



Défi: Classification de Textes Français Subjectifs

Michel Génèreux Marina Santini

Natural Language Technology Group

{M.Genereux,M.Santini}@brighton.ac.uk



University of Brighton

Plan de l'exposé

- Méthodologie
 - Algorithme d'apprentissage
 - Traits, dictionnaire bilingue et méthodes de normalisation
- Mise au point
 - Combinaisons de traits et normalisation
 - Combinaison retenue et résultats
- Classification des quatre corpus
- Discussion et conclusion

Apprentissage automatique

- (Semi-)Supervisé (pour la classification)
 - Support Vector Machines*
 - Peut tenir compte d'une grande quantité de traits
 - Logiciel Weka
 - Multi-classes
 - Sélection de traits
- Non supervisé (pour la normalisation des vecteurs de traits)
 - Orientation sémantique (statistiques sur un gros corpus)
 - Ressources lexicales (WordNet)

Traits (1)

- **Groupe 1:** unigrammes provenant de catégories grammaticales pouvant exprimer une émotion ou opinion: **Adjectifs, Noms, Verbes et Adverbes** selon TreeTagger
 - termes entre *ne...fin* de phrase non comptés
- **Groupe 2:** quatorze facettes linguistiques fonctionnelles, empruntées à une approche pour la classification selon le genre
 - nominal, prédicat, passé, passif, pronoms à la 1ère/2ième/3ième personne, verbes d'activité (*saisir*), verbes de communication (*raconter*), verbes mentaux (*savoir*), verbes causatifs (*aider*), verbes d'occurrence (*devenir*), verbes existentiels (*devoir*), verbes aspectuels (*garder*)

Traits (2)

- **Groupe 3:** termes à connotations émotives

- **WordNet-Affect**

- *POSITIF* joie rayonné exalté allègrement réjouir euphorisant

- *NÉGATIF* crainte effrayé terrible alarme impatient timide

- *NEUTRE* apathie impassibilité rêveur langoureusement

- **Big-Six**

- *COLÈRE* ombrage offense folie irritation enragé indignation

- *JOIE* culte adoration chaleur triomphe unité sympathie

- *TRISTESSE* ennui poids apitoiement douleur oppression misère

- *DÉGOÛT* répugnance horreur nausée maladie révolusion

- *PEUR* agitation effroi cercueil timidité suspense hésitation

- *SURPRISE* admiration étonnement stupeur terreur merveille

Dictionnaire bilingue

En raison du manque de ressources en français pour certains calculs de normalisation nécessitant l'accès à un corpus (Waterloo) ou lexique (WordNet, SentiWordNet et liste de P. Turney) en anglais, certaines manipulations ont nécessité la création d'un dictionnaire bilingue anglais-français. Ce dictionnaire se compose de 1244 termes traduits manuellement et provenant des groupes de traits 2 et 3. Ce dictionnaire est nécessaire pour le calcul des facteurs de normalisation *so-pmi-ir*, *so-sim*, *sen* et *hum* (i.e. impliquant des approches non supervisées accédant à un corpus)

Facteurs de normalisation (1)

S'applique à chaque valeur de traits pour un document:

Binaire: Document = [0,1,0,1,0,0,1,1,0,1,...]

Fréquence: Document = [0,2,13,0,0,0,34,3,0,0,...] ajusté à 1000 mots

IDF: Permet d'évaluer l'importance d'un terme i , la supposition étant que l'importance d'un terme diminue à mesure qu'il apparaît dans une proportion grandissante de documents faisant partie du corpus.

$$idf_i = \log \frac{D}{d_i}$$

Facteur de normalisation (2)

SO-PMI-IR^{DIC}: permet de calculer l'orientation sémantique de termes en calculant leur degré d'association avec une liste de mots de référence positifs P (*bon, gentil, excellent, positif, chanceux, correcte* et *supérieur*) et négatifs N (*mauvais, méchant, pauvre, négatif, malchanceux, fautif* et *inférieur*)

$$\sum_p^P A(\text{terme}, p) - \sum_n^N A(\text{terme}, n) = \log_n \frac{D * (nd(\text{terme}_1 \text{ AUTOUR } \text{terme}_2) + 1/D)}{(nd(\text{terme}_1) + 1) * (nd(\text{terme}_2) + 1)}$$

SO-SIM^{DIC}: l'orientation sémantique est obtenue en comptant le nombre d'arcs séparant deux termes dans WordNet: P (*bon-bonifier, gain-gagner, excellence-exceller, supériorité-surpasser*) et N (*mauvais-empirer, perte-perdre, pauvreté-appauvrir, négation-nier*).

Facteur de normalisation (3)

SentiWordNet^{DIC}: ressource qui permet d'obtenir une valeur d'*objectivité* et de subjectivité (*positif* et *négatif*) d'une liste d'un sous-ensemble de termes de WordNet.

TERME	SCORE POSITIF	SCORE NÉGATIF	SCORE NEUTRE
abandonné	0.000	0.000	1.000
accablant	0.250	0.250	0.500
adoration	0.625	0.000	0.375
affligé	0.125	0.625	0.250
affreux	0.000	0.625	0.375
agacer	0.000	0.625	0.375

Facteur de normalisation (4)

HUMain^{DIC}: liste de termes annotés manuellement comme:

NÉGATIF abandon accablant affligé affligeant **ou**

POSITIF acclamation accomplir admiration

BINF: normalisation hybride permettant de normaliser de façon binaire le groupe 1 et par la fréquence les groupes 2 et 3.

Mise au point

Combinaisons de traits et de méthodes de normalisation mise à l'épreuve dans le but d'en extraire les mieux adaptées

Corpus: Relectures

Groupe 1: 500 traits les plus fréquents

Traits/Normalisation	binaire	fréquence	idf	so-pmi-ir	so-sim	sen	vrai	binf
Adjectifs	1ABHI	2 FG	3G	4	5	6	7	J
Noms	1 B	2	3	4	5	6	7	
Verbes	1 B I	2	3	4	5	6	7	
Facettes	1	2C	3	4	5	6	7	
WordNet-Affect	1	2 D K	3	4	5	6	7	J
Big-Six	1	2 E	3	4	5	6	7	
Adverbes	HI							J

Mise au point: résultats

Validation croisée 10-fois

Combinaison	H	J	A	I	1	F	3	2	B
Exactitude	50.6	50.3	49.4	47.2	45.6	45.3	44.7	44.4	43.7
Combinaison	E	G	4	D	6	5	K	7	C
Exactitude	42.9	42.7	42.7	42.7	42.2	42.2	42.1	41.7	41.4

- Choix 1: combinaison H
 - Traits: adjectifs, adverbes
 - Normalisation: binaire
- Choix 2: combinaison J
 - Traits: adjectifs, adverbes et WordNet-Affect
 - Normalisation: hybride (binaire et fréquence)
- Choix 3: combinaison A
 - Traits: adjectifs
 - Normalisation: binaire

Expériences (1)

- Chacun des quatre corpus fût classé en utilisant H, J et A
- Pour des raisons de rapidité de traitement:
 - le corpus 4 (débat) fût amputé de 80%
 - les corpus 1, 2 et 4 n'utilisent que les 100 traits les plus fréquents pour le groupe 1

Expériences (2)

Combinaison	Validation-Croisée 3-fois				Matrice de confusion			Classe
	Exactitude	Précision	Rappel	F-score	zéro (309)	un (615)	deux (1150)	
H	56.2	0.425	0.246	0.311	76	61	172	zéro
		0.415	0.270	0.327	50	166	399	un
		0.618	0.803	0.699	53	173	924	deux
J	55.4	0.392	0.236	0.295	73	64	172	zéro
		0.399	0.259	0.314	55	159	401	un
		0.615	0.797	0.695	58	175	917	deux
A	55.7	0.382	0.220	0.279	68	34	207	zéro
		0.392	0.115	0.178	53	71	491	un
		0.593	0.884	0.710	57	76	1017	deux
Moyenne	55.8	0.470	0.429	0.423	Fin de l'entraînement			
H avec données de test		0.48	0.43	0.45	Tel que communiqué par DEFT 07			
J avec données de test		0.49	0.44	0.46	Tel que communiqué par DEFT 07			
A avec données de test		0.50	0.39	0.44	Tel que communiqué par DEFT 07			
Moyenne Test		0.49	0.42	0.45				
Tous les participants		0.53	0.48	0.50	Tel que communiqué par DEFT 07			
		±0.10	±0.07	±0.07				

Corpus 1: Critiques $(55.8-33)/33 = 69\%$ plus exact que par chance

F-score à l'intérieur de l'écart-type autour de la moyenne

Bias en faveur des classes les plus nombreuses

Expériences (3)

Combinaison	Validation-Croisée 3-fois				Matrice de confusion			Classe
	Exactitude	Précision	Rappel	F-score	zéro (497)	un (1166)	deux (874)	
H	64.6	0.610	0.453	0.520	225	216	56	zéro
		0.639	0.703	0.699	113	820	233	un
		0.673	0.681	0.677	31	248	595	deux
J	62.1	0.568	0.435	0.493	216	224	57	zéro
		0.616	0.679	0.646	127	792	247	un
		0.651	0.649	0.650	37	270	567	deux
A	63.2	0.551	0.467	0.505	232	231	34	zéro
		0.617	0.669	0.642	160	780	226	un
		0.695	0.677	0.686	29	253	592	deux
Moyenne	63.3	0.624	0.604	0.613	Fin de l'entraînement			
H avec données de test		0.64	0.60	0.62	Tel que communiqué par DEFT 07			
J avec données de test		0.64	0.61	0.63	Tel que communiqué par DEFT 07			
A avec données de test		0.61	0.59	0.60	Tel que communiqué par DEFT 07			
Moyenne Test		0.63	0.60	0.62				
Tous les participants		0.69 ±0.10	0.64 ±0.09	0.66 ±0.09	Tel que communiqué par DEFT 07			

Corpus 2: Jeux vidéo (63.3-33)/33 = 92% plus exact (meilleur rés.)

F-score à l'intérieur de l'écart-type autour de la moyenne

Expériences (4)

Combi- naison	Validation-Croisée 3-fois				Matrice de confusion			Classe
	Exactitude	Précision	Rappel	F-score	zéro (227)	un (278)	deux (376)	
H	50.2	0.435	0.445	0.440	101	70	56	zéro
		0.439	0.482	0.460	63	134	81	un
		0.602	0.551	0.575	68	101	207	deux
J	50.1	0.441	0.445	0.443	101	72	54	zéro
		0.436	0.482	0.458	59	134	85	un
		0.597	0.548	0.571	69	101	206	deux
A	47.9	0.427	0.436	0.431	99	76	52	zéro
		0.397	0.442	0.418	68	123	87	un
		0.590	0.532	0.559	65	111	200	deux
Moyenne	49.4	0.475	0.485	0.484	Fin de l'entraînement			
H avec données de test		0.47	0.47	0.47	Tel que communiqué par DEFT 07			
J avec données de test		0.46	0.46	0.46	Tel que communiqué par DEFT 07			
A avec données de test		0.43	0.44	0.43	Tel que communiqué par DEFT 07			
Moyenne Test		0.45	0.46	0.45				
Tous les participants		0.48	0.46	0.47	Tel que communiqué par DEFT 07			
		±0.05	±0.05	±0.05				

Corpus 3: Relectures $(49.4-33)/33 = 50\%$ plus exact

F-score à l'intérieur de l'écart-type autour de la moyenne

Expériences (5)

Combinaison	Validation-Croisée 3-fois				Matrice de confusion		Classe
	Exactitude	Précision	Rappel	F-score	zéro (2080)	un (1380)	
H	62.7	0.643	0.852	0.733	1773	307	zéro
		0.564	0.288	0.381	983	397	un
J	63.4	0.650	0.849	0.736	1766	314	zéro
		0.576	0.309	0.403	953	427	un
A	64.6	0.651	0.886	0.751	1843	237	zéro
		0.623	0.284	0.390	988	392	un
Moyenne	63.6	0.618	0.578	0.566	Fin de l'entraînement		
H avec données de test		0.56	0.52	0.54	Tel que communiqué par DEFT 07		
J avec données de test		0.57	0.54	0.55	Tel que communiqué par DEFT 07		
A avec données de test		0.60	0.54	0.57	Tel que communiqué par DEFT 07		
Moyenne Test		0.58	0.53	0.55			
Tous les participants		0.65 ±0.06	0.63 ±0.06	0.64 ±0.06	Tel que communiqué par DEFT 07		

Corpus 4: Débats $(63.6-50)/50 = 27\%$ plus exact (pire résultat)

F-score à l'intérieur de l'écart-type autour de la moyenne

Discussion (1)

- Mise au point: dans le cadre d'une approche supervisée SVM, unigrammes (adjectifs et adverbes) avec normalisation binaire et WordNet-Affect avec normalisation fréquence donnent les meilleurs résultats
- Impact limité des méthodes de normalisation alternatives: faible dimension du dictionnaire bilingue?
- Facettes linguistiques: sous-ensemble exploratoire
 - permet d'éviter l'*overfitting* et offre une interprétation fonctionnelle: e.g. 1ère personne suggère un texte argumentaire (commentaires ou opinions)
 - pronom pers. 2ème pers. → jeux vidéo, 1ère pers. → débats
 - verbes d'activité → jeux vidéo, mentaux/communication → relectures

Discussion (2)

- F-scores varient entre 54% et 57% pour deux classes et entre 43% et 63% pour trois classes: ces résultats sont largement au-dessus de résultats obtenus par chance
- Étant donné le petit nombre de traits provenant des groupes 2 et 3, leur contribution est honorable
- Mise au point avec Relectures (langage neutre) difficile à classer en soi et non optimale pour les débats

Conclusion

- Approche classique (SVM, unigramme) augmentée de traits affectifs et fonctionnelles, normalisations inspirés d'approches non-supervisées
- Résultats compétitifs
- Contribution limitée:
 - des groupes 2 et 3, attribuables à un nombre limité de traits
 - des méthodes de normalisation, partiellement dû à un dictionnaire limité