Trois recettes d'apprentissage automatique pour un système d'extraction d'information et de classification de recettes de cuisine

Résultats de la campagne d'évaluation DEFT 2013

Eric Charton *Marie-Jean Meurs* *Ludovic Jean-Louis*

Michel Gagnon

POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

UNIVERSITÉ Concordia UNIVERSITY

Wikimeta

Plan

Les campagnes DEFT

Historique.

Création en 2005

Travailler sur des thématiques régulièrement renouvelées relevant de la fouille de textes en langue française.

Principes et méthodes

Proposer des corpus originaux (si possible libres de droits).

Confronter, sur un même corpus, des méthodes et logiciels d'équipes différentes.

Méthodes de fouille de textes étudiées.

- Classification supervisée et non supervisée
- Segmentation
- Extraction d'information
- Appariements

Historique.

Quelques exemples :

- 2005 identifier un locuteur de discours politique
- 2006 segmentation thématique de texte politique
- 2007 détection d'opinion
- 2008 classification en genre et thème
- 2009 fouille d'opinion
- 2010 identification de période et lieu de publication d'un article
- 2011 appariement d'un article avec son résumé
- etc ... (http://deft.limsi.fr/)

Enjeux.

Historiquement

Évaluer des méthodes symboliques et numériques, les comparer.

En pratique

La plupart des campagnes ont vu des méthodes statistiques obtenir les meilleurs résultats.

Des laboratoires majoritairement francophones.

Défis ouverts à des équipes de chercheurs seniors et juniors (étudiants en Master, Doctorat, Docteurs depuis moins d'un an).

Classement final par tâche ou sur l'ensemble des tâches selon les années.

De 5 à 10 équipes selon les éditions, parfois internationales.

Voir http://fr.wikipedia.org/wiki/Défi_fouille_de_texte

Méthode d'évaluation et calendrier.

- Définition des tâches et des métriques ∼ **Janvier**
- Diffusion des corpus d'apprentissage ∼ **Février**
- Diffusion des corpus de test ∼ **Avril**
- **3 jours** pour appliquer les méthodes sur le corpus de test
- Publication des résultats lors de l'**atelier DEFT** (généralement pendant TALN)

La campagne DEFT 2013

Tâches proposées.

Fouille de recettes de cuisine en français

Tâche 1 Classification de recettes selon leur difficulté

Tâche 2 Classification de recettes selon le type de plat

Tâche 3 Appariement du texte d'une recette à son titre

Tâche 4 Extraction des ingrédients contenus dans chaque recette

Corpus.

Recettes de cuisine du site Marmiton.org

Entraînement	13.864	recettes (19,2 MB)
Test T1	2.309	recettes (3 MB)
Test T2	2.307	recettes (2,9 MB)
Test T4	2.306	recettes (2,2 MB)
Format	XML	

Exemple : la recette du *"Gigôt bitume"*

```
<recette id="17129">
<titre>Gigot bitume</titre>
<type>Plat principal</type>
<niveau>Difficile</niveau>
<cout>Moyen</cout>
<ingredients>
  <p>6 gigots d'agneau</p>
  <p>fines herbes, dont thym</p>
  <p>sel, poivre</p>
</ingredients>
<preparation>
Matériel : 1 chantier de bâtiment ou de travaux publics
Envelopper les gigots assaisonnés dans plusieurs couches serrées de papier
kraft-aluminium utilisé en bâtiment. Entourer généreusement de fil de fer
pour assurer que l'ensemble ne se défasse pas. Plonger pendant 25 mn les
gigots ainsi préparé dans le baril de bitume brûlant destiné à l'étanchéité
de la terrasse du bâtiment ou au revêtement de la route en construction.
Retirer et défaire avec précaution.
Excellent plat traditionnel de BTP mais qui ne relève pas de la cuisine
familiale. Je le donne pour information suite à l'appel aux gourmands.
Rouge
</preparation>
</recette>
```

Wikimeta

Tâches de classification :

classer des recettes

✓ selon leur difficulté

✓ selon leur type

Tâche 1 : détection du niveau de difficulté.

- niveaux : Très facile, Facile, Moyennement difficile, Difficile
- étiquette de difficulté
  - pas d'annotation classique (pas de guide fourni aux auteurs)
  - procédé collaboratif (auteur entièrement libre)
- critère subjectif, lié à l'expérience de l'auteur

  ⇒ définition de difficulté :
  - extrêmement variable selon les fiches
  - difficile à modéliser

  ⇒ tâche de détection d'opinion

Corpus d'entraînement, corpus de test.

Répartition des recettes selon leur niveau de difficulté

Niveau	corpus d'entraînement		corpus de test	
	#recettes	% du corpus	#recettes	% du corpus
Très facile	6962	50,2	1132	49,0
Facile	5752	41,5	968	41,9
Moyennement difficile	1068	7,7	189	8,2
Difficile	80	0,6	20	0,9
Total	13862*		2309	

* Deux recettes sont incorrectement étiquetées *tres-facile*.

⇒ fort déséquilibre entre les classes :
"Très facile" et "Facile" : plus de 90% des recettes
"Moyennement difficile" et "Difficile" : moins de 10% et 1%

Caractéristiques discriminantes, espace des données.

- nombre de mots du titre
- nombre de mots des consignes de préparation
- nombre d'ingrédients
- coût
- sous-ensemble de mots discriminants du vocabulaire spécifique de la classe *Moyennement difficile*
- sous-ensemble de trigrammes de mots discriminants
- nombre de verbes dans les consignes pour 3 familles de verbes

$\Rightarrow$ Espace des données : 13.864 vecteurs, dimension 78

Système.

Arbres de régression logistique Logistic Model Tree* (LMT)

- classifieur probabiliste
- résultats pertinents avec peu de données d'apprentissage
- arbre de décision standard de taille réduite
- fonctions de régression logistique** au niveau des feuilles

$$ln\frac{P(Y=j|X=x)}{P(Y=J|X=x)} = \beta_j^T.x \text{ pour } j = 1..J-1$$

$$\text{ie. } P(Y=j|X=x) = \frac{e_j^{\beta_j^T.x}}{1+\sum_{l=1}^{J-1} e_j^{\beta_l^T.x}}, j = 1..J-1 \text{ et } P(Y=J|X=x) = \frac{1}{1+\sum_{l=1}^{J-1} e_j^{\beta_l^T.x}}$$

- minimize les erreurs d'entraînement
- limite le sur-apprentissage

* *Logistic model trees*, Landwehr et al., Machine Learning, 2005
** *Logistic regression, AdaBoost and Bregman distances*, Collins et al., Machine Learning, 2002

Résultats.

Corpus d'entraînement, validation croisée de pas 5

Classe	Précision	Rappel	F-mesure
Très facile	0,671	0,777	0,720
Facile	0,587	0,561	0,574
Moyennement difficile	0,549	0,147	0,232
Difficile	0,400	0,073	0,124
Moyenne pondérée	0,625	0,635	0,618

Matrice de confusion	Classe estimée			
Classe réelle	Très facile	Facile	Moyennement difficile	Difficile
Très facile	5406	1534	22	0
Facile	2446	3228	76	2
Moyennement difficile	197	707	157	7
Difficile	12	33	31	6

Corpus de test

Macro évaluation			Micro évaluation		
Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
0,682	0,375	0,484	0,625	0,625	**0,625**

6 équipes, micro-mesures entre 0,625 à 0,489, moyenne = 0,569, médiane = 0,589.

Tâche 2 : détection du type de plat.

- types : Entrée, Plat principal, Dessert
- étiquette de type choisie par l'auteur
- critère peu subjectif

  $\Rightarrow$ répartition des classes plus homogène que pour la tâche 1
- presque une recette sur deux décrit un plat principal

Type de plat	corpus d'entraînement		corpus de test	
	#recettes	% du corpus	#recettes	% du corpus
Entrée	3246	23,4	562	24,4
Plat principal	6449	46,5	1084	47,0
Dessert	4169	30,1	661	28,6
Total	13864		2307	

Caractéristiques discriminantes, espace des données.

- nombre de mots du titre
- nombre de mots des consignes de préparation
- nombre d'ingrédients
- coût
- ingrédients normalisés et sélectionnés par analyse en composantes principales (ACP)
- sous-ensemble de trigrammes de mots discriminants
- nombre de verbes dans les consignes pour 3 familles de verbes

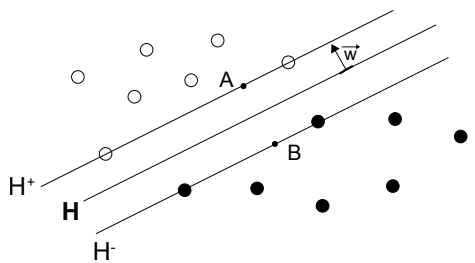
$\Rightarrow$ Espace des données : 13.864 vecteurs, dimension 1.287

Wikimeta

Système.

Séparateur à Vaste Marge* (SVM, Support Vector Machine) à noyau linéaire

- hyperplan séparateur optimal de l'espace des données
- classifieur non probabiliste
- SVM multiclasses (k), one-versus-one, $\frac{k(k-1)}{2}$ modèles binaires
- mêmes paramètres pour tous les modèles
- aptitude à traiter les problèmes de grande dimension
- résultats pertinents avec peu de données d'apprentissage



* *The Nature of Statistical Learning Theory*, Vladimir Vapnik, 1995

Résultats.

Corpus d'entraînement, avec une validation croisée de pas 5

Classe	Précision	Rappel	F-mesure
Plat principal	0,834	0,854	0,844
Dessert	0,967	0,982	0,974
Entrée	0,694	0,648	0,670
Moyenne pondérée	0,841	0,844	0,842

Matrice de confusion	Classe estimée		
Classe réelle	Plat principal	Dessert	Entrée
Plat principal	5507	60	882
Dessert	29	4094	46
Entrée	1064	80	2102

Corpus de test

Macro évaluation		
Précision	Rappel	F-mesure
0,850	0,843	0,847

Micro évaluation		
Précision	Rappel	F-mesure
0,856	0,856	**0,856**

5 équipes, micro-mesures entre 0,889 et 0,746, moyenne = 0,833, médiane = 0,849

Extraire les ingrédients des recettes

à partir de leurs descriptions

Tâche 4 : extraire les ingrédients d'une recette.

- Établir la liste des ingrédients d'une recette à partir de son titre et sa description
- On dispose d'une liste d'ingrédients normalisée (960 entrées)

Recette

Tagliatelles aux crevettes, parfumées au pesto rosso
Cuire les tagliatelles comme indiqué sur le paquet. A côté, faire blondir l'ail dans l'huile d'olive. Ajouter le Pesto Rosso, le jus de citron et les crevettes, chauffer le tout encore 2 minutes. Quand les tagliatelles sont cuites, prélever une cuillerée à soupe d'eau de cuisson et l'incorporer à la sauce. Incorporez les pâtes à la sauce. On peut décorer avec des feuilles de basilic, de la roquette mixée qui donne bon goût, ou avec des tomates séchées.

Ingrédients

Forme initiale	Normalisation
500 g de tagliatelles	tagliatelle
1 pot de pesto rosso	pesto-rosso
2 cuillères à soupe d'huile d'olive	huile-d-olive
3 gousses d'ail hachées	ail
crevettes décortiquées	crevette
jus d'un demi citron pas trop gros	citron

Tâche 4 : Spécificité de la tâche

- Les ingrédients normalisés ne sont pas tous dans le corpus
- Les recommandations des auteurs sont des sources de bruit
- Environ 40% des ingrédients n'apparaissent pas de façon explicite dans les recettes

Cas de figure où les ingrédients sont implicites :

Substitution par un verbe : saler ⇒ sel

Substitution par un nom : légumes ⇒ carottes

Omission volontaire : *Mélangez tous les ingrédients et faire chauffer dans une poêle à crêpes.*

Wikimeta

Tâche 4 : Méthode

- Normalisation des descriptions
- Détection des ingrédients
- Sélection des ingrédients

Tâche 4 : Normalisation des descriptions

- Lemmatisation des descriptions (TreeTagger) et désaccentuation des mots

Recette

tagliatelles au crevette, parfumer au pesto rosso
cuire le tagliatelles comme indiquer sur le paquet . avoir cote , faire blondir le ail dans le huile de olive . ajouter le pesto rosso , le jus de citron et le crevette , chauffer le tout encore @card@ minute . quand le tagliatelles etre cuire , prelever un cuilleree a soupe d' eau de cuisson et le incorporer a le sauce . incorporer le pate a le sauce . on pouvoir décorer avec du feuille de basilic , de le roquette mixer qui donner bon gout , ou avec du tomate seche .

Ingrédients à trouver

Forme initiale	Normalisation
500 g de tagliatelles	tagliatelle
1 pot de pesto rosso	pesto-rosso
2 cuillères à soupe d'huile d'olive	huile-d-olive
3 gousses d'ail hachées	ail
crevettes décortiquées	crevette
jus d'un demi citron pas trop gros	citron

Tâche 4 : Détection des ingrédients

Objectif : Identifier tous les ingrédients d'une description

- Extraction à partir d'expressions régulières
  - 30 expressions spécifiques (beurrée|beurrez|beurrer ⇒ beurre)
  - expressions générées à partir de la liste normalisée
- 1 ingrédient non détecté
- Bruit ⇒ *pesto, huile, jus de citron, pate, roquette, eau, ...*

Recette

tagliatelles au crevette, parfumer au pesto rosso*

cuire le tagliatelles comme indiquer sur le paquet . avoir cote , faire blondir le ail dans le huile de olive* . ajouter le pesto rosso , le jus de citron* et le crevette , chauffer le tout encore @card@ minute . quand le tagliatelles etre cuire , prelever un cuilleree a soupe d' eau de cuisson et le incorporer a le sauce . incorporer le pate a le sauce . on pouvoir décorer avec du feuille de basilic , de le roquette mixer qui donner bon gout , ou avec du tomate seche .

Ingrédients à trouver

Forme initiale	Normalisation
500 g de tagliatelles	tagliatelle
1 pot de pesto rosso	pesto-rosso
2 cuillères à soupe d'huile d'olive	huile-d-olive
3 gousses d'ail hachées	ail
crevettes décortiquées	crevette
jus d'un demi citron pas trop gros	citron

Wikimeta

Tâche 4 : Sélection des ingrédients

Objectif : Choisir les ingrédients les plus pertinents parmi ceux détectés

Sélection initiale (*baseline*) à partir de deux critères :

- fréquence dans le document
- position de la première occurrence

Tâche 4 : Sélection des ingrédients

Sélection par classifieur : 7 critères exploités

- la présence/absence de l'ingrédient dans le titre de la recette
- la forme normalisée de l'ingrédient
- le nombre d'occurrences de l'ingrédient dans la recette
- la position de la première occurrence de l'ingrédient dans le document : pos_{first}
- la position de la dernière occurrence dans le document : pos_{last}
- le taux de recouvrement du document : $rec_{last} = (pos_{last} - pos_{first})/taille(recette)$
- la profondeur : $prof = (1 - pos_{first})/taille(recette)$

Résultats.

Entraînement du modèle de sélection des ingrédients, validation croisée de pas 5

Classe	Précision	Rappel	F-Mesure
Sélectionné	0,806	0,846	0,825
Non sélectionné	0,806	0,759	0,782
Moyenne pondérée	0,806	0,806	0,806

Résultat global de l'extraction des ingrédients

Système	corpus d'entraînement	corpus de test
Baseline	0,5659	0,5678
modèle	0,6702	**0,6462**

5 équipes, MAP entre 0,6662 et 0,4649, moyenne = 0,5916 médiane =0,6287

Conclusions et reproductibilité

Conclusions et perspectives.

- Des tâches diversifiées appliquées sur un corpus construit collaborativement
  - Crée une difficulté de modélisation sur les tâches de classification
  - La tâche 1 est une tâche de détection d'opinion
  - La tâche 2 est une tâche de classification
- La tâche d'extraction d'information est influencée par la méthode de construction des références

Reproductibilité.

Pour l'équipe Wikimeta :

- données d'entrainement (fichiers ARFF) publiques[1]
- logiciels d'extraction de paramètres en accès libre (Java)
- classifieurs utilisés proposés par Weka, logiciel libre

⇒ Expériences **intégralement reproductibles**

(étudiants, futurs participants à DEFT, etc.)

[1] Les corpus sont généralement diffusés par ELDA ou l'organisation.