'un échantillon qualitatif ne s'appliquent pas à toute la population avec certitude, car on ne recherche pas des exemples représentatifs de la population.

Les Données Quantitatives

Exemple : le comptage des verbes dans un corpus français.

Les données n'ont pas toutes la même importance.

- Les données sont classées, comptées, résumées avec des statistiques.
- Les données à basse fréquence sont souvent considérées comme moins importantes (mais sont-elles moins nombreuses ?).
- Les données sont des échantillons, donc les généralisations s'appliquent à toute la population avec un certain degré de certitude.

Tokens, types et distributions

Pour classer des occurrences (les tokens), il faut d'abord établir un schéma, qu'on appelle une classification (les types). Une fois la classification établie, on peut classer chaque occurrence selon un type. Chaque type aura alors un certain nombre d'effectifs. L'ensemble des comptages d'effectifs de la classification s'appelle une distribution.

Exemple 1 Si les vocables du langage sont les types de données, et les occurrences des mots, les tokens, alors il s'agit d'une distribution des fréquences des mots. Par exemple, la phrase

La fille a vu son père, mais le père n'a pas vu la fille.

aura la distribution suivante :

type	a	fille	la	le	mais	n'	pas	père	son	vu
token	2	2	2	1	1	1	1	2	1	1

Tokens, types et distributions(suite)

Exemple 2 Soit une classification dont les types sont les étiquettes des parties du discours (partsofspeech tags ou POS tags en anglais). Les tokens sont les mots dans un texte. Alors, il s'agit d'une distribution d'étiquettes.

Par exemple, voici la distribution des 15 étiquettes les plus fréquentes dans le corpus Brown, étiqueté avec les étiquettes du Penn Treebank :

1.	161397	NN	6.	58262	,	11.	46684	VBD
2.	136714	IN	7.	55912	NNS	12.	38097	CC
3.	116454	DT	8.	55645	.	13.	36887	VB
4.	76586	JJ	9.	52037	RB	14.	29435	VBN
5.	62020	NNP	10.	47303	PRP	15.	26135	TO

Tokens, types et distributions(suite)

Exemple 3 Si les mots et les signes de ponctuation sont les types de la classification, et leurs occurrences les tokens de la classification, alors il s'agit d'une distribution de lexèmes. Ou un dictionnaire des fréquences.

Par exemple, voici un extrait de la distribution des mots et signes de ponctuation dans le corpus Brown :

1.	69836	the	7.	23157	a	66.	1961	said
2.	58260	,	8.	21314	in	70.	1815	about
3.	49249	.	9.	10777	that	80.	1600	time
4.	36365	of	10.	10182	is	89.	1332	man
5.	28826	and	11.	9968	was	93.	1292	like
6.	26126	to	12.	9801	he	99.	1125	made

Morale

finalement...

FOXTROT



La Loi de Zipf

Si l'on dresse une table de l'ensemble des mots différents d'un texte quelconque, classés par ordre de fréquences décroissantes, on constate que la fréquence d'un mot est inversement proportionnelle à son rang dans la liste, ou, autrement dit, que le produit de la fréquence de n'importe quel mot par son rang est constant, ce que traduit la formule

$$f * r = C, \text{ où } f \text{ est la fréquence et } r \text{ le rang.}$$

La loi de Zipf stipule donc que la fréquence du second mot le plus fréquent est la moitié de celle du premier, la fréquence du troisième mot le plus fréquent, son tiers, etc. Cette égalité, qui n'est vraie qu'en approximation, est indépendante des locuteurs, des types de textes et des langues -> un des nombreux traits généraux des énoncés linguistiques

Extrait de la section Linguistique et Statistique de l'Encyclopaedia Universalis version 3.0 sur CD-ROM.

Exemple

Pour le deuxième paragraphe du texte précédant, on a la distribution suivante. On indique le rang, la fréquence et le mot.

1 - 8 ,	16 - 1 général	16 - 1 donc	16 - 1 moitié
2 - 4 de	16 - 1 isolée	16 - 1 linguistiques	16 - 1 pas
2 - 4 des	16 - 1 langues	16 - 1 tiers	16 - 1 premier
2 - 4 .	16 - 1 indépendante	16 - 1 toute	16 - 1 locuteurs
2 - 4 la	16 - 1 Il	16 - 1 trait	16 - 1 loi
6 - 3 n'est	16 - 1 d'autres	16 - 1 stipule	16 - 1 mais
6 - 3 du	16 - 1 d'un	16 - 1 série	16 - 1 première
8 - 2 fréquent	16 - 1 constatation	16 - 1 textes	16 - 1 s'agisse
8 - 2 mot	16 - 1 approximation	16 - 1 troisième	16 - 1 second
8 - 2 plus	16 - 1 celle	16 - 1 véritablement	16 - 1 semble
8 - 2 est	16 - 1 ainsi	16 - 1 égalité	16 - 1 qu'en
8 - 2 fréquence	16 - 1 et	16 - 1 énoncés	16 - 1 qu'il
8 - 2 le	16 - 1 etc	16 - 1 une	16 - 1 qui
8 - 2 que	16 - 1 La	16 - 1 vraie	
8 - 2 Cette	16 - 1 Zipf	16 - 1 son	

Exemple (suite)

Remarques :

- il y a 88 mots au total.
- le mot le plus fréquent apparaît 8 fois, le deuxième 4 fois,
- il y a 42 mots qu'on trouve une seule fois (hapax legomena)

URL : <http://users.info.unicaen.fr/giguette/java/zipf.html>

Les n-grammes

De la même façon que nous sommes intéressés aux distributions des fréquences des mots individuels, nous sommes aussi, même plus, intéressés à récolter les distributions des fréquences des séquences à deux, trois, quatre mots à la fois.

Les n-grammes

Un n-gramme (néologisme à partir des termes « bigramme », « trigramme », etc.) est une séquence de taille fixée d'un texte.

Exemples

Les n-grammes des mots sont toutes les séquences de n mots dans le corpus.

le chat mange la souris

- bigrammes (n-grammes de longueur 2)
(le chat) (chat mange) (mange la) (la souris)
- trigrammes (n-grammes de longueur 3)
(le chat mange) (chat mange la) (mange la souris)

Identificateur de langues

Arrive-t-on à identifier une langue avec seulement les fréquences de n-grammes ?

Essayons !

La distribution des bigrammes

Voici pour trois langues inconnues, les fréquences d'apparition des 10 bigrammes les plus fréquents :

ES	DE	LE	EN	RE	NT	ON	ER	TE	EL
3318	2409	2366	2121	1885	1694	1646	1514	1484	1382
TH	HE	IN	ER	AN	RE	ES	ON	ST	NT
3020	2496	2078	1821	1676	1467	1345	1318	1290	1267
EN	ER	CH	DE	TE	ND	EI	IE	IN	GE
3956	3818	2647	2386	2167	1990	1935	1702	1579	1521

La distribution des trigrammes

Voici pour trois langues inconnues, les fréquences d'apparition des 10 trigrammes les plus fréquents :

ENT	LES	EDE	DES	QUE	AIT	LLE	SDE	ION	EME
900	801	630	609	607	542	509	508	477	472

THE	AND	ING	ENT	ION	NTH	TER	INT	OFT	THA
2069	819	607	487	428	381	367	357	355	355

DER	ICH	EIN	NDE	SCH	DIE	TEN	END	CHE	UND
1025	959	939	812	812	804	662	611	607	586

Les n-grammes

À quoi servent les n-grammes ?

À travers les distributions des n-grammes on arrive à approximer certaines régularités langagières. Par exemple,

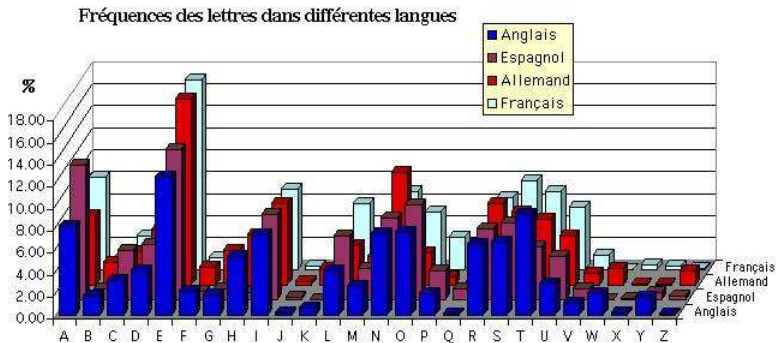
- voisins d'un mot dans un texte
- classification distributionnelle (syntaxique) des mots
- représentation d'un document
- représentation d'une langue

Les distributions de n-grammes

Les systèmes de chiffrement simples sont facilement cassable par une méthode d'analyse de fréquences des lettres, car pour chaque langue, certaines lettres sont beaucoup plus utilisées que d'autres.

Les distributions d'unigrammes

Voici pour le français, l'anglais, l'allemand et l'espagnol les fréquences d'apparition des lettres.



Les n-grammes

Les unités de comptage ne sont pas nécessairement les mots.

- Les spécialistes de parole s'occupent de n-grammes de phonèmes.
- En syntaxe, les bigrammes de catégorie morpho-syntaxique sont des couples du type (Nom-Verbe) ou (Adjectif-Nom), parmi d'autre, indiquant combien de fois un nom est suivi d'un verbe dans le corpus, ou un adjectif suivi d'un nom.
- En cryptographie on s'occupe d'unigrammes, de bigrammes et de trigrammes de lettres.

Résumé

- Données qualitatives vs quantitatives
 - quantitatives : attention au détail, même importance
 - qualitative : résumé, échantillonnage, représentativité
- Classification : on établit un schéma (les types), et on classe les instances (tokens).
- Distribution des tokens par type
- Loi de Zipf : le produit entre le rang et la fréquence est constant
- Les n-grammes en tant qu'unités linguistiques, les distributions de n-grammes
- Fréquence relative pour normaliser échantillons de tailles différentes

Résumé

- Le TALN tente de donner à un ordinateur la faculté de « comprendre » les langues naturelles (anglais, français, etc.)
- Ses applications sont l'interfaçage avec les grandes bases des données, la traduction automatique ou assistée, la génération automatique des documents, la recherche et le filtrage documentaire, le résumé automatique
- Propriétés des LN : ambiguës et implicites
- Approche corpus :
 - ressources linguistiques en grande quantité
 - acquisition automatique de connaissances langagières
 - accent sur la performance et l'évaluation systématique