Akio

Beyond Customer eXpectations

22/01/2019

Une étude empirique de classification textuelle multi-étiquette pour la relation-client

Gil Francopoulo, Léon-Paul Schaub

Lynda Ould Younès

SOMMAIRE

La relation-client

- Définition CRM (gestion de la relation-client)
  - Analyse des informations clients
    - Fiches
    - Données
  - Outils de gestion (Akio Unified, Zoho, Salesforce..)
- Gestion de l'interaction client
  - Voix du client
  - Différents canaux : écrits/ vocaux
  - Masse immense de données non structurées

- le TAL dans la relation-client
  - automatisation des interactions
  - chatbot
  - **text mining**
    - **classification fine**
    - **s'affranchir de la langue**
  - ...

Corpus

- Caractéristiques
  - corpus “client”
    - mal écrit
    - informel
    - équivalent public introuvable
  - séparé en 6 secteurs d’activité :
    - hôtels-restaurants
    - transports
    - banques
    - sites de rencontre
    - e-commerce
    - assurances
  - dans plusieurs langues nativement
  - avec des topics exclusifs et d’autres transversaux (livraison VS site web)

Prétraitement et analyse

- utilisation de TagParser [Francopoulo 2018]
  - tokeniseur, correcteur
  - tagger-chunker
  - analyse syntaxique dépendances
  - négation
  - DEN + co-références
- Normalisation
  - certains symboles
  - anonymisation

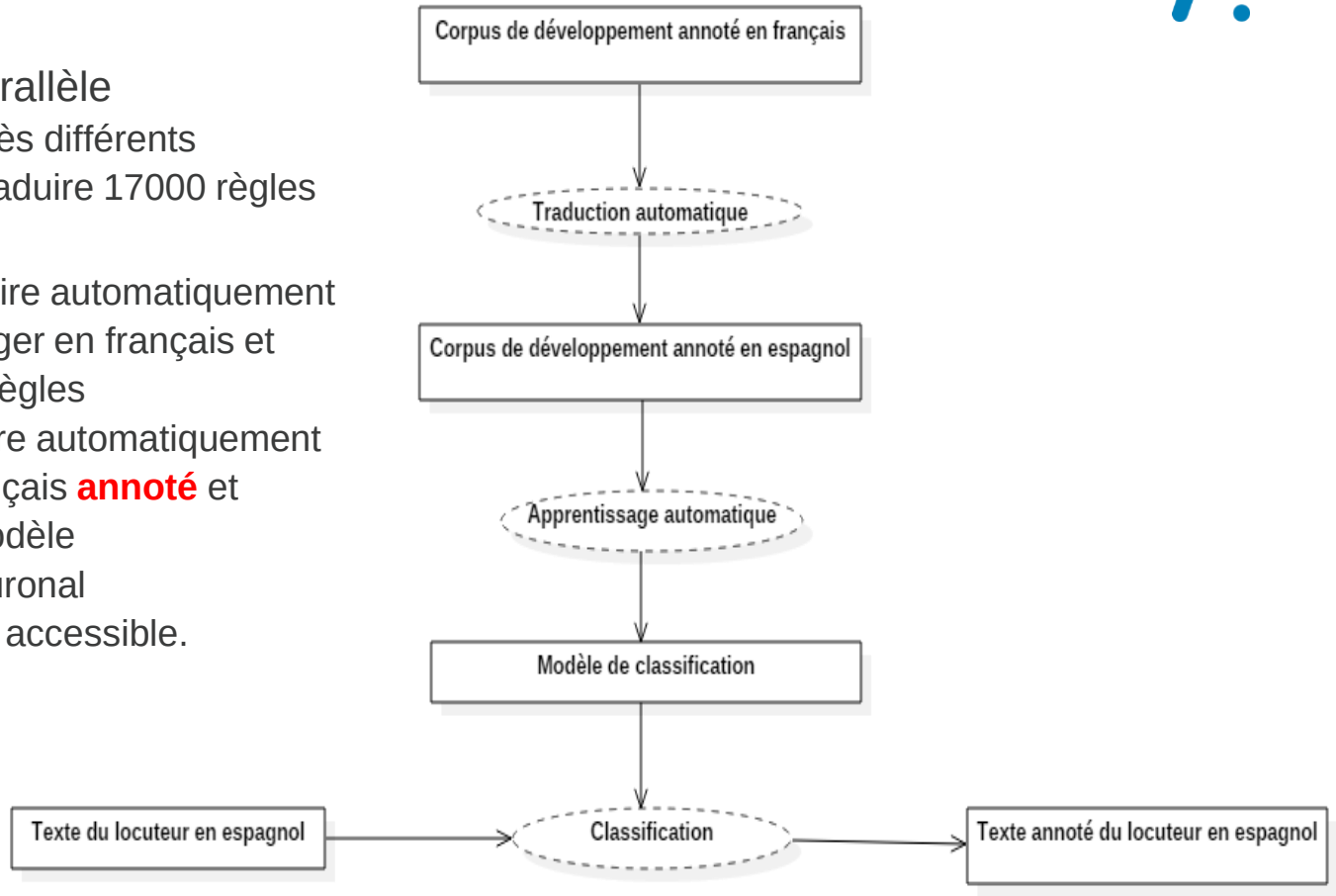
- Système symbolique complet
  - pour le français uniquement
  - basé sur de la sémantique de surface
  - 3 Axes d'analyse
    - modalités/ sujets/ opinion
- Quelques chiffres :
  - textes de petite taille (5 lignes)
  - 17 500 règles (e-commerce = 9000)
  - F-mesure -> 85%
  - plus de deux ans
  - 150 catégories juste pour e-commerce

Comment obtenir un système complet pour les autres langues ?

sous-corpus	*mode de construction*	*F-mesure (évaluation des règles symboliques)*	*nombre de textes (en milliers)*	*nombre de mots (en millions)*
GOLD	annotation manuelle	1.0	8,757	0,402
SILVER	application de règles symboliques sur textes inconnus	0.85	15,028	0,586
BRONZE	expansion du GOLD par remplacement synonymique	1.0	15,269	1,193
total	N/D	N/D	3,9	2,2

Classification multilingue

- Pas de corpus parallèle
  - textes natifs très différents
  - contraint de traduire 17000 règles
- Deux solutions :
  - En prod, traduire automatiquement un texte étranger en français et appliquer les règles
  - En dev, traduire automatiquement un corpus français **annoté** et générer un modèle statistique/neuronal
  - 2e option plus accessible.



Algorithmes d'apprentissage automatique

- Quatre types de corpus
  - brut + fléchi
  - lemmatisé + lemmatisé filtré
- Trois grands types d'algorithmes :
  - statistiques (WEKA)
    - SMO (SVM)
    - SVM seul
    - Bayésiens naïfs
    - Classifieur SGD
  - neuronaux (TensorFlow)
    - CNN et Bi-LSTM
  - FastText (Facebook)
- Deux représentations de données :
  - BOW
  - Word embeddings

Méthodologie :

- Séparation du GOLD utilisation de SILVER et génération du BRONZE :
  - apprentissage/test = 90/10
  - les 10% restent intouchés
  - génération du BRONZE avec les 90% d'apprentissage
  - apprentissage avec 90% GOLD + SILVER + BRONZE
  - maximum 1 semaine d'apprentissage

Résultats et temps de traitement

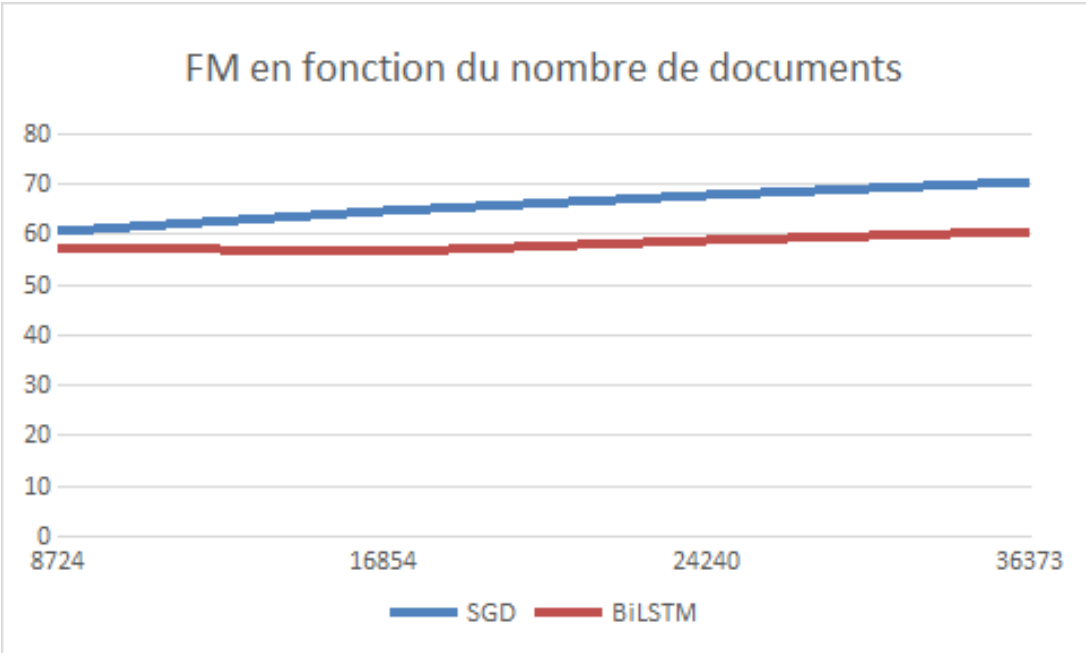
Baseline RuleBased = 0,85

Performances (en F-mesure)

nom	brut	fléchi corrigé	lemme corrigé	lemme corrigé filtré
Bayésiens naïfs	0,3	0,33	0,34	0,37
CNN	0,45	0,42	0,44	0,42
FastText	0,52	0,50	0,46	0,46
Bi-LSTM	0,60	0,60	0,59	0,604
SVM	0,70	0,69	0,64	0,62
class. SGD	0,73	0,733	0,71	0,70
SMO	0,74	**0,75**	0,72	0,71

Temps de traitement

Nom	apprentissage le plus lent (brut)	apprentissage le plus rapide (lemmes)	temps d'inférence
FastText	0h15 min	0h15	2 sec
CNN	1h03 min	0h37	2 sec
Bayésiens naïfs	2h30 min	0h50	**7 min**
SVM	4h50	1h44	39 sec
class. SGD	6h	2h	31 sec
Bi-LSTM	27h	14h30	10 sec
SMO	**5j 16h**	**15h**	21 sec



Analyse des résultats

- Au niveau des chiffres :
  - meilleure **précision** et **FM** pour les algorithmes **statistiques**
  - meilleur **rappel** pour les algorithmes **neuronaux**
- Pour l'instant résultats proportionnels à la taille du corpus -> plafond de verre non atteint
- qualité des traductions possible explication au faible rendement
  - problèmes de traduction pour **ironie, sarcasme et insultes**
- spécificité du corpus possible explication :
  - les **Word Embeddings** de Wikipedia pas assez représentatifs
  - à l'avenir faire une étude **texto/lexicométrique** pour connaître les concordances, cooccurrences, AFC…
- La lemmatisation n'améliore pas les résultats
- Paradoxal : statistique **meilleur** mais **plus long** que les réseaux de neurones

Futurs développements

Prétraitement

- corpus pré-entraîné “maison”
- enrichir les vecteurs avec :
  - P.O.S
  - marqueurs syntaxiques
  - n-grammes

Utiliser des techniques d’optimisation :

- bagging
- boosting
- mécanisme d’attention
- co-apprentissage/adversarial

Améliorer les traductions

Combiner différentes couches de réseaux :

- selon l’état de l’art
- pour obtenir un bon modèle CRM
- utiliser en sortie un CRF

Essayer de trouver un corpus public pour mieux évaluer le système

Conclusion

- Classification multilingue possible à partir d'un corpus français annoté
  - Stats VS neuronal : winner is **Statistical algorithms**
  - Malgré la grande finesse (**plus de 150 labels** possibles pour chaque texte)
  - Nécessité de plus de prétraitements linguistiques malgré les réseaux de neurones
  - La taille du corpus a un impact direct et proportionnel sur les performances pour tout algo

→ **IL FAUT AUGMENTER LA TAILLE DU CORPUS**