Centre de Recherche
en Ingénierie
Multilingue

Mémoire de Master 2 professionnel
Ingénierie linguistique
Traductique et gestion de l'information

Analyse textométrique et sémantique de textes numériques japonais pour la constitution de corpus de textes pro et anti-tabac

Valérie COLLEC CLERC
19 octobre 2010

Sommaire

- Objectifs
- Démarche
- Langue japonaise et TAL
- Segmentation
- Analyse lexicométrique
- Analyse sémantique
- Catégorisation
- Evaluation
- Bilan
- Questions

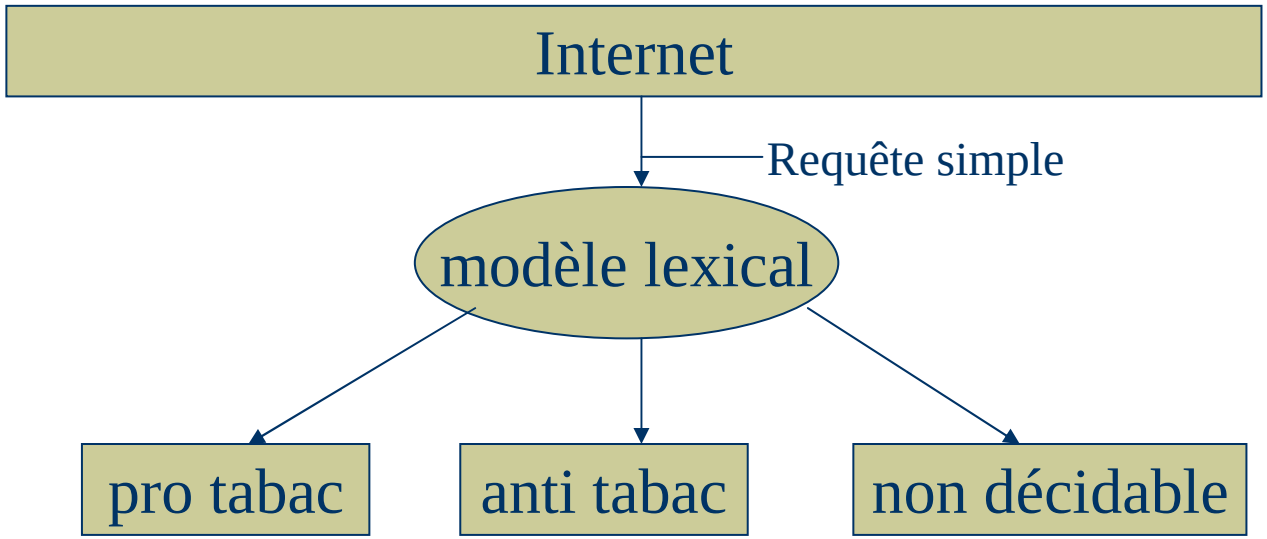
Objectif : Analyse sémantique et son application

Notre objectif: analyse sémantique de textes pro et anti tabac en langue japonaise

=> Validation par un outil prototype de catégorisation

.

Prototype de catégorisation



Principe: repérer des opinions par la présence d'éléments lexicaux caractéristiques

Priorité: peu de textes mais bien identifiés

Démarche

Constitution d'un corpus expérimental de textes représentatifs

Détermination d'éléments lexicaux caractéristiques par une analyse lexicométrique

Détermination de classes lexicales par une analyse sémantique des contextes associés

Postulat: une approche purement statistique de l'analyse de corpus est insuffisante pour catégoriser des domaines sociolinguistiques complexes.

Contenu des travaux

Montrer l'intérêt de l'approche sur un modèle restreint mais significatif:

- une trentaine de textes pour le corpus expérimental
- Une vingtaine d'éléments lexicaux pour la catégorisation.

Mettre en évidence les problèmes spécifiques liés à la langue japonaise.

Langue japonaise et TAL
Difficultés de l'analyse automatique

Confusion liée aux **codages multiples** non compatibles (JIS, EUC-JP, Shift-JIS, Unicode, codage occidental…)
文字化け

Ambigüité lexicale: **absence de séparateurs** entre morphèmes dans les phrases

Ambigüité sémantique: **complexité linguistique** du japonais (énoncés contextuels, hétéronymes, ..)
私は本です /je/thème/livre/est/ / je suis un livre/??

Langue japonaise et TAL
Solutions possibles

Confusion liée aux **codages multiples non compatibles**

➔*Codage pivot UTF8*

➔*Prétraitements/post-traitements entre les outils utilisés (Lexico3, Chasen, KH-Coder, …)*

Ambigüité lexicale :**absence de séparateurs**

➔*Outils de segmentation (Chasen, MeCab,…)*

Ambigüité sémantique: **complexité linguistique**

➔*Analyse sémantique*

Problèmes de segmentation

室内を含め，喫煙可能な場所でも灰皿がなければ喫煙しない。その場合にも決して「置き煙草」はしない。

Segmentation attendue:
室内 | を | 含め | ， | 喫煙可能 | な | 場所 | で | も | 灰皿 | が | なければ | 喫煙しない | 。 | その | 場合 | に | も | 決して | 「 | 置き煙草 | 」 | は | しない | 。

Segmentation obtenue par Chasen:
室内 | を | 含め | ， | 喫煙 | 可能 | な | 場所 | で | も | 灰皿 | がなけれ | ば | 喫煙 | し | ない | 。 | その | 場合 | に | も | 決して | 「 | 置き | 煙草 | 」 | は | し | ない | 。

Segmentation obtenue par MeCab:
室 | 内 | を | 含め | ， | 喫煙 | 可能な | 場所 | でも | 灰皿 | が | なければ | 喫煙 | し | ない | 。 | その | 場合 | に | も | 決して | 「 | 置き | 煙草 | 」 | は | し | ない | 。

室 | 内 sur-segmentation (souvent récupérable, si mot composé)
でも　sous-segmentation (généralement irrécupérable, confusion)

Evaluation des outils de segmentation

Comparaison entre Chasen le plus connu et Mecab le plus récent, sur un texte représentatif et analyse des divergences

Taux de rappel (éléments lexicaux) = 75%/74%
Taux de précision (éléments lexicaux) = 61%/63%

Taux de rappel des coupures = sous-segmentation) = 97%/97%
Taux de précision des coupures= sur-segmentation) = 75%/77%

Les deux logiciels ont tendance à sur-segmenter et les sous-segmentations sont rares. Une analyse complémentaire de la gravité des segmentations divergentes entre les deux n'a pas permis pas de les séparer.

Constitution du corpus expérimental
Principes

Base: un ensemble de sites fourni par une personne de langue maternelle japonaise.

Compléments par des requêtes (Google Japan, et Yahoo.jp)
mots clés de domaines basés sur notre expérience en français:
Tabac たばこ,
Aimer/Détester couples antonymiques (嫌煙, 愛煙) (禁煙, 喫煙).

Sélection des textes selon leur contenu (posture d'énonciation et acteurs).

Constitution du corpus expérimental
Textes pro-tabac

➔ Difficulté: trouver des textes clairement pro-tabac ou pro-fumeur sur les sites japonais

=> Les textes anti-fumeurs sont nombreux et expriment des opinions de degrés variables (envie d'arrêter de fumer, gêné par le tabac, détestation des fumeurs),

=> Les textes anti-fumeurs sont plus rares et ne sont pas ouvertement militants. Ils s'expriment par des biais détournés.

➔ Conséquence: axer notre démarche sur la posture pro-fumeur.

Analyse lexicométrique
Fonctions utilisées des logiciels

Lexico 3 : non spécifique au japonais,
- nécessite segmentation et prétraitement des textes

Permet l'analyse de **plusieurs textes** (balises de séparation)

➔ **Analyse de spécificités** (textes pro fumeur / anti fumeurs)

KH-Coder : spécifique au japonais
– analyse morphosyntaxique intégrée (Chasen)

Permet l'analyse des segments **d'un seul texte** => utilisation d'une concaténation des textes pro-tabac et anti-tabac

➔**Occurrences des segments par partie du discours**

➔**Segments répétés** (cohérents)

➔**Recherche de termes (**avec **TermExtract)**

Analyse sémantique
Exemples

Une analyse sémantique analyse est menée pour les différents des contextes des éléments lexicaux retenus à partir de l'analyse lexicométrique

Exemple: Recherche de contextes de 煙草 dans le corpus pro-tabac

Partie : AviProFum1508_08__Kachijiten, Nombre de contextes : 5
と 思い ます . 私 は 喫煙 者 です が , **煙草** を 吸わ ない 友人 に は 許可 を 得る , または
一時期 は ノン ニコチン , ノン タール の **煙草** (500 円) を 吸っ て い ました が ,好み
め て しまい ました . でも 何故か 今 の **煙草** の ほう が 非 喫煙 者 に 好か れ て い ます
ます . 知ら ない お 兄さん に まで ,『 **煙草** は 嫌い です が それ は いい かも 』と
オ . 友人 の 優し さ に 感謝 を し つつ , **煙草** を 吸っ て いる 日々 です . コンビニ で

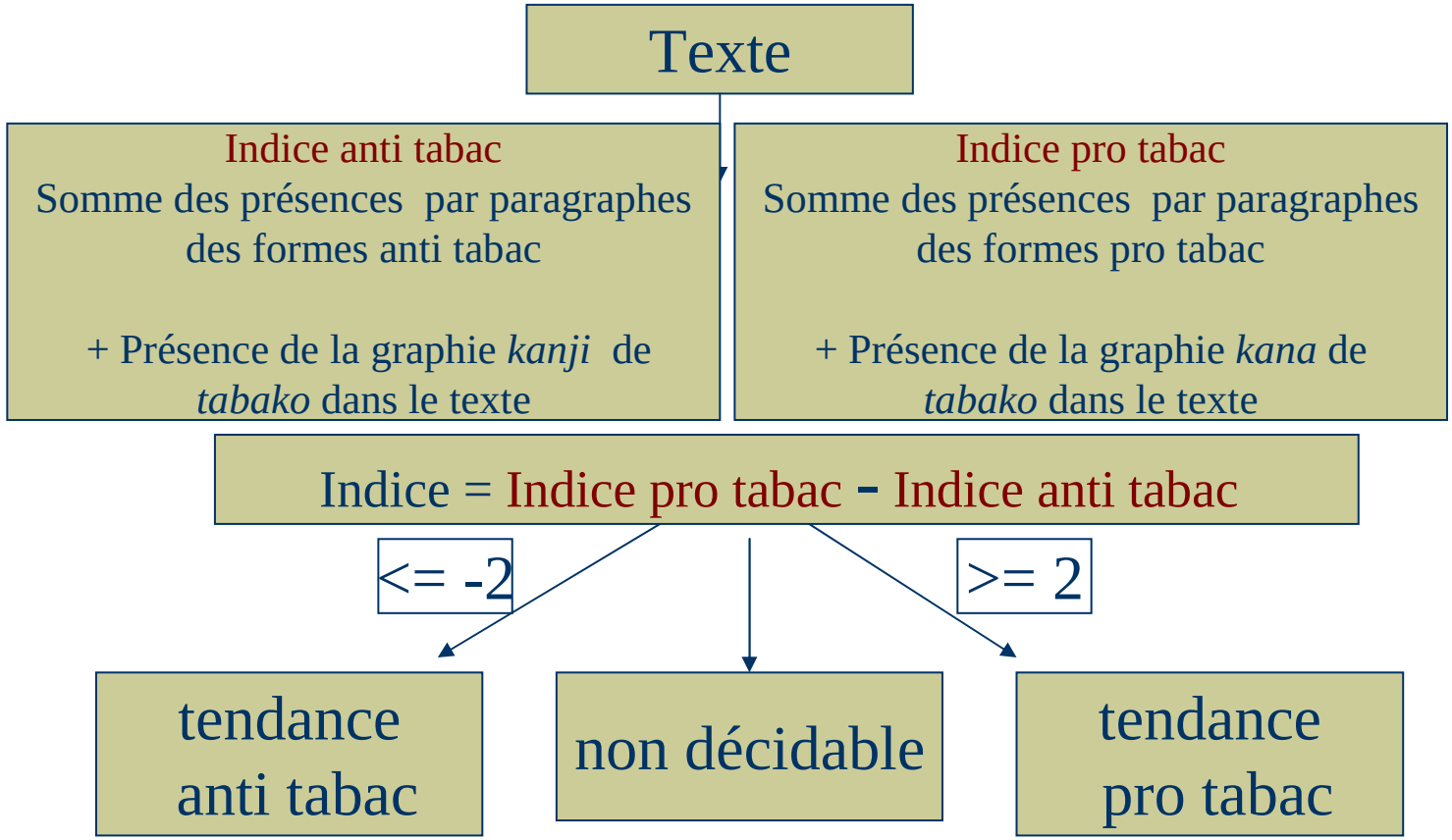
Catégorisation
Classes sémantiques retenues

Anti-tabac	Pro-tabac
nuisance	responsabilité
victimes	mélioratif
substances chimiques	

graphie: *kana* たばこ, タバコ

graphie *kanji* :煙草

Catégorisation
Algorithme



Evaluation
Corpus expérimental

Rappel AviProFum = Nombre de décisions pro tabac correctes / Nombre de textes du corpus AviProFum

Précision AviProFum = Nombre de décisions pro tabac correctes / Nombre de décisions prises sur le corpus AviProFum

Corpus AviAntiFum	Rappel =20%	Précision=80%
Corpus AviProFum	Rappel =67%	Précision=86%
Corpus total	Rappel = 39%	Précision=87%

L'évaluation sur le corpus expérimental donne une indication du comportement de l'outil de catégorisation:

⇒ Priorité donnée aux textes pro-tabac

⇒ Priorité donnée à la précision sur le rappel

Evaluation
Blog japonais

Evaluation sur Ameba

Aspiration de 100 textes du site, sélectionnés sur le mot clé たばこ avec extraction de la contribution principale.

Décision sur 30% des textes. Ces textes ont fait l'objet d'une annotation manuelle => 10 sont ambigus (mélanges d'opinions) et sont éliminés.

Décisions pro-tabac :	Précision = 78 %
Décisions anti-tabac :	Précision = 67 %
Total des décisions :	Précision = 71%

Bilan

- Analyse sémantique dans le cadre de textes japonais
- Construction d’un prototype simple
- La catégorisation permet une présélection des textes pro-tabac.

Questions/Réponses

Merci de votre attention …

Si vous avez des questions …