Étude expérimentale de la classification interlingue pour la gestion de la relation client

Gil Francopoulo*, Léon-Paul Schaub**
Lynda Ould Younes***

*AKIO
gfrancopoulo@akio.com,
http://www.akio.com
**AKIO + LIMSI-CNRS
lpschaub@akio.com
***AKIO
louldyounes@akio.com

1 Introduction

La gestion de la relation client est l'analyse des données des interactions des clients. Ce qui est important pour l'industrie, ce n'est pas de déterminer si un document exprime globalement une opinion positive ou négative, mais au contraire de détecter si le client exprime des opinions sur des sujets précis. Par exemple, il peut être satisfait des frais de livraison, tout en étant très mécontent du délai de livraison. Les langues couvertes sont le français (considéré comme la langue native), l'anglais (natif mais moins développé donc non utilisé lors de l'apprentissage), l'espagnol, l'allemand, le portugais et l'italien. Le logiciel s'appelle AKIO Analytics.

2 Prétraitement, catégories et flux de données

Concernant le prétraitement, nous avons un pipeline linguistique comprenant un tokeniseur, un correcteur orthographique et grammatical, un tagger-chunker statistique, un analyseur syntaxique en dépendance, un annotateur de la négation. L'entrée peut être de quatre types : a) la chaîne brute d'origine, b) une suite de formes fléchies corrigées, c) une suite pleine de lemmes, d) une suite filtrée de lemmes corrigés. Le filtrage des lemmes consiste à ne prendre que les parties du discours comme les noms ou les adverbes de négation et d'ignorer d'autres mots comme les déterminants. Nous gérons trois types de catégories : les modalités d'expression, les thèmes et les opinions. Les catégories sont précises et nombreuses (179 catégories). Une modalité d'expression sera par exemple "question", un thème sera StoreDelivery et une opinion sera MissingItemNeg. Nous traduisons le corpus de développement automatiquement du français vers l'espagnol, nous transférons les marques de catégories depuis le document français et ensuite, de manière monolingue, en espagnol, nous apprenons un modèle de classification qui sera appliqué en exploitation.

3 Comparaisons des temps et de la qualité

nom	multi-étiquettes	ordre	CUDA	temps d'appr. chaînes brutes	temps d'appr. lemmes corrigés	temps d'inférence
NB	non	BOW	non	2 h 30	50 mn	7 mn
SGD	non	BOW	non	6 h	2 h	31 s
SVM	non	BOW	non	4 h 50	1 h 44	39 s
SMO	non	BOW	non	5 jours 16 h	15 h	21 s
FastText	oui	WE	non	15 mn	15 mn	2 s
BiLSTM	oui	WE	oui	7 jours 12 h	2 h	10 s
CNN	oui	WE	oui	45 mn	40 mn	2 s

TAB. 1 – *Temps de traitement.*

Les mesures de qualité des différentes sessions sont présentées dans le tableau 2 en fonction du niveau linguistique de l'entrée de la catégorisation.

nom	chaînes brutes			formes fléchies corrigées			lemmes corrigés non filtrés			lemmes corrigés filtrés		
	R	P	FM	R	P	FM	R	P	FM	R	P	FM
NB	68	19	30,0	74	21	32,6	72	22	34,2	73	25	37,4
SGD	68	79	72,9	70	76	73,3	69	73	71,0	69	72	70,4
SVM	58	87	69,8	57	87	68,7	50	88	64,0	48	87	61,6
SMO	68	83	74,4	70	81	75,2	67	78	72,2	65	79	71,3
FastText	45	61	51,6	45	55	49,7	44	47	45,7	45	48	46,4
BiLSTM	74	36	48,7	76	37	50,0	77	38	51,4	78	40	53,2
CNN	67	40	50,8	65	36	46,8	70	38	49,4	69	32	44,7

TAB. 2 – *Qualité.*

Concernant le choix des options, si on se focalise sur les FM au dessus de 70, l'option des formes brutes avec SMO n'est pas réaliste car le temps d'apprentissage est trop long. En l'état actuel de nos évaluations, nous optons pour le classifieur SGD avec les lemmes corrigés filtrés, étant entendu que ce choix pourrait être remis en question à la lumière de futurs développements.