Quel modèle pour détecter une opinion ?

Rodrigo Acuna-Agost & Eric Charton
DEFT'07: DEfi Fouille de Texte 2007

Séminaire du LIA, 14 Juin 2007

Plan

Plan

Plan

Plan

Introduction

Deft : défi fouille de texte en langue française

Deft et le LIA

- Deft'07 : le troisième défi
- Deux participations du LIA à DEFT ...

Deft'05 *Segmentation de discours*

- Une équipe senior et une équipe junior
- L'équipe senior remporte Deft
- L'équipe junior classée première pour la tâche 1

Deft'07 *Recherche d'opinion*

- Une équipe senior et une équipe junior

Description de la tâche

La tâche proposée lors de DEFT'07 a consisté à attribuer automatiquement une classe d'opinion à des textes - critiques, commentaires ou interventions - regroupés dans 4 corpus

Description des corpus

Corpus

Corpus 1

Critiques de films, livres, spectacles et bandes dessinées

Corpus 2

Tests de jeux vidéo

Corpus 3

Relectures d'articles de Conférences

Corpus 4

Débats parlementaires

FIG.: Description des corpus DEFT'07

Description des corpus

Corpus

Corpus 1 – 3000 textes

0: mauvais
1: moyen
2: bien

Corpus 2 – 4000 textes

0: mauvais
1: moyen
2: bien

Corpus 3 – 1000 textes

0: rejet de l'article
1: acceptation (modifications majeures)
2: acceptation (modifications mineures)

Corpus 4 – 28000 textes

0: contre
1: pour

FIG.: Description des corpus DEFT'07

Trois méthodes pour détecter une opinion dans un corpus

Recherche par similarité cosine

Principe

- Méthode de base de la recherche documentaire
- Projeter un document à retrouver et sa requête dans un espace vectoriel en utilisant le poids des mots qu'ils contiennent
- Poids des mots obtenus par TF.IDF
- Mesurer le cosinus de l'angle entre le vecteur d'une requête et celui des documents à retrouver
- $cosine(\vec{R}, \vec{D}) = \frac{\vec{R}.\vec{D}}{\|\vec{R}\|.\|\vec{D}\|}$
- Produire une liste de documents pertinents classés par score de similarité cosine

Recherche par similarité cosine

Adaptation du modèle à la recherche d'opinion

- Le modèle du moteur de recherche *requête-document* est inversé
- La requête devient une classe k et contient des mots ou suites de mots caractéristiques de l'opinion
- Les k classes d'opinions sont construites d'après le corpus d'entrainement
- Mesure du cosinus de l'angle entre le vecteur du poids des mots du document à classer et celui des k classes
- On cherche à maximiser le caractère descriptif d'une opinion, en paramétrisant la sélection des mots contenus dans les classes

Recherche par similarité cosine

Similarité cosine : paramètres de construction des classes

- Adoption de n_grammes (bon_article, mauvais_article, film_excellent, etc ...)
- Suppression des mots outils non porteurs d'opinion (le, les, des, un, ceci, etc ...)
- Réduction des mots à leurs lemmes (bon, bonne ...)
- Suppression des noms propres (détection par les majuscules)
- Suppression des mots communs à plus d'une classe
- Conservation de portions spécifiques du texte susceptibles d'être porteuses d'opinion

Recherche par similarité cosine

Optimisation du classifieur

- En s'appuyant sur le corpus d'entrainement, recherche des paramètres qui maximisent le F Score, pour chacune des classes
- $\vec{A}(k) = Argmax_{A_i}\left(cosine(\vec{D}, \vec{A_i})\right)$

Recherche par similarité cosine

Résultats

Corpus	Lem	Adico	$\cup = 0$	npr	pct	ngm	Train	DEFT07
Avoir..	oui	oui	non	oui	20/30	3	0.50	0.37 (0.48)
Jeux..	oui	oui	oui	non	20/30	3	0.73	0.62 (0.65)
Relect..	oui	oui	non	oui	50/50	3	0.37	0.43 (0.46)
Débats	non	non	oui	non	20/30	3	0.63	0.61 (0.64)

TAB.: Résultats par corpus et options utilisées pour les obtenir (lemmatisation, antidictionnaire, intersections de classes, noms propres, pourcentages de textes retenus en tête et fin, n_grammes).

Régression logistique

Recherche par régression logistique

Hypothèses

- La classification de chaque texte dépend des mots qui le composent
- Il existe une série de *mots critiques* qui prédominent dans chacune des catégories
- La probabilité qu'un texte appartienne à une certaine catégorie dépend du nombre de *mots critiques* présents
- Il est possible d'estimer cette probabilité en utilisant un modèle de régression logistique

Recherche par régression logistique

Hypothèses

- La classification de chaque texte dépend des mots qui le composent
- Il existe une série de *mots critiques* qui prédominent dans chacune des catégories
- La probabilité qu'un texte appartienne à une certaine catégorie dépend du nombre de *mots critiques* présents
- Il est possible d'estimer cette probabilité en utilisant un modèle de régression logistique

Recherche par régression logistique

Hypothèses

- La classification de chaque texte dépend des mots qui le composent
- Il existe une série de *mots critiques* qui prédominent dans chacune des catégories
- La probabilité qu'un texte appartienne à une certaine catégorie dépend du nombre de *mots critiques* présents
- Il est possible d'estimer cette probabilité en utilisant un modèle de régression logistique

Recherche par régression logistique

Hypothèses

- La classification de chaque texte dépend des mots qui le composent
- Il existe une série de *mots critiques* qui prédominent dans chacune des catégories
- La probabilité qu'un texte appartienne à une certaine catégorie dépend du nombre de *mots critiques* présents
- Il est possible d'estimer cette probabilité en utilisant un modèle de régression logistique

Recherche par régression logistique

Ensembles et des index

i : Index des documents
j : Index des classes
C : Ensemble de catégories.
T : Ensemble de documents d'apprentissage

Variable de Dépendance

θ_{ij} : estime la probabilité que le document i soit apparenté à la classe j

Variables Explicatives

z_{ij} : Le nombre de mots critiques dans le texte i de la catégorie j
y_i : Le nombre total de mots dans le texte i

Recherche par régression logistique

Ensembles et des index

i	:	Index des documents
j	:	Index des classes
C	:	Ensemble de catégories.
T	:	Ensemble de documents d'apprentissage

Variable de Dépendance

θ_{ij} : estime la probabilité que le document i soit apparenté à la classe j

Variables Explicatives

z_{ij} : Le nombre de mots critiques dans le texte i de la catégorie j
y_i : Le nombre total de mots dans le texte i

Recherche par régression logistique

Ensembles et des index

i	:	Index des documents
j	:	Index des classes
C	:	Ensemble de catégories.
T	:	Ensemble de documents d'apprentissage

Variable de Dépendance

θ_{ij} : estime la probabilité que le document i soit apparenté à la classe j

Variables Explicatives

z_{ij} : Le nombre de mots critiques dans le texte i de la catégorie j
y_i : Le nombre total de mots dans le texte i

Recherche par régression logistique

Modèle de régression logistique

$$\theta_{ij} = \frac{e^{\left(\alpha_j + \gamma_j y_i + \sum_{k \in C} \beta_j^k z_{ik}\right)}}{1 + e^{\left(\alpha_j + \gamma_j y_i + \sum_{k \in C} \beta_j^k z_{ik}\right)}}$$

Coefficients

α_j : Constante de l'équation
β_j^k : Coefficient des variables de prédiction z_{ik}
γ_j : Coefficient des variables de prédiction y_{ij}

Classification de chaque texte

$$j^* = \arg\{\max_{j \in C} \theta_{ij}\}$$

Recherche par régression logistique

Modèle de régression logistique

$$\theta_{ij} = \frac{e^{\left(\alpha_j + \gamma_j y_i + \sum_{k \in C} \beta_j^k z_{ik}\right)}}{1 + e^{\left(\alpha_j + \gamma_j y_i + \sum_{k \in C} \beta_j^k z_{ik}\right)}}$$

Coefficients

α_j : Constante de l'équation
β_j^k : Coefficient des variables de prédiction z_{ik}
γ_j : Coefficient des variables de prédiction y_{ij}

Classification de chaque texte

$$j^* = \arg\{\max_{j \in C} \theta_{ij}\}$$

Recherche par régression logistique

Modèle de régression logistique

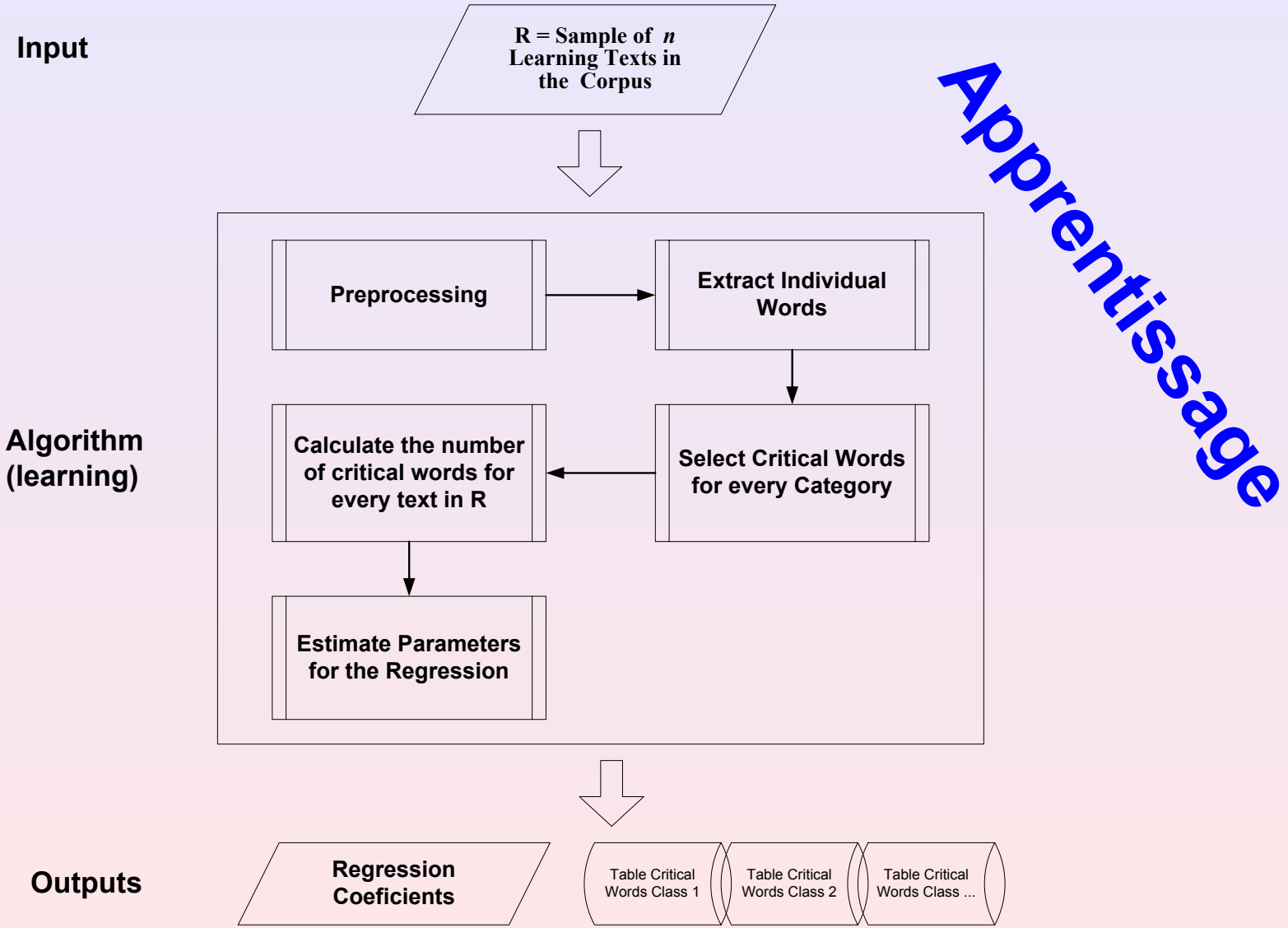
$$\theta_{ij} = \frac{e^{\left(\alpha_j + \gamma_j y_i + \sum_{k \in C} \beta_j^k z_{ik}\right)}}{1 + e^{\left(\alpha_j + \gamma_j y_i + \sum_{k \in C} \beta_j^k z_{ik}\right)}}$$

Coefficients

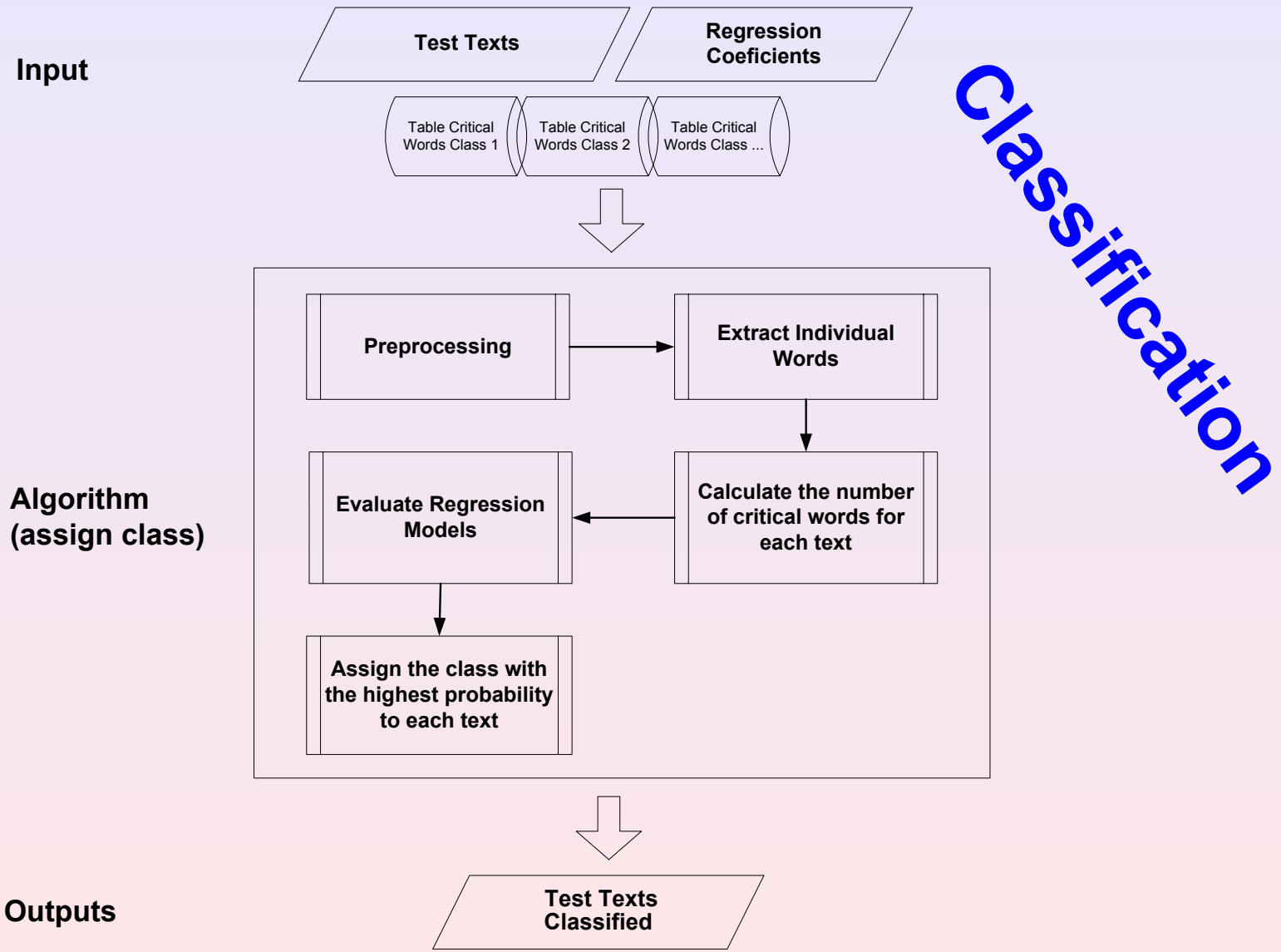
α_j : Constante de l'équation
β_j^k : Coefficient des variables de prédiction z_{ik}
γ_j : Coefficient des variables de prédiction y_{ij}

Classification de chaque texte

$$j^* = \arg\{\max_{j \in C} \theta_{ij}\}$$


Input
R = Sample of n Learning Texts in the Corpus
Apprentissage
Algorithm (learning)
Preprocessing
Extract Individual Words
Calculate the number of critical words for every text in R
Select Critical Words for every Category
Estimate Parameters for the Regression
Outputs
Regression Coeficients
Table Critical Words Class 1
Table Critical Words Class 2
Table Critical Words Class ...

Classification



Résultats Méthode Régression Logistique

Corpus 1 : Critiques de films, livres, spectacles et bandes dessinées

Item	Valeur
$\|T\|$	2074
n	2074
$\|C\|$	3
$F - Score$ Corpus de test (cette méthode)	0.50
$F - Score$ Corpus de test (moyenne DEFT'07)	0.48

Résultats Méthode Régression Logistique

Corpus 2 : Tests de jeux vidéo

Item	Valeur
$\|T\|$	2537
n	60
$\|C\|$	3
$F - Score$ Corpus de test (cette methode)	0.46
$F - Score$ Corpus de test (moyenne DEFT07)	0.65

Résultats Méthode Régression Logistique

Corpus 3 : Relectures d'articles de Conférences

Item	Value
$\|T\|$	881
n	881
$\|C\|$	3
$F - Score$ Corpus de test (cette méthode)	0.47
$F - Score$ Corpus de test (Moyenne DEFT07)	0.47

Résultats Méthode Régression Logistique

Corpus 4 : Débats parlementaires

Item	Valeur
$\|T\|$	17299
n	1000
$\|C\|$	2
$F - Score$ Corpus de test (cette méthode)	0.55
$F - Score$ Corpus de test (moyenne DEFT07)	0.64

Calcul de compacité

Recherche par calcul de compacité

Principe d'un système de question réponse

- La question est nettoyée des mots outils et étiquetée. Exemple : *Qui a gagné le Tour de France* devient *Qui*[*pers*] *gagné Tour_de_France* [*prod*]
- Recherche dans le corpus d'un segment contenant le mot objet de la question pour obtenir une *entité réponse candidate* (*ERC*). Recherche dans l'*ERC* d'une entité nommée cible et calcul de son score de compacité.

Calcul de compacité

$$Compacité(ER_{c_i}) = \frac{\sum_{X \in MQ} P_{X,ER_{c_i}}}{\mid MQ \mid} \quad (1)$$

Calcul de compacité pour une recherche d'opinion

Création des questions

Construction pour chaque classe d'opinion d'une liste de questions possibles d'après les affirmations du corpus d'entrainement. Exemple : Dans la classe d'opinion "bonne", l'affirmation *cet article est de qualité* est étiquetée en *article* [*mot centroïde*] *qualité* [*qualificatif possible*]

Recherche de réponses

- Recherche dans le document à classer de tous les segments avec le mot centroïde (ex :*article*) et des mots proches
- Addition des scores de compacité obtenus pour tous les segments issus du document d'après toutes les questions possibles contenues dans les classes

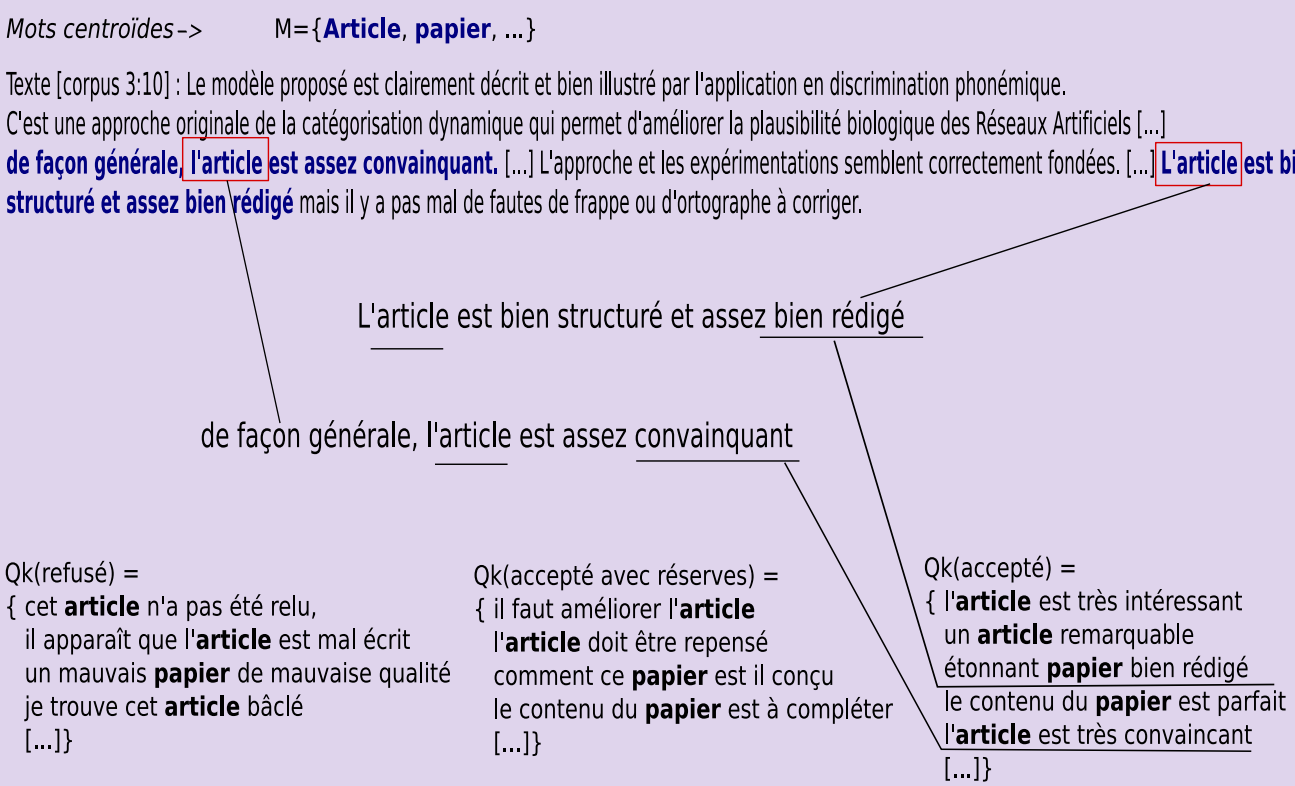
Recherche par calcul de compacité

Localisation des affirmations à transformer en questions

- Identifier dans le corpus d'entrainement les mots susceptibles d'être le centre d'une phrase exprimant l'opinion (centroïdes)
- Méthode : statistiques de mots les plus fréquents, après suppression des mots outils et des noms propres
- Exemple : dans le corpus relectures, le mot "Article" apparaît en première position

Recherche par calcul de compacité

Compacité exemple



Recherche par calcul de compacité

Résultats : corpus débat

Les mots centroïdes de plus fortes occurences : $\{Loi, projet\}$.

Classe	0 (Défavorable)	1 (Favorable)	F-Score
Précision	0.55	0.48	0.51 (0.64)
Rappel	0.89	0.11	

TAB.: Scores obtenus sur le corpus Débats (entre parenthèses, le score moyen obtenu par les participants de DEFT07

Recherche par calcul de compacité

Résultats : corpus relectures

Les mots centroïdes de plus fortes occurences : {*Article*, *Papier*}

Classe	0 (Défavorable)	1 (Moyen)	2 (Favorable)	F-Score
Précision	0.54	0.09	0.66	0.42 (0.46)
Rappel	0.19	0.46	0.57	

Recherche par calcul de compacité

Résultats : corpus A voir A lire

Les mots centroïdes de plus fortes occurences : {*Film*, *Roman*}

Classe	0 (Défavorable)	1 (Moyen)	2 (Favorable)	F-Score
Précision	0.43	0.35	0.49	0.40 (0.48)
Rappel	0.14	0.10	0.86	

Conclusions et perspectives

Conclusions

Conclusions

Ces trois méthodes dont deux originales permettent d'obtenir des résultats proches de la moyenne des participants de Deft'07. De multiples possibilités d'améliorations existent.

Perspectives

Amélioration du modèle de régression

- Ajouter d'autres variables explicatives
- Passer d'une exploration des mots critiques à une étude des phrases critiques
- Introduire une utilisation des modèles n-grammes
- Utiliser une taille d'échantillon appropriée

Perspectives

Amélioration du modèle de régression

- Ajouter d'autres variables explicatives
- Passer d'une exploration des mots critiques à une étude des phrases critiques
- Introduire une utilisation des modèles n-grammes
- Utiliser une taille d'échantillon appropriée

Perspectives

Amélioration du modèle de régression

- Ajouter d'autres variables explicatives
- Passer d'une exploration des mots critiques à une étude des phrases critiques
- Introduire une utilisation des modèles n-grammes
- Utiliser une taille d'échantillon appropriée

Perspectives

Amélioration du modèle de régression

- Ajouter d'autres variables explicatives
- Passer d'une exploration des mots critiques à une étude des phrases critiques
- Introduire une utilisation des modèles n-grammes
- Utiliser une taille d'échantillon appropriée

Perspectives

Amélioration du modèle par calcul de compacité

- Affiner la méthode d'étiquetage du corpus
- Orienter l'étiquetage morphosyntaxique vers l'expression d'une idée
- Améliorer la détection des mots centroïdes par rapport à l'idée ou l'opinion recherchée
- Mieux différencier les classes (phrases redondantes)

Perspectives

Amélioration du modèle par calcul de compacité

- Affiner la méthode d'étiquetage du corpus
- Orienter l'étiquetage morphosyntaxique vers l'expression d'une idée
- Améliorer la détection des mots centroïdes par rapport à l'idée ou l'opinion recherchée
- Mieux différencier les classes (phrases redondantes)

Perspectives

Amélioration du modèle par calcul de compacité

- Affiner la méthode d'étiquetage du corpus
- Orienter l'étiquetage morphosyntaxique vers l'expression d'une idée
- Améliorer la détection des mots centroïdes par rapport à l'idée ou l'opinion recherchée
- Mieux différencier les classes (phrases redondantes)

Perspectives

Amélioration du modèle par calcul de compacité

- Affiner la méthode d'étiquetage du corpus
- Orienter l'étiquetage morphosyntaxique vers l'expression d'une idée
- Améliorer la détection des mots centroïdes par rapport à l'idée ou l'opinion recherchée
- Mieux différencier les classes (phrases redondantes)

Fin