Classification supervisée multi-étiquette en actes de dialogue : analyse discriminante et transformations de Schoenberg

Christelle Cocco[1]

[1] Université de Lausanne – christelle.cocco@unil.ch

Abstract

This work studies the multi-label classification of turns in simple English Wikipedia talk pages into dialog acts. The treated dataset was created and multi-labeled by (Ferschke et al., 2012). The first part analyses dependences between labels, in order to examine the annotation coherence and to determine a classification method. Then, a multi-label classification is computed, after transforming the problem into binary relevance. Regarding features, whereas (Ferschke et al., 2012) use features such as uni-, bi-, and trigrams, time distance between turns or the indentation level of the turn, other features are considered here: lemmas, part-of-speech tags and the meaning of verbs (according to WordNet). The dataset authors applied approaches such as Naive Bayes or Support Vector Machines. The present paper proposes, as an alternative, to use Schoenberg transformations which, following the example of kernel methods, transform original Euclidean distances into other Euclidean distances, in a space of high dimensionality.

Résumé

Ce travail étudie la classification supervisée multi-étiquette en actes de dialogue des tours de parole des contributeurs aux pages de discussion de *Simple English Wikipedia* (Wikipédia en anglais simple). Le jeu de données considéré a été créé et multi-étiqueté par (Ferschke et al., 2012). Une première partie analyse les relations entre les étiquettes pour examiner la cohérence des annotations et pour déterminer une méthode de classification. Ensuite, une classification supervisée multi-étiquette est effectuée, après recodage binaire des étiquettes. Concernant les variables, alors que (Ferschke et al., 2012) utilisent des caractéristiques telles que les uni-, bi- et trigrammes, le temps entre les tours de parole ou l'indentation d'un tour de parole, d'autres descripteurs sont considérés ici : les lemmes, les catégories morphosyntaxiques et le sens des verbes (selon *WordNet*). Les auteurs du jeu de données ont employé des approches telles que le *Naive Bayes* ou les Séparateurs à Vastes Marges (SVM) pour la classification. Cet article propose, de façon alternative, d'utiliser et d'étendre l'analyse discriminante linéaire aux transformations de Schoenberg qui, à l'instar des méthodes à noyau, transforment les distances euclidiennes originales en d'autres distances euclidiennes, dans un espace de haute dimensionnalité.

Mots-clés : Actes de dialogue, pages de discussion de Wikipédia, classification multi-étiquette, *WordNet*, analyse discriminante, transformations de Schoenberg.

1. Introduction

Les articles de Wikipédia sont créés par ses contributeurs, qui partagent leurs informations et leurs critiques sur des pages de discussion, chaque article étant lié à une page de discussion. Ces discussions fournissent une base de données que (Ferschke et al., 2012) ont segmentée, pour *Simple English Wikipedia* (Wikipédia en anglais simple), en *tours de parole* (interventions successives des intervenants). Ils ont ensuite annoté ces tours de parole avec des actes de dialogue (section 2). De nombreux travaux se sont intéressés à la classification de dialogues écrits ou oraux en actes de dialogue (*dialog acts*) ou en actes de langage (*speech acts*), servant à caractériser la fonction d'un énoncé dans un dialogue (Austin, 1962 ; Searle

1969). Les actes de dialogue peuvent être différents selon le but de la classification (pour une comparaison des principaux actes de dialogue et de langage utilisés, voir par exemple Goldstein et Sabin, 2006). (Ferschke et al., 2012) utilisent leur propre jeu d'étiquettes d'actes de dialogue avec pour but de comprendre les « efforts de coordination pour l'amélioration d'un article ». Dans un second temps, ils ont procédé à une classification multi-étiquette. En général, un acte de dialogue est attribué à chaque énoncé, ce qui conduit à une classification ordinaire mono-étiquette. Dans ce jeu de données, chaque tour de parole peut se voir attribuer un ou plusieurs actes de dialogue, ce qui conduit à une classification multi-étiquette (cf. section 4.1.2) des tours de paroles en actes de dialogue. Pour examiner la cohérence de ces annotations et pour déterminer une méthode de classification, une première partie analyse les relations entre les étiquettes (section 3).

Concernant les actes de dialogue, (Colineau et Caelen, 1995) distinguent quatre types de marqueurs : linguistiques (morphologiques, syntaxiques et lexicaux), prosodiques, situationnels (phases du dialogue et règles d'enchaînement préférentiel) et du geste. Ici, le jeu de données contient exclusivement des textes écrits, sans annotation des actions qui découlent du dialogue ; ainsi seuls les marqueurs linguistiques et situationnels peuvent être employés. (Ferschke et al., 2012) utilisent les deux types de marqueurs, *i.e.* uni-, bi- et trigrammes (linguistiques) et temps entre les tours de parole, leur indentation, etc. (situationnels), puis les combinent. Ce travail propose d'utiliser trois autres caractéristiques (*features*), toutes de nature linguistique, et de les étudier *séparément* pour mieux comprendre l'impact de chacune d'entre elles, *sans visée de performance globale*. Les trois types de caractéristiques employées sont (section 4.1.1) : **les lemmes** (unigrammes), donnant des résultats légèrement meilleurs que les mots bruts dans la classification de *chats* en actes de dialogues (Kim et al., 2010) ; **les catégories morphosyntaxiques** (CMS), dont l'intérêt pour la classification en actes de dialogue est démontré dans plusieurs travaux (voir par exemple (Cohen et al., 2004) ou (Boyer et al., 2010)) ; et **le sens des verbes selon *WordNet*** (Fellbaum, 1998). Deux articles, l'un étudiant la classification de messages sur des forums (Qadir et Riloff, 2011), l'autre la classification d'e-mails (Goldstein et Sabin, 2006), concluent que des classes de verbes (selon des listes prédéfinies) aident à la reconnaissance de certains actes de langage. L'idée, un peu différente ici, est de voir si les classes recréées à l'aide de WordNet permettent de telles reconnaissances dans le jeu de données étudié.

Finalement, une autre originalité de ce travail concerne la méthode de classification. Alors que les auteurs du jeu de données ont employé des approches classiques, telles que le *Naive Bayes* ou les *Séparateurs à Vastes Marges* (SVM), cet article utilise l'analyse discriminante linéaire afin de l'étendre aux transformations de Schoenberg qui, à l'instar des méthodes à noyau, transforment les dissimilarités euclidiennes originales en d'autres dissimilarités euclidiennes, dans un espace de haute dimensionnalité (section 4.1.3). Les résultats ainsi obtenus sont exposés dans la section 4.2, puis les extensions possibles de la méthode sont discutées dans la section 5.

2. Données

Les données utilisées dans ce projet sont celles de (Ferschke et al., 2012) et mises librement à disposition sur Internet (`http://www.ukp.tu-darmstadt.de/data/wikidiscourse`). Elles concernent les pages de discussion de Wikipédia en anglais simple. Une partie de ces pages de discussion ont été extraites, segmentées automatiquement en tours de parole (1 450 au total), et classifiées en actes de dialogue par trois annotateurs pour constituer un corpus de référence (pour la structure des données et le détail, voir (Ferschke et. al., 2012)). Les

étiquettes se divisent en quatre groupes principaux : interpersonnel (*Interpersonal*), critique d'articles (*Article Criticism*), contenu de l'information (*Information Content*) et performativité (*Explicit Performative*), lesquelles se subdivisent en un jeu de 17 étiquettes.

Les étiquettes **interpersonnelles** « décrivent l'attitude qui est exprimée envers les autres participants dans la discussion et/ou les commentaires ». Ces étiquettes se divisent en trois sous-étiquettes : « une approbation ou un rejet partiel » (ATTP), « une attitude négative envers un autre participant ou un rejet » (ATT-) et « une attitude positive envers un autre participant ou une approbation » (ATT+). Les étiquettes de **critique d'articles** « dénotent les commentaires qui identifient des insuffisances dans l'article. La critique peut porter sur l'article entier ou sur une partie de l'article ». Cet ensemble se subdivise en sept parties : « les insuffisances de langage ou de style » (CL), « un contenu incomplet ou manque de détail » (CM), « d'autres sortes de critiques » (CO), « des problèmes objectifs » (COBJ), « des problèmes structurels » (CS), « un contenu inapproprié ou inutile » (CU) et « le manque de précision ou d'exactitude » (CW). Les étiquettes sur **le contenu de l'information** « décrivent la direction de la communication ». Elles se divisent en trois catégories : « correction de l'information » (IC), « apport d'information » (IP) et « demande d'information » (IS). Quant aux étiquettes de **performativité**, elles concernent « l'annonce, le rapport ou la suggestion d'activités d'édition ». Elles se divisent en quatre sous-catégories : « engagement à une action dans le futur » (PFC), « rapport d'une action accomplie » (PPC), « référence explicite ou indicateur » (PREF) et « suggestion, recommandation ou demande explicite » (PSR).[1]

3. Lien entre étiquettes

Chaque tour de discussion pouvant avoir plusieurs étiquettes ou classes, il semblait pertinent de commencer par déterminer s'il existe des liens entre ces étiquettes. En plus de permettre une meilleure compréhension de l'annotation et de sa cohérence, cette première étude permet de choisir une méthode de classification multi-étiquette appropriée, *i.e.* prenant en compte ou non le lien entre les étiquettes (cf. section 4.1.2).

3.1. Méthode

Pour chaque paire d'étiquettes ou classes *l* et *l'*, une table de contingence 2×2 a été créée, représentant le nombre d'absences et de présences (codées 0 et 1) de chaque classe:

		Classe *l'*	
Classe *l*		Présence	Absence
Présence		n_{11}	n_{10}
Absence		n_{01}	n_{00}

Il s'agit alors de mesurer le lien qui existe entre les deux classes. Pour ce faire, il existe des indices particulièrement adaptés aux accords entre deux partitions binaires (voir par exemple (Warrens, 2008)), dont deux ont été utilisés ici. Le premier est simplement la corrélation de Pearson appliquée à deux variables binaires (Yule, 1912). Cet indice est connu sous le nom de *coefficient phi*, en raison de son rapport avec le chi carré ($\phi^2 = \chi^2/N$ où $N = n_{11} + n_{00} + n_{10} + n_{01}$):

[1] Les définitions de ce paragraphe sont une traduction personnelle des définitions données dans (Ferschke et. al., 2012). Des exemples de tours de parole appartenant à chacune des étiquettes et extraits du jeu de données se trouvent dans leur article.

$$\phi_{ll'} = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{11}+n_{10})(n_{01}+n_{00})(n_{11}+n_{01})(n_{10}+n_{00})}}$$

On a $\phi_{ll'} = 1$, si et seulement si chaque tour de parole appartenant (respectivement n'appartenant pas) à la classe *l* appartient (respectivement n'appartient pas) aussi à la classe *l'*. Inversement, $\phi_{ll'} = -1$, indique que tous les tours de parole apparaissant dans la classe *l* n'apparaissent pas dans la classe *l'*, et vice-versa. Et si $\phi_{ll'} = 0$, alors il n'y a pas de lien entre les deux classes. Il est possible de tester sa significativité en le comparant à la valeur critique $\sqrt{\chi^2_{1-\alpha/2}[1]}$, qui vaut 0.059 pour $\alpha = 5\%$.

Le second indice de dépendance est le *Q* de Yule et est défini comme (Yule, 1900):

$$Q_{ll'} = \frac{n_{11}n_{00} - n_{10}n_{01}}{n_{11}n_{00} + n_{10}n_{01}}$$

Si $Q_{ll'} = 1$, alors la classe *l* est incluse dans la classe *l'* ou inversement, tandis que si $Q_{ll'} = -1$, soit aucun tour de parole n'appartient simultanément aux deux classes, soit tous les tours de parole appartiennent à au moins une des deux classes. $Q_{ll'} = 0$ s'interprète comme $\phi_{ll'} = 0$.

Dans un second temps, à partir de la matrice des corrélations entre toutes les classes $\Phi = (\phi_{ll'})$, une analyse en composantes principales (ACP) a été effectuée afin de visualiser les relations entre les différentes étiquettes et étudier la diversité de ces dernières.

	ATTP	ATT-	ATT+	CL	CM	CO	COBJ	CS	CU
ATTP		-0.039	-0.051	-0.051	-0.028	-0.028	0.047	-0.049	-0.026
ATT-	**-1**		-0.055	-0.107*	-0.053	-0.047	0.008	-0.071*	-0.026
ATT+	**-1**	-0.527		-0.089*	-0.013	-0.010	0.022	-0.051	-0.030
CL	-0.707	**-1**	-0.532		0.018	-0.046	0.056	0.043	-0.004
CM	-0.477	-0.590	-0.084	0.086		0.031	-0.003	0.123*	0.010
CO	**-1**	**-1**	-0.099	-0.464	0.253		0.003	-0.020	-0.032
COBJ	0.564	0.115	0.229	0.415	-0.042	0.059		-0.009	0.067*
CS	**-1**	-0.809	-0.364	0.183	0.503	-0.222	-0.130		0.001
CU	**-1**	-0.455	-0.383	-0.034	0.098	**-1**	0.632	0.009	
CW	**-1**	-0.381	-0.301	-0.034	-0.064	-0.417	0.229	-0.271	0.473
IC	0.008	0.204	-0.670	0.817	-0.152	0.279	-0.105	-0.118	-0.333
IP	0.842	0.723	0.232	0.722	0.605	0.287	0.638	0.663	0.760
IS	-0.288	-0.358	-0.534	0.132	0.284	0.410	-0.387	0.042	0.281
PFC	0.435	-0.424	0.584	-0.370	0.074	**-1**	0.180	-0.320	-0.059
PPC	-0.196	-0.597	-0.144	-0.742	-0.736	-0.576	-0.311	-0.776	-0.523
PREF	0.347	0.058	-0.415	-0.594	-0.594	-0.207	**-1**	-0.648	-0.139
PSR	-0.722	-0.562	-0.168	0.683	0.810	0.583	0.418	0.845	0.528

	CW	IC	IP	IS	PFC	PPC	PREF	PSR
ATTP	-0.034	0.001	0.080*	-0.026	0.046	-0.023	0.026	-0.075*
ATT-	-0.030	0.033	0.118*	-0.050	-0.034	-0.098*	0.005	-0.099*
ATT+	-0.032	-0.080*	0.056	-0.089*	0.137*	-0.035	-0.033	-0.043
CL	-0.005	0.353*	0.190*	0.036	-0.048	-0.188*	-0.053	0.303*
CM	-0.007	-0.021	0.118*	0.067*	0.010	-0.133*	-0.038	0.309*
CO	-0.024	0.036	0.040	0.072*	-0.044	-0.070*	-0.011	0.125*
COBJ	0.017	-0.008	0.059*	-0.030	0.013	-0.032	-0.025	0.062*
CS	-0.028	-0.018	0.138*	0.009	-0.034	-0.151*	-0.044	**0.358***
CU	0.057	-0.025	0.084*	0.042	-0.004	-0.061*	-0.007	0.103*
CW		0.222*	0.120*	0.021	0.033	-0.084*	0.034	0.060*
IC	0.758		0.176*	-0.072*	-0.053	-0.124*	0.013	0.159*
IP	0.855	**0.925**		-0.115*	0.099*	**-0.306***	0.089*	0.295*
IS	0.128	-0.438	-0.322		-0.023	-0.149*	-0.031	0.002
PFC	0.267	-0.588	0.632	-0.157		-0.064*	-0.007	-0.024
PPC	-0.570	-0.622	-0.627	-0.563	-0.389		-0.066*	-0.293*
PREF	0.329	0.123	0.776	-0.301	-0.096	-0.551		-0.050
PSR	0.277	0.497	0.802	0.005	-0.127	-0.825	-0.366	

Table 1. Coefficients $\phi_{ll'}$ suivis d'une étoile pour les valeurs significatives (au-dessus de la diagonale grise) et $Q_{ll'}$ (en-dessous de la diagonale) pour toutes les paires d'étiquettes. Les valeurs maximales et minimales de chaque coefficient sont notées en gras

3.2. Résultats

Pour les *coefficients phi*, la table 1 montre que la valeur maximale de 0.358 est obtenue pour la paire d'étiquettes (ou classes) CS et PSR, ce qui signifie que les tours de parole classés comme parlant de problèmes structurels sont aussi classés comme étant une suggestion, une recommandation ou une demande explicite, et inversement, ce qui semble cohérent. Quant à la valeur minimale de -0.306, elle se produit entre les classes IP et PFC. Cela suggère qu'en général, si un tour de parole apporte de l'information, il ne propose pas en même temps un engagement à une action dans le futur. En ce qui concerne le Q de Yule, la valeur maximale de 0.925 est atteinte pour les classes IP et IC, ce qui signifie qu'une des classes est presque incluse dans l'autre ; en fait, IC est presque incluse dans IP, car cette dernière a été assignée à la grande majorité des tours de parole (Ferschke et al., 2012). Ainsi, la plupart des tours de parole proposant une correction de l'information, amènent aussi de l'information.

De plus, comme il a été exposé dans la section précédente, une ACP a été effectuée sur la matrice des corrélations Φ. Le diagramme des valeurs propres de la figure 1 (gauche) montre qu'un faible pourcentage de la variance totale est expliqué par les deux premiers facteurs (moins de 22%), ce qui signifie que les étiquettes sont diversifiées et que l'information qu'elles contiennent peut difficilement être compressée. Le cercle des corrélations (figure 1 droite) est difficilement interprétable, un phénomène attendu au vu de la non significativité d'un grand nombre de *coefficients phi*.

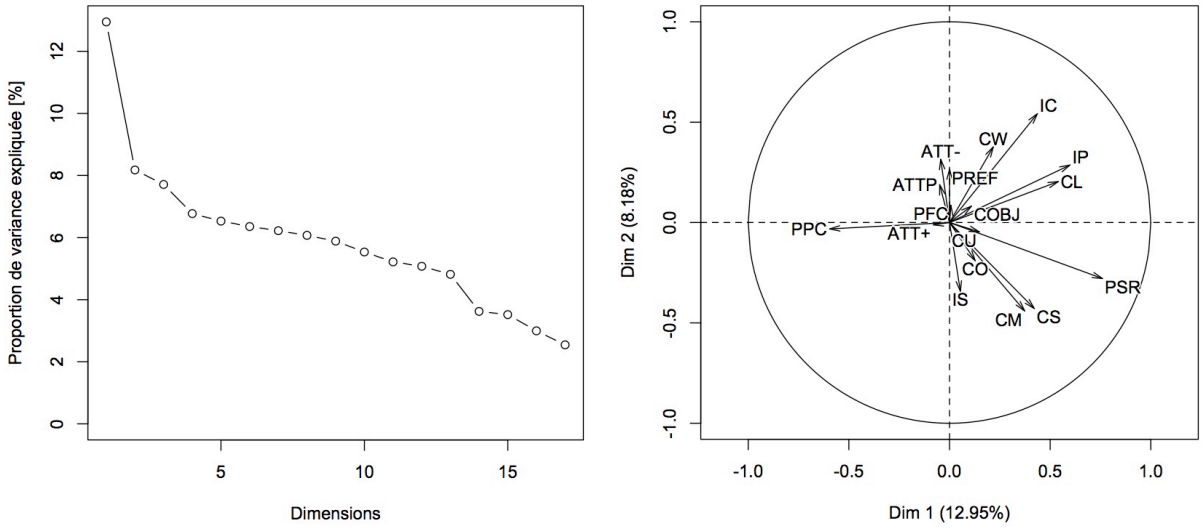


Figure 1. ACP sur la matrice des corrélations. Valeurs propres (gauche). Cercle des corrélations (droite)

4. Classification

4.1. Méthode

4.1.1. Prétraitement et caractéristiques

Comme déjà mentionné, les caractéristiques (*features*) utilisées dans ce travail sont uniquement linguistiques, afin de mieux comprendre l'importance spécifique et individuelle de ces dernières. La première étape a donc consisté à « nettoyer » les données de (Ferschke et al., 2012), pour en enlever les balises HTML (concernant principalement la mise en forme), les ponctuations découlant de la mise en forme du texte, les informations concernant les

utilisateurs, l'heure à laquelle le tour de parole a été posté et les symboles indiquant l'indentation du tour de parole par rapport au premier tour de parole de la discussion.

Ensuite, trois types de caractéristiques ont été extraites pour chaque tour de parole : les catégories morphosyntaxiques (CMS), les lemmes et le sens des verbes (selon *WordNet* (Fellbaum, 1998)). Les CMS et les lemmes ont été extraits à l'aide de *TreeTagger* (Schmid, 1994). L'extraction du sens des verbes a été effectuée à l'aide de *WordNet* et *TreeTagger*. Pour chaque tour de parole, les lemmes des mots considérés comme des verbes par *TreeTagger* ont été soumis à *WordNet*. Dans ce dernier, le premier sens du verbe proposé, pour des raisons d'automatisation, a été retenu, puis l'hyperonyme le plus général a été conservé. Ainsi, c'est ce dernier hyperonyme qui sera retenu comme caractéristique de ce tour de parole. Les verbes modaux ne sont pas traités par *WordNet*. Cependant, au vu de leur importance supposée pour la classification en actes de dialogue, il semblait intéressant de les ajouter explicitement au même titre que les hyperonymes traités par *WordNet*. A ce stade, trois tables de contingence sont créées : tours de parole - CMS, tours de parole - lemmes et tours de paroles - verbes, comptant le nombre de chaque caractéristique par tour de parole.

Les tours de parole qui n'étaient pas étiquetés ont été supprimés ; il s'agissait généralement de tours de parole soit trop longs et contenant de tout, soit écrits en français ou encore mal segmentés. Les tours de parole ne contenant aucune des caractéristiques décrites plus haut ont également été supprimés. Au final, la base de données a été réduite de 1 450 à 1 324 tours de parole, contenant, pour chaque table, 5 198 lemmes distincts, 57 CMS et 155 sens de verbes.

4.1.2. Classification multi-étiquette

Deux types d'approche sont couramment pratiqués pour la classification multi-étiquette (Tsoumakas et al., 2010) : le premier (*problem transformation*) consiste à recoder le jeu de données pour le transformer en problème de classification ordinaire, sans modification des algorithmes de classification ; le second (*algorithm adaptation*) adapte les algorithmes pour qu'ils puissent directement traiter des données multi-étiquette. Pour ce travail, il a été choisi d'utiliser le premier traitement, *i.e.* le recodage des données. Parmi les nombreux recodages possibles, celui du recodage binaire (*Binary Relevance* (BR)) a été choisi. Cela signifie que chaque tour de parole sera classé de façon binaire, *i.e.* comme faisant partie ou non d'une classe donnée (avec un classifieur pour chaque étiquette). Bien que ce recodage soit parfois critiqué, car il ne prend pas en compte les dépendances entre les étiquettes, il a ici plusieurs avantages : rendre les résultats comparables à ceux de (Ferschke et al., 2012) qui utilisent le même principe ; il a le mérite, en plus d'avoir une complexité computationnelle faible, d'être simple, intuitif, résistant au surapprentissage des combinaisons d'étiquettes et de pouvoir traiter les étiquetages irréguliers (Read et al., 2011) ; et il est particulièrement adapté aux situations où il n'y a pas de dépendance entre les étiquettes, ce qui semble être le cas ici (cf. section 3.2). De plus, (Luaces et al., 2012) proposent un indice qui mesure la dépendance entre toutes les étiquettes comme la moyenne des corrélations $\phi_{ll'}$ pour chaque paire d'étiquettes l et l', pondérée par le nombre d'individus (ici les tours de parole) communs $|l \cap l'|$:

$$\text{dépendance} = \frac{\sum_{l<l'} \phi_{ll'} |l \cap l'|}{\sum_{l<l'} |l \cap l'|}$$

Pour le jeu d'étiquettes de ce travail, cette dépendance vaut 0.10, ce qui correspond à la valeur la plus faible trouvée par (Luaces et al., 2012) dans la vingtaine de jeux de données qu'ils examinent, et qui confirme donc le choix de la méthode BR.

Finalement, étant donné que le nombre de tours de parole appartenant ou non à une étiquette est très variable, à l'instar de (Ferschke et al., 2012), des échantillons équilibrés ont été constitués, contenant, pour une étiquette donnée, un nombre identique de tours de parole lui appartenant ou non. Cet échantillon a été constitué une fois pour toutes pour chacune des classes. Ce choix a été fait pour éviter que les étiquettes les plus fréquentes soient plus facilement attribuées lors de la classification et inversement. Cependant, alors que le principe de la méthode BR est de sélectionner un individu, de le faire passer dans les différents classifieurs et d'obtenir toutes les étiquettes de cet individu, le choix des échantillons équilibrés conduit à une série de classifications séparées, ce qui influencera le choix de la méthode d'évaluation des résultats (cf. section 4.1.4).

4.1.3. Analyse discriminante et transformations de Schoenberg

Pour la classification, l'analyse discriminante linéaire (Fisher, 1936) a été appliquée, combinée avec une validation croisée sur 5 sous-échantillons (contre 10 utilisés par les auteurs de la base de données). Plus précisément, pour chacune des trois tables, une classification discriminante binaire pour chaque étiquette a été effectuée cinq fois.

A partir de chacune des tables de tours de parole ($i = 1,...,n$) - caractéristiques ($k = 1,...,p$), les dissimilarités du $chi2$ entre les paires de tours de parole peuvent être calculées dans l'échantillon d'apprentissage :

$$D_{ij} = \sum_{k=1}^{p} \frac{1}{\rho_k} \left(g_{ik} - g_{jk} \right)^2 \quad \text{avec} \quad g_{ik} = \frac{n_{ik}}{n_{i\bullet}} \quad \text{et} \quad \rho_k = \frac{n_{\bullet k}}{n_{\bullet\bullet}} \tag{1}$$

où n_{ik} est le nombre de fois qu'apparaît la caractéristique k dans le tour de parole i, $n_{i\bullet} = \Sigma_k n_{ik}$, $n_{\bullet k} = \Sigma_i n_{ik}$ et $n_{\bullet\bullet} = \Sigma_{ik} n_{ik}$. De la même manière, et en utilisant les poids des caractéristiques ρ_k calculés en (1), il est possible de calculer les dissimilarités D_{xj} du $chi2$ entre un tour de parole de l'échantillon de test x et un tour de parole de l'échantillon d'apprentissage j.

Deux critères d'analyse discriminante peuvent être utilisés. Le premier (*plus proches voisins*) attribue le nouveau tour de parole x (échantillon de test) au groupe contenant les individus d'apprentissage les plus proches de x en moyenne, *i.e.* à

$$\underset{g}{\operatorname{argmin}} \sum_{j=1}^{n_g} f_j^g D_{xj}$$

où $f_j^g = I(j \in g) / n_g$ est la distribution des tours j dans le groupe g, contenant n_g individus.

Le second critère (*plus proche centroïde*) attribue le tour test x au groupe d'apprentissage dont le centroïde est le plus proche, i.e. à

$$\underset{\tilde{g}}{\operatorname{argmin}} D_{x\tilde{g}}$$

où $\tilde{g}$ dénote le profil moyen des n_g individus constituant le groupe g. Les deux critères sont liés par le théorème de Huygens, valide pour toute dissimilarité euclidienne carrée :

$$\sum_{j=1}^{n_g} f_j^g D_{xj} = D_{x\tilde{g}} + \Delta_g \qquad \Delta_g = \frac{1}{2} \sum_{ij} f_i^g f_j^g D_{ij}$$

où Δ_g est l'inertie du groupe g, mesurant sa dispersion.

Les deux critères ci-dessus peuvent être étendus en considérant des *transformations de Schoenberg* (Schoenberg, 1938) de la forme $\tilde{D} = \varphi(D)$, ayant la propriété de conserver le caractère euclidien des dissimilarités (Bavaud, 2011), et en appliquant les règles

discriminantes à partir de $\tilde{D}_{xj}$ et $\tilde{D}_{ij}$.[2] Les transformations de Schoenberg peuvent être considérées comme des alternatives aux méthodes SVM, car toutes deux s'appuient sur un plongement de haute dimensionnalité des données de départ. Parmi les nombreuses transformations possibles visant à améliorer la classification, on utilise ici la transformation de puissance $\tilde{D} = D^q$, où $0 < q \leq 1$ (Bavaud, 2011). Dans ce qui suit, les classifications ont été produites avec les deux critères et q allant de 0.5 à 1, avec des incréments de 0.1 (pour un autre exemple d'application voir Cocco, 2012). Après expérimentation, le premier critère donne des résultats un peu moins bons que ceux du second, quoique comparables. Ainsi, seuls les résultats associés au second critère seront présentés dans ce qui suit.

4.1.4. Évaluation

Pour l'évaluation, en raison du choix de la méthode binaire et des échantillons équilibrés (cf. section 4.1.2) qui engendre des classifications séparées, il faut utiliser des méthodes (Tsoumakas et al., 2010) basées sur les étiquettes (*label-based*) et non sur les individus - tours de parole - (*example-based*). Parmi les différentes possibilités, les méthodes standards ont été appliquées: précision, rappel et F-mesure. Plus précisément, pour chaque étiquette ou classe ($l = 1,...,m$), les résultats des 5 validations croisées ont été agrégés pour former une matrice de confusion :

Classe c_l	Référence	
Décision	OUI	NON
OUI	VP_l	FP_l
NON	FN_l	VN_l

comptant le nombre de tours de parole attribués à l'étiquette l par la classification supervisée et classés dans l'étiquette l dans le corpus de référence (les vrais positifs VP_l), le nombre de tours de parole attribués à l'étiquette l par la classification supervisée et non classés dans l'étiquette l dans le corpus de référence (les faux positifs FP_l), le nombre de tours de parole non attribués à l'étiquette l par la classification supervisée et classés dans l'étiquette l dans le corpus de référence (les faux négatifs FN_l) et le nombre de tours de parole non attribués à l'étiquette l par la classification supervisée et non classés dans l'étiquette l dans le corpus de référence (les vrais négatifs VN_l). Ensuite, les mesures standards d'évaluation sont calculées, soit la précision (P_l), le rappel (R_l) et la F-mesure (F_l) :

$$P_l = \frac{VP_l}{VP_l + FP_l} \qquad R_l = \frac{VP_l}{VP_l + FN_l} \qquad F_l = \frac{2P_lR_l}{P_l + R_l} \tag{2}$$

Dans un second temps, pour évaluer la performance de la classification sur l'ensemble des classes, deux types de moyennes des mesures de (2) ont été calculées, la macro-moyenne:

$$P_{macro} = \frac{\sum_{l=1}^{m} P_l}{m} \qquad R_{macro} = \frac{\sum_{l=1}^{m} R_l}{m} \qquad F_{macro} = \frac{2P_{macro}R_{macro}}{P_{macro} + R_{macro}}$$

et la micro-moyenne :

$$P_{micro} = \frac{\sum_{l=1}^{m} VP_l}{\sum_{l=1}^{m} (VP_l + FP_l)} \qquad R_{micro} = \frac{\sum_{l=1}^{m} VP_l}{\sum_{l=1}^{m} (VP_l + FN_l)} \qquad F_{micro} = \frac{2P_{micro}R_{micro}}{P_{micro} + R_{micro}}$$

[2] Il faut noter que $\tilde{D}_{ij} := \varphi(D_{ij})$, $\tilde{\Delta}_g := \frac{1}{2}\sum_{ij} f_i^g f_j^g \tilde{D}_{ij}$, mais que $\tilde{D}_{x\tilde{g}} := \sum_j f_j^g \tilde{D}_{xj} - \tilde{\Delta}_g \neq \varphi(D_{x\tilde{g}})$.

Tandis que la macro-moyenne intègre toutes les classes avec le même poids et ne s'occupe donc pas de la taille des classes, la micro-moyenne prend tous les documents, ici les tours de parole, avec le même poids, donnant donc plus d'importance aux classes qui comptent le plus grand nombre de tours de parole (Yang, 1999).

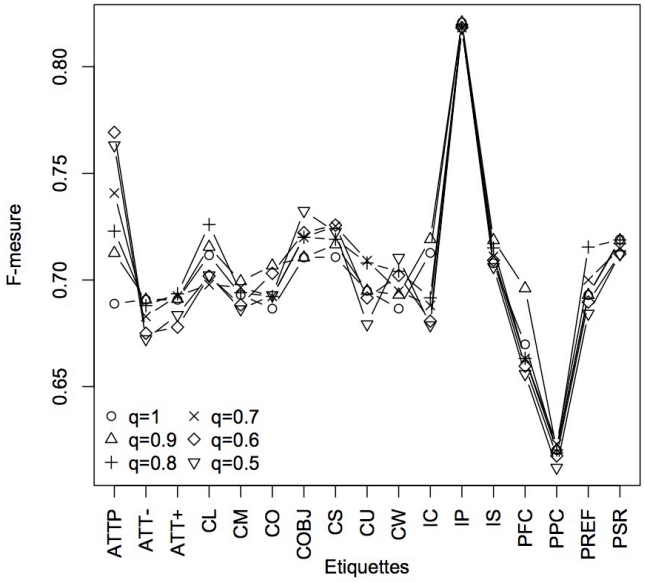
4.2. Résultats

Comme expliqué dans la section 4.1.2, des échantillons équilibrés ont été sélectionnés, et seule une partie des tours de paroles a donc été utilisée pour la classification binaire de chaque étiquette. Le nombre de tours de parole (exemples positifs et négatifs) utilisé pour chaque étiquette est donné ci-dessous :

Étiquette	Taille	Étiquette	Taille	Étiquette	Taille
ATTP	64	COBJ	54	IS	424
ATT-	174	CS	258	PFC	154
ATT+	276	CU	82	PPC	694
CL	422	CW	140	PREF	88
CM	228	IC	260	PSR	798
CO	96	IP	554		

A partir de ces échantillons, l'analyse discriminante selon les *plus proches voisins* et le *plus proche centroïde* a été effectuée, pour chacun des critères proposés dans la section 4.1.1, *i.e.* les lemmes, les CMS et le sens des verbes selon *WordNet*. Comme déjà mentionné, seuls les résultats pour *le plus proche centroïde* sont exposés (figures 2 à 4).

Globalement, dans les 3 figures, il est remarquable que la classification de l'étiquette IP donne la meilleure F-mesure et qu'elle varie peu avec la puissance q. Ce résultat semble cohérent avec ceux de (Ferschke et al., 2012) qui obtiennent toujours la meilleure F-mesure pour cette étiquette, quelle que soit la méthode utilisée. Un autre point récurrent pour les trois cas lors de l'évaluation pour l'ensemble des étiquettes (graphiques de droite), est que les micro- et macro- moyennes donnent des résultats très similaires pour la précision, le rappel et la F-mesure. Ceci est dû au fait que les résultats pour chacune des étiquettes sont très proches, indépendamment de la taille de l'échantillon. Aussi, le rappel est toujours plus élevé que la précision. Cela signifie que le nombre de faux positifs (tours de parole n'appartenant pas à une classe mais étiquetés comme y appartenant) est plus élevé que le nombre de faux négatifs (tours de parole appartenant à une classe mais étiquetés comme n'y appartenant pas).



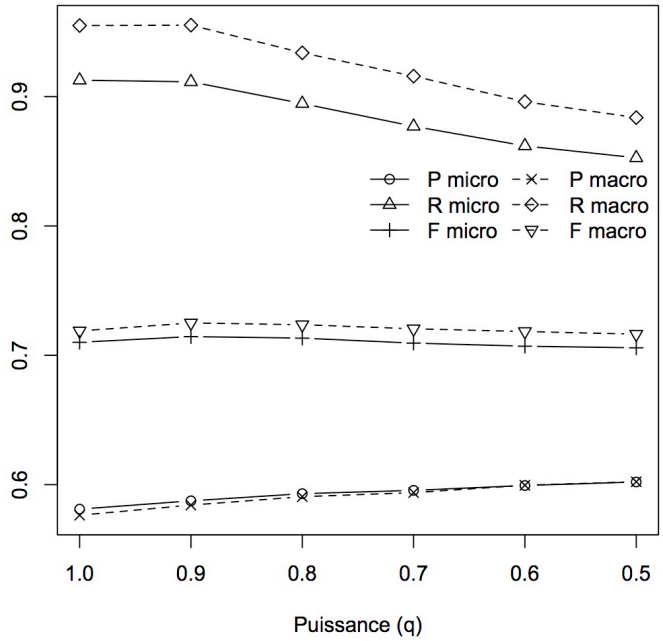
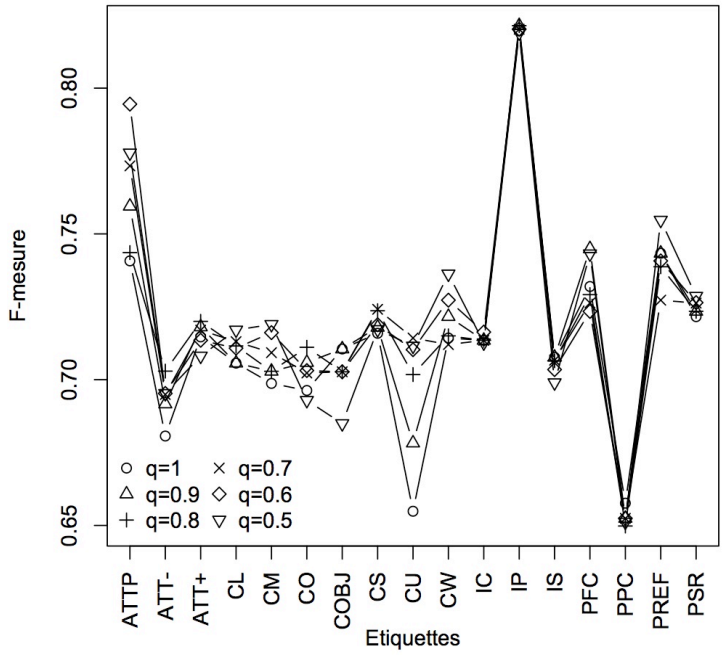


Figure 2. Classification avec les lemmes. F-mesure pour chaque étiquette selon la puissance q (haut). Macro et micro-moyennes pour la précision, le rappel et la F-mesure selon q (bas)

Avec les lemmes comme caractéristiques (figure 2), l'utilisation de la transformation de puissance q améliore la précision de la classification de l'ensemble des étiquettes (P_{macro} = 0.58 pour q = 1 et P_{macro} = 0.60 pour q = 0.5), mais diminue le rappel. Au final, la meilleure F_{macro} = 0.73 est obtenue pour q = 0.9. En regardant les résultats obtenus pour chaque étiquette, l'intérêt de la puissance q est plus marqué : q = 0.5 donne les meilleures F-mesures pour les étiquettes COBJ et CW; q = 0.9, pour CM, CO, IC, IS et PFC; et l'amélioration de la F-mesure obtenue pour l'étiquette ATTP est importante passant de 0.69 pour q = 1 à 0.77 pour q = 0.6.



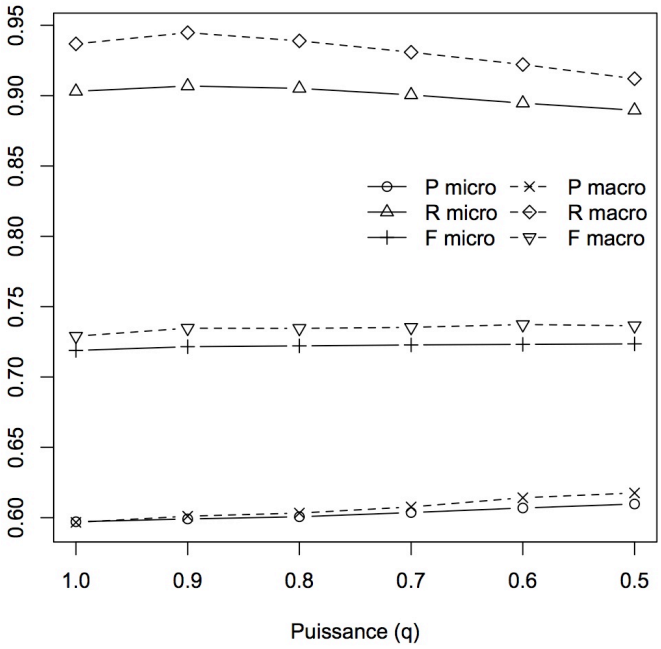
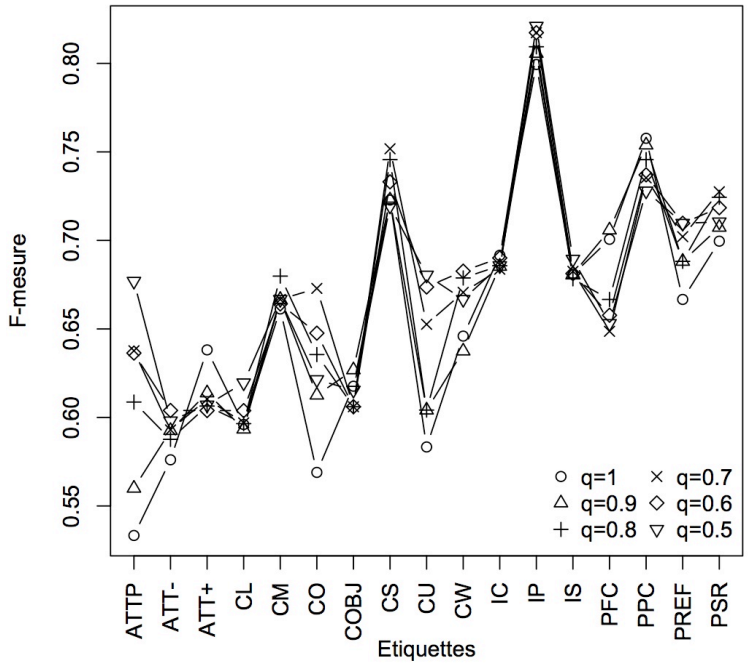


Figure 3. Classification avec les CMS. F-mesure pour chaque étiquette selon la puissance q (haut). Macro et micro-moyennes pour la précision, le rappel et la F-mesure selon q (bas)

Avec les CMS (figure 3), comme pour les lemmes, l'amélioration apportée par la transformation de puissance q est difficilement visible sur le résultat global (graphique de droite), même si l'on remarque que le meilleur résultat $F_{macro} = 0.74$ est obtenu pour $q = 0.6$. Ce dernier est aussi le meilleur sur l'ensemble des résultats. A nouveau, au niveau des étiquettes, la transformation q améliore les résultats pour la plupart des étiquettes (toutes sauf les étiquettes COBJ, IS et PPC), et en particulier pour les étiquettes ATTP ($F = 0.74$ pour $q = 1$ et $F = 0.79$ pour $q = 0.6$) et CU ($F = 0.65$ pour $q = 1$ et $F = 0.71$ pour $q = 0.7$).



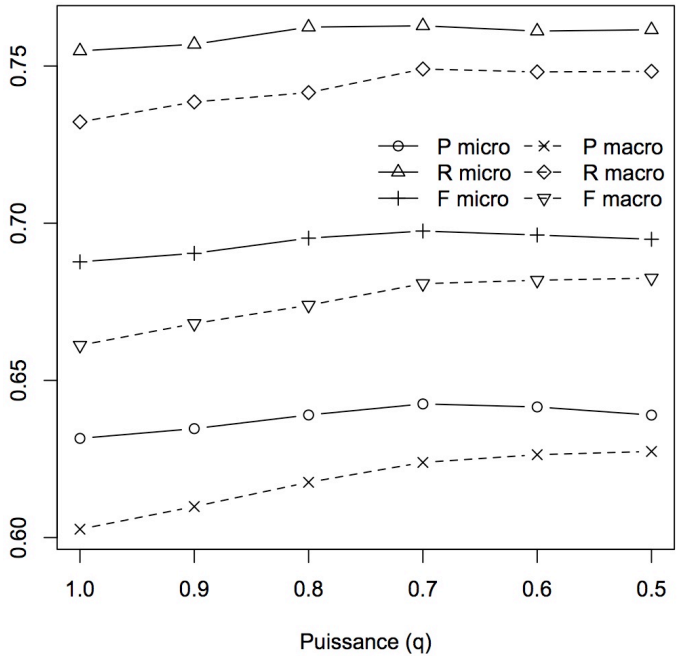


Figure 4. Classification avec le sens des verbes. F-mesure pour chaque étiquette selon la puissance q (haut). Macro et micro-moyennes pour la précision, le rappel et la F-mesure selon q (bas)

Au niveau de la classification pour l'ensemble des étiquettes avec le sens des verbes selon *WordNet* (figure 4, à droite), la première différence avec les autres caractéristiques est qu'ici les micro-moyennes donnent de meilleurs résultats que les macro-moyennes, car la majorité des étiquettes contenant plus de tours de parole donnent de meilleurs résultats (figure 4, à gauche). De plus, la meilleure classification est obtenue pour $q = 0.7$ ($F_{micro} = 0.70$). Concernant la classification par étiquette, comme pour les autres caractéristiques, la figure 4 de gauche montre que la puissance q améliore les résultats pour plusieurs étiquettes (toutes sauf ATT+, IC, et PPC). Un autre point plus important est que l'étiquette PPC, qui obtient des résultats assez faibles avec les lemmes ($F = 0.62$ pour $q = 0.7$) et les CMS ($F = 0.66$ pour $q = 1$), est bien mieux classifiée ici ($F = 0.76$ pour $q = 1$). Il en est de même pour l'étiquette CS ($F = 0.75$ pour $q = 0.7$, contre $F = 0.73$ pour $q = 0.6$ avec les lemmes et $F = 0.72$ pour $q = 0.7$ avec les CMS). Ainsi, même si le sens des verbes donne de moins bons résultats sur l'ensemble de la classification, cette caractéristique est plus discriminante pour ces deux étiquettes.

Comme attendu, la valeur maximale de la F-mesure de 0.74 (macro avec les CMS), est plus faible que celle obtenue par (Ferschke et al., 2012), qui trouvent une F-mesure maximale de 0.82 par micro-moyenne (et de 0.73 par macro-moyenne). Cependant, elle reste tout à fait comparable et élevée, étant donné qu'ici les résultats se calculent avec une caractéristique à la fois, et sans combiner les meilleurs résultats obtenus pour chaque caractéristique. Par contraste, (Ferschke et al., 2012) assemblent toutes les caractéristiques, en font une sélection (*feature selection*) et combinent les meilleurs résultats obtenus avec différentes méthodes.

5. Conclusion et développements futurs

La première partie concernant le lien entre les étiquettes a montré que l'annotation semblait cohérente et que les liens, mêmes s'ils existent, ne sont en majorité pas significatifs. Cette dernière constatation a permis de choisir la méthode de classification multi-étiquette, à savoir BR. En associant ce choix à l'analyse discriminante et les transformations de puissance de

Schoenberg, la classification a donné de bons résultats pour les trois caractéristiques linguistiques choisies, mais plus particulièrement avec les CMS. Cependant, au vu des meilleurs résultats du sens des verbes selon *WordNet* avec certaines étiquettes, et des lemmes sur d'autres étiquettes, il serait intéressant de combiner ces caractéristiques, par exemple en mélangeant les distances avec différents poids β :

$$D_{ij}^{\text{tot}} = \beta_{\text{lemme}} D_{ij}^{\text{lemme}} + \beta_{\text{CMS}} D_{ij}^{\text{CMS}} + (1 - \beta_{\text{lemme}} - \beta_{\text{CMS}}) D_{ij}^{\text{verbe}}$$

De la même manière, il serait possible d'y ajouter des caractéristiques situationnelles, telles que celles utilisées par (Ferschke et al., 2012), soit le temps entre les tours de parole, l'indentation entre les tours de parole, etc. De plus, la transformation de puissance montre une amélioration pour toutes les caractéristiques qui pourrait être aussi utilisée *avant* de mélanger les différentes distances. Il serait aussi intéressant d'explorer d'autres transformations de Schoenberg, susceptibles de donner de meilleurs résultats.

Une toute autre approche consisterait à explorer les liens entre étiquettes, même s'ils ne sont pas très importants ici, en utilisant une méthode *d'adaptation de l'algorithme* (au sens de la section 4.1.2) d'analyse discriminante (Park et Lee, 2008) afin qu'il puisse traiter globalement l'entièreté des étiquettes de chaque tour de parole.

Finalement, pour apprécier l'impact des caractéristiques proposées dans ce travail sur la performance, il faudrait les ajouter à celles utilisées par (Ferschke et al., 2012) en utilisant les algorithmes employés par ces auteurs et disponibles dans WEKA (Hall et al., 2009).

Références

Austin J. L. (1962). *How to do Things with Words.* Oxford University Press.

Bavaud F. (2011). On the Schoenberg Transformations in Data Analysis: Theory and Illustrations. *Journal of Classification*, 28(3): 297-314.

Boyer K., Ha E. Y., Phillips R., Wallis M., Vouk M. et Lester J. (2010). Dialogue Act Modeling in a Complex Task-Oriented Domain. In *Proc. of the SIGDIAL 2010 Conference,* pp. 297-305.

Cocco C. (2012). Discourse Type Clustering using POS n-gram Profiles and High-Dimensional Embeddings. In *Proc. of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 55-63.

Cohen W. W., Carvalho V. R. et Mitchell T. M. (2004). Learning to Classify Email into "Speech Acts". In *Proc. of EMNLP 2004*, pp. 309-316.

Colineau N. et Caelen J. (1995). Étude de marqueurs dans les actes de dialogue dans un corpus de conception. In *Actes de 01Design'05: Aspects communicatifs en conception, 4ème table ronde sur la conception,* pp. 127-139.

Fellbaum C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Ferschke O., Gurevych I. et Chebotar Y. (2012). Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. In *Proc. of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 777-786.

Fisher R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2): 179-188.

Goldstein J. et Sabin R. E. (2006). Using Speech Acts to Categorize Email and Identify Email Genres. In *Proc. of the 39th Annual Hawaii International Conference on System Sciences,* pp. 50b.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. and Witten I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations Newsletter,* 11(1): 10-18.

Kim S. N., Cavedon L. et Baldwin T. (2010). Classifying Dialogue Acts in One-on-One Live Chats. In *Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 862-871.

Luaces O., Dìez J., Barranquero J., del Coz J. J. et Bahamonde A. (2012). Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence*, 1(4): 303-313.

Park C. H. and Lee M. (2008). On applying linear discriminant analysis for multi-labeled problems. *Pattern Recognition Letters*, 29(7): 878-887.

Qadir A. et Riloff E. (2011). Classifying Sentences as Speech Acts in Message Board Posts. In *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing,* pp. 748-758.

Read J., Pfahringer B., Holmes G. et Frank E. (2011). Classifier chains for multi-label classification. *Machine Learning*, 85(3): 333-359.

Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proc. of the International Conference on New Methods in Language Processing*, pp. 44-49.

Schoenberg I. J. (1938). Metric Spaces and Positive Definite Functions. *Transactions of the American Mathematical Society*, 44(3): 522-536.

Searle J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Tsoumakas G., Katakis I. et Vlahavas I. (2010). Mining Multi-label Data. In Maimon O. and Rokach L. editors, *Data Mining and Knowledge Discovery Handbook.* Springer US.

Warrens M. J. (2008). On Association Coefficients for 2x2 Tables and Properties That Do Not Depend on the Marginal Distributions. *Psychometrika*, 73(4): 777-789.

Yang Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, 1(1-2): 69-90.

Yule G. U. (1900). On the Association of Attributes in Statistics: With Illustrations from the Material of the Childhood Society, &c. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 194(252-261): 257-319.

Yule G. U. (1912). On the Methods of Measuring Association Between Two Attributes. *Journal of the Royal Statistical Society,* 75(6): 579-652.